Cari dan kumpulkan satu Jurnal tentang Fuzzy C-Means

IJCCS, Vol.6, No.1, January 2012, pp. 101~110
ISSN: 1978-1520

Optimasi Cluster Pada Fuzzy C-Means Menggunakan Algoritma Genetika Untuk Menentukan Nilai Akhir

Putri Elfa Mas`udia*[1], Retantyo Wardoyo[2]
[1]Universitas Muhammadiyah Malang; Jl. Raya Tlogomas 246, Malang, (0341) 463513
[2]Jurusan Ilmu Komputer dan Elektronika, FMIPA UGM, Yogyakarta
e-mail: ***[1]aishsalma@gmail.com**, [2]rw@ugm.ac.id

Abstrak

Nilai akhir mahasiswa dapat ditentukan dengan berbagai cara, beberapa diantaranya menggunakan range nilai, standart deviasi, dll. Dalam penelitian ini akan ditawarkan sebuah metode baru untuk menentukan nilai akhir mahasiswa menggunakan clustering dalam hal ini adalah Fuzzy C-Means.

Fuzzy C-Means digunakan untuk mengelompokkan sejumlah data dalam beberapa cluster. Tiap data memiliki derajat keanggotaan pada masing-masing cluster antara 0-1 yang diukur melalui fungsi objektif. Pada Fuzzy C-Means ini fungsi objektif diminimumkan menggunakan iterasi yang biasanya terjebak dalam optimum lokal. Algoritma genetika diharapkan dapat menangani masalah tersebut karena algoritma genetika berbasis evolusi yaitu dapat mencari individu terbaik melalui operasi genetika (seleksi, crossover, mutasi) dan dievaluasi berdasarkan nilai fitness.

Penelitian ini bertujuan untuk mengoptimasi titik pusat cluster pada Fuzzy C-Means menggunakan algoritma genetika. Hasilnya, bahwa dengan menggunakan GFS didapatkan fungsi objektif yang lebih kecil daripada menggunakan FCM, walaupun membutuhkan waktu yang relative besar. Meskipun selisih antara FCM dan GFS tidak terlalu besar namun hal tersebut berpengaruh pada anggota cluster

Kata kunci— *clustering, Fuzzy C-Means, algoritma genetika*

Abstract

The final grade of students could be determined in various ways, some of which use a range of values, deviation standard, etc. In this study will be offered a new method for determining final grades of students by using the clustering method. In this research the clustering method that will be used is the Fuzzy C-Means (FCM).

Fuzzy C-Means is used to group a number of data in multiple clusters. Each data has a degree of membership (the range value of membership degree is 0-1). Membership degree is measured through the objective function. In Fuzzy C-Means, objective function is minimized by using iteration and is usually trapped in a local optimum. Genetic algorithm is expected to handle these problems. The operation of genetic algorithm based on evolution that is able to find the best individuals through genetic operations (selection, crossover and mutation) and evaluated based on fitness values.

This research aims to optimize the cluster center point of FCM by using genetic algorithms. The result of this research shows that by combining the Genetic Algorithm with FCM could obtained a smaller objective function than using FCM, although it takes longer in execution time. Although the difference of objective function that produced by FCM and FCM-Genetic Algorithm combination is not too big each other, but it takes effect on the cluster members.

Keywords— *clustering, fuzzy c-means, genetic algorithm*

Received March 2nd,2012; Revised March 15th, 2012; Accepted June 20th, 2012

1. PENDAHULUAN

Terdapat berbagai cara untuk menentukan nilai akhir huruf mahasiswa, diantaranya adalah menggunakan skala pasti maupun standar deviasi, Tapi dengan cara yang seperti itu mungkin akan menyulitkan dosen jika terjadi kondisi-kondisi tertentu yang memerlukan pertimbangan dosen, pada sistem ini ditawarkan alternatif baru untuk penilaian dimana dosen diberi kebebasan (hak istimewa) untuk menentukan standart nilai menggunakan metode *clustering* dalam hal ini menggunakan *fuzzy clustering*. *Cluster* adalah sekelompok sesuatu yang mempunyai kesamaan sifat [1]. *Fuzzy clustering* memainkan peran yang paling penting dalam pencarian struktur dalam data [2].

Metode ini dipilih karena dengan metode ini, bisa ditentukan jumlah *cluster* yang akan dibentuk. Dengan penentuan jumlah *cluster* di awal, bisa diatur keragaman nilai akhir atau pelabelan nilai sesuai dengan *cluster*nya.

Pada *fuzzy c-means* nilai awal titik pusat *cluster* dibangkitkan secara acak sehingga terjadi optimum lokal,dimana proses selanjutnya bergantung pada nilai awal yang dibangkitkan secara acak, disini akan digunakan algoritma genetika untuk mengoptimasi nilai awal titik pusat *cluster*.

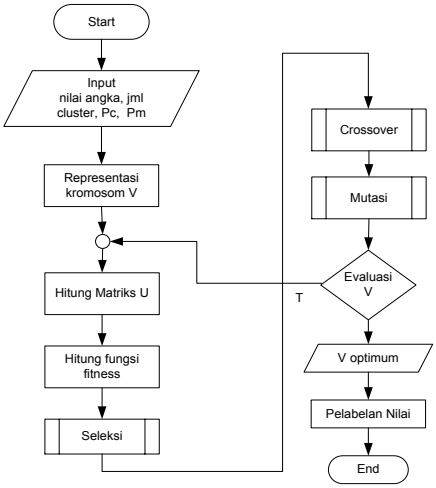
Algoritma genetika (GA) sebagai teknik optimasi dapat diterapkan pada *clustering* yang berbasis optimasi fungsi tujuan. Pada pendekatan GA untuk fuzzy *clustering* fungsi *fitness* diambil dari fungsi objektif yang diminimumkan, yaitu Jm(U,V) [3].

Pada pendekatan algoritma genetika, Pada setiap generasi, kromosom dievaluasi berdasarkan nilai fungsi *fitness* [4], untuk pencarian titik pusat cluster dilakukan dengan cara meng-evolusikan matrik pusat *cluster* (seleksi, crossover dan mutasi)dengan fungsi *fitness* menggunakan fungsi objektif yang terdapat pada *fuzzy c-means*

2. METODE PENELITIAN

Secara umum sistem yang akan dibuat dalam penelitian ini adalah sistem untuk menentukan nilai akhir huruf mahasiswa dengan menggunakan perhitungan *Fuzzy clustering* dan algoritma genetika dimana algoritma genetika digunakan untuk optimasi titik pusat *cluster* pada *Fuzzy C-means*.

Dengan menggunakan algoritma genetika, nilai awal titik pusat *cluster* V dibangkitkan secara acak, kemudian nilai V tersebut digunakan untuk menghitung *matriks u*. Selanjutnya nilai V ini akan dievolusikan menggunakan seleksi, *crossover* dan mutasi untuk mendapatkan nilai V yang paling optimum. Deskripsi umum sistem tersebut ditunjukkan pada Gambar 1

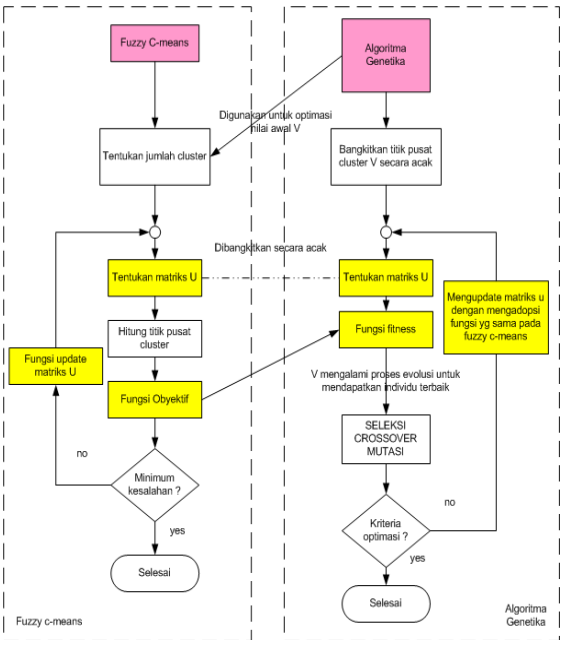


Gambar 1. Gambaran Umum Sistem

Langkah-langkah kegiatan yang akan dilakukan untuk permasalahan ini adalah:

1. Representasi solusi dan spesifikasi masukan dan keluaran.
2. Penentuan ukuran parameter genetika meliputi ukuran populasi, jumlah generasi, probabilitas *crossover*, probabilitas mutasi, maxIterasi.
3. Menentukan fungsi *fitness* yang digunakan, dalam hal ini fungsi *fitness* yang digunakan adalah fungsi objektif yang meminimumkan jarak antar data dengan titik pusat *cluster* (V) yang terdapat pada *Fuzzy C-means*.
4. Menentukan encoding kromosom yang digunakan yaitu *real encoding*.
5. Pemilihan fungsi seleksi yaitu menggunakan metode rank, *crossover* menggunakan *one cut point*, dan mutasi dengan mengubah gen secara acak.
6. Melakukan pengupdatetan *matriks u* yang merupakan derajat keanggotaan data pada masing-masing titik pusat *cluster* (V).

Langkah-langkah tersebut direpresentasikan pada Gambar 2.



Gambar 2. Penerapan Algoritma Genetika pada *Fuzzy C-means*

2.1 Perancangan Sistem

Pada penelitian ini digunakan metode GFS (*Genetic Fuzzy System*) untuk menentukan nilai akhir kuliah mahasiswa. Dari data nilai mahasiswa akan dikelompokkan menjadi beberapa *cluster*, dalam hal ini dosen berhak menentukan jumlah *cluster* yang diinginkan berdasarkan nilai terendah dan nilai tertinggi pada data nilai mahasiswa. Kemudian dari hasil *cluster* yang didapatkan, dosen dapat memberi pelabelan nilai yang sesuai

Dalam *cluster*ing, persoalan sebenarnya adalah persoalan untuk mencari titik pusat *cluster* yang paling optimum berdasarkan fungsi objektif. Pada penelitian ini, persoalan ini akan diselesaikan dengan menggunakan algoritma genetika. Dalam algoritma genetika, solusi dari persoalan direpresentasikan dengan kromosom.

2. 1.1 Representasi Kromosom

Pada penelitian ini nilai awal kromosom akan dibangkitkan secara acak sebanyak n *cluster* (V) dan sebanyak jumlah populasi kromosom m. Jadi satu gen mewakili satu cluster. Banyaknya jumlah *cluster* ditentukan oleh dosen berdasarkan nilai minimal dan maksimal dari

data nilai mahasiswa, sedangkan jumlah awal kromosom yang digunakan pada penelitian ini adalah 50, hal ini dikarenakan algoritma genetika membutuhkan ruang solusi yang cukup besar dan ini berdasarkan rekomendasi dari De Jong 1975 yang terdapat pada dasar teori.

Tabel 1. Contoh Representasi Kromosom

	v1	v2	v3	v4	...	vn
Kromosom 1	1.4	5.0	7.1	1.5	...	0.4
Kromosom 2	0.5	4.4	6.0	1.3	...	2.6
...	...	...	...	...	...	...
Kromosom m	1.7	2.1	6.3	5.5	...	2.7

2.1.2 Hitung Matrik U

Matriks u adalah derajat keanggotaan data pada tiap-tiap *cluster*, Berdasarkan rumus perhitungan matriks u membutuhkan input berupa data nilai dan nilai titik pusat cluster. Rumus untuk menghitung matriks u ditunjukkan pada Rumus 1

$$\mu_{ik} = \frac{\left[\sum_{j=1}^{m}(X_{ij} - V_{kj})^2\right]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{m}(X_{ij} - V_{kj})^2\right]^{\frac{-1}{w-1}}} \quad (1)$$

2. 1.3 Fungsi Fitness

Fungsi *fitness* merupakan ukuran kinerja suatu individu agar tetap bertahan hidup dalam lingkungannya. Dalam algoritma genetika fungsi *fitness* adalah fungsi objektif dari masalah yang akan dioptimasi.

Fungsi fitness yang digunakan dalam penelitian ini mengadopsi rumus fungsi objektif yang terdapat pada Fuzzy C-means yang meminimumkan jarak antara data dengan titik pusat cluster. Rumus fungsi objektif yang digunakan dalam fungsi fitness ditunjukkan pada Rumus 2

$$P_t = \sum_{i=1}^{n}\sum_{k=1}^{c}\left(\left[\sum_{j=1}^{m}(X_{ij} - V_{kj})^2\right](\mu_{ik})^w\right) \quad (2)$$

2.1.4 Seleksi

Pada penelitian ini, seleksi kromosom dilakukan dengan menggunakan metode *rank* yang bertujuan untuk mendapatkan prosentase yang lebih merata. Seleksi ini menggunakan *fitness* skala dalam memilih kromosom dari populasi. Untuk menghitung *fitness* skala dari suatu kromosom, maka perlu diketahui terlebih dahulu jumlah seluruh *fitness* dalam populasi. Prosedur seleksi adalah sebagai berikut:

1. Menghitung nilai *fitness* (f_k) tiap kromosom
2. Urutkan nilai *fitness* secara *descending* menggunakan library value comparator.
3. Beri prosentase pada tiap kromosom dengan membagi tiap kromosom dengan jumlah total kromosom dikalikan 100%
4. Hitung akumulasi prosentase
5. Pilih induk yang menjadi kandidat untuk di-*crossover* dengan cara:

a. Bangkitkan suatu bilangan [0,1]
b. Memilih kromosom dengan nilai kumulatif q_k yang lebih besar dan yang paling mendekati dari nilai random sebagai parent.

6. Lakukan langkah 5 sebanyak ukuran populasi.

2.1.5 Crossover

Metode *crossover* yang digunakan pada penelitian ini adalah metode *one cut point* karena mudah diterapkan pada kromosom yang diencodingkan secara real. Tidak semua kromosom mengalami *crossover*, kromosom yang di *crossover* sesuai dengan probabilitas *crossover* (Pc). Jika terdapat dua kromosom yang nilai randomnya lebih kecil dari Pc, maka langsung di *crossover*, jika lebih dari dua kromosom maka di*crossover* secara berurutan dan hanya dipilih satu *offspring* terbaik yang dapat menggantikan induk.

Prosedur persilangan satu titik sebagai berikut:

1. Menentukan Pc dan nilai acak p, dimana Pc adalah peluang *crossover* dan p adalah nilai yang dibangkitkan secara acak. Hanya p yang lebih kecil dari Pc yang akan di *crossover*. Nilai awal Pc yang digunakan pada penelitian ini adalah 0,6 yang merupakan rekomendasi dari De jong yang mengacu pada dasar teori pada Bab III
2. Tentukan posisi *crossover* dengan cara membangkitkan bilangan acak dengan batasan 1 sampai n-1, dimana n adalah panjang kromosom. Misalkan didapat posisi *crossover* adalah 20 maka kromosom induk akan dipotong mulai gen ke 21 sampai gen ke n, kemudian potongan gen tersebut saling ditukar antar induk.
3. Posisi *cut-point crossover* dipilih menggunakan bilangan acak 1-(n-1) sebanyak jumlah *crossover* yang terjadi.

2.1.6 Mutasi

Metode mutasi dilakukan dengan cara uniform, yaitu dengan mengganti satu gen yang dipilih secara acak dengan suatu nilai baru dalam hal ini nilai baru tersebut juga ditentukan secara acak. Jumlah kromosom yang mengalami mutasi dalam satu populasi ditentukan oleh parameter peluang mutasi (p_m).

Langkah-langkah mutasi sebagai berikut:

1. Menentukan Pm dan p, dimana Pm adalah peluang mutasi dan p adalah nilai yang dibangkitkan secara acak. Nilai awal Pm yang digunakan pada penelitian ini adalah 0,001 yang merupakan rekomendasi dari De jong yang mengacu pada dasar teori yang terdapat pada Bab III
2. Untuk memilih posisi gen yang akan mengalami mutasi dilakukan dengan cara membangkitkan bilangan acak (p) antara 0 sampai 1. Jika $p < p_m$, maka gen akan dimutasi, misal p_m kita temukan 10% maka diharapkan ada 10% dari populasi yang mengalami mutasi.
3. Bangkitkan bilangan acak untuk mengubah nilai gen yang ditentukan secara acak dengan nilai yang acak.

3. HASIL DAN PEMBAHASAN

Pembahasan pada sistem ini meliputi perbandingan *clustering* menggunakan FCM dan GFS dilihat dari nilai objektif, hasil pengelompokan data, waktu dan analisis pengaruh parameter genetika terhadap nilai objektif dan lama proses sistem. Pengaruh parameter yang dipelajari dalam penelitian ini dikelompokkan menjadi 4 bagian yaitu:

1. Hasil *clustering* menggunakan FCM
2. Pengaruh parameter ukuran populasi.
3. Pengaruh parameter jumlah iterasi
4. Pengaruh kombinasi parameter Pc dan Pm

3.1 Hasil Percobaan

Hasil percobaan pertama adalah hasil clustering dengan menggunakan parameter popsize, Pc, Pm rekomendasi dari De jong, yaitu populasi = 50, Pc = 0,6, Pm=0,001, maksimal iterasi = 100 dan jumlah cluster = 4 Data tersebut ditampilkan dalam Tabel 2 berikut ini.

Tabel 2 Data Nilai Mahasiswa Percobaan 1

Nim	Nama	Nilai	*Cluster*	Pusat *Cluster*	Nilai Huruf
1031140052	Aditya Dani Rosanto	90.0	4	94.11973	A
1031140065	Adhi Susanto	72.0	2	73.054245	B
1031140003	Arynda Putri Mertabuan	35.0	3	24.373365	D
1031140099	Avinda Ika Nadiasari	25.0	3	24.373365	D
1031140059	Bagus Rakhmadian	63.0	1	73.054245	C
1031140009	Deka Cahya Endradewi	83.0	2	73.054245	B
1031140056	Dian Munika Ayu	44.0	1	48.716026	C
1031140083	Diana Mayangsari	10.0	3	24.373365	D
1031140076	Elviera Savitrie P	53.0	1	48.716026	C
1031140024	Farah Azizah Ansi	95.0	4	94.11973	A
1031140092	Firman Adiyansyah	80.0	2	73.054245	B
1031140041	Ira Reza Permata	75.0	2	73.054245	B
1031140072	M. Hamzah Mashuri	65.0	1	73.054245	C
1031140066	Madarina Imanisari	95.0	4	94.11973	A
1031140097	Mokh Zafril	51.0	1	48.716026	C
1031140033	Nadia Putri	50.0	1	48.716026	C
1031140046	Putri Ardiani	25.0	3	24.373365	D
1031140017	Raj Nur	45.0	1	48.716026	C
1031140045	Revidya Dewantari	75.0	2	73.054245	B
1031140022	RR Eka Regina	100.0	4	94.11973	A

Dari hasil yang ditunjukkan pada Tabel 6.1, didapatkan fungsi objektif sebesar 800,557 dan membutuhkan waktu 869 ms

Percobaan selanjutnya sistem dijalankan dengan menggunakan data parameter genetika dengan jumlah *cluster* yang sama, tetapi parameter berbeda. Jumlah *cluster* = 4, ukuran populasi 50, probabilitas *crossover* (Pc) 0.3, probabilitas mutasi (Pm) 0.9, pembobot 2 dan max iterasi 300.

Tabel 3. Data Nilai Mahasiswa Percobaan 2

Nim	Nama	Nilai	*Cluster*	Pusat *Cluster*	Nilai Huruf
1031140052	Aditya Dani Rosanto	90.0	2	94.11973	A
1031140065	Adhi Susanto	72.0	3	73.054245	B
1031140003	Arynda Putri Mertabuan	35.0	4	24.373365	D
1031140099	Avinda Ika Nadiasari	25.0	4	24.373365	D
1031140059	Bagus Rakhmadian	63.0	3	73.054245	B
1031140009	Deka Cahya Endradewi	83.0	3	73.054245	B
1031140056	Dian Munika Ayu	44.0	1	48.716026	C
1031140083	Diana Mayangsari	10.0	4	24.373365	D
1031140076	Elviera Savitrie P	53.0	1	48.716026	C
1031140024	Farah Azizah Ansi	95.0	2	94.11973	A
1031140092	Firman Adiyansyah	80.0	3	73.054245	B
1031140041	Ira Reza Permata	75.0	3	73.054245	B
1031140072	M. Hamzah Mashuri	65.0	3	73.054245	B
1031140066	Madarina Imanisari	95.0	2	94.11973	A
1031140097	Mokh Zafril	51.0	1	48.716026	C
1031140033	Nadia Putri	50.0	1	48.716026	C
1031140046	Putri Ardiani	25.0	4	24.373365	D
1031140017	Raj Nur	45.0	1	48.716026	C
1031140045	Revidya Dewantari	75.0	3	73.054245	B
1031140022	RR Eka Regina	100.0	2	94.11973	A

Dari hasil percobaan tersebut terlihat sedikit perbedaan pengelompokkan nilai. Fungsi objektif terakhir yang dihasilkan pada percobaan kedua adalah 662.817 dan membutuhkan waktu 2795 ms, nilai ini jauh lebih baik dibandingkan percobaan sebelumnya, data inilah yang digunakan sebagai acuan untuk membandingkan hasilnya dengan metode FCM.

3.2 Hasil Perbandingan dengan FCM

Dilakukan perbandingan menggunakan FCM dan GFS dengan data yang sama dan jumlah cluster yang sama.Metode GFS dijalankan dengan menggunakan data parameter genetika dengan jumlah *cluster* = 4, ukuran populasi 50, probabilitas *crossover* (Pc) 0.3, probabilitas mutasi (Pm) 0.9, pembobot 2 dan max iterasi 300. Data ini ditentukan secara acak. Fungsi objektif terakhir yang dihasilkan adalah 662.817 dan membutuhkan waktu 2795 ms.

Pengelompokan data nilai mahasiswa menggunakan FCM *(Fuzzy C-Means)* dilakukan dengan nilai parameter jumlah *cluster* = 4, pembobot = 2, maxIterasi = 100.Dari percobaan tersebut diketahui bahwa hasil *cluster* data nilai mahasiswa hasilnya sama baik menggunakan FCM maupun GFS, sedangkan dilihat dari waktu yang digunakan untuk mengklaster data, FCM relative lebih singkat disbanding GFS.

Fungsi objektif yang dihasilkan FCM lebih besar dari GFS yaitu 662.82, hal ini dikarenakan proses penyelesaian optimasi secara klasik sering menghantar pada penyelesaian yang merupakan optimum lokal sedangkan algoritma genetika akan selalu berevolusi untuk menemukan individu yang lebih baik tergantung dari parameter masukan.

3.3 Pengaruh Parameter Ukuran Populasi

Percobaan pertama adalah dengan mengambil jumlah *cluster* = 4, ukuran populasi = 10, Pc = 0.4, dan Pm = 0.8, jumlah iterasi = 100. Satu nilai parameter yang sama dihitung sebanyak 5 kali. Percobaan selanjutnya adalah dengan menambah ukuran populasi dan tetap menggunakan nilai parameter yang sama seperti pada percobaan pertama. Setelah dilakukan 10 kali percobaan dengan ukuran populasi yang berbeda diperoleh hasil seperti Tabel2

Tabel 4 Pengaruh Parameter Ukuran Populasi

Ukuran Populasi	Nilai Objektif (*B*)	Waktu Proses (ms)	Ukuran Populasi	Nilai Objektif (*B*)	Waktu Proses (ms)
10	682,817	188	50	663,330	1058
	753,889	188		670,899	1035
	668,92	186		663,210	1162
	678,316	191		665,880	933
	674,107	190		664,535	1079
20	674,643	374	80	662,889	1822
	667,502	458		663,540	1740
	666,024	382		663,730	1701
	670,111	379		663,880	1620
	690,206	360		663,117	1729
30	669,843	566	100	663,09	2610
	663,751	650		663,089	2256
	668,497	575		663,477	2238
	666,702	398		662,997	2265
	667,801	450		663,137	2647
40	667,162	779	150	662,879	3406
	671,35	772		662,861	3505
	664,501	820		662,863	3762
	676,153	726		663,070	3065
	665,313	650		662,910	3889

Dari tabel diatas dapat dilihat bahwa semakin besar jumlah populasi maka nilai objektifnya semakin kecil dan stabil dan membutuhkan waktu yang relative lebih besar. Hal ini dikarenakan semakin banyak jumlah populasi kemungkinan ruang solusi juga semakin besar

dan peluang untuk memperoleh individu yang lebih baik semakin besar walaupun harus dibayar dengan waktu yang lebih banyak.

3.4 Pengaruh Parameter Jumlah Iterasi

Percobaan pertama adalah dengan mengambil jumlah iterasi awal = 100, ukuran populasi = 50, Pc = 0.8, dan Pm = 0.2. Percobaan selanjutnya adalah dengan menambah jumlah iterasi dan tetap menggunakan nilai parameter lain yang sama seperti pada percobaan pertama. Setelah dilakukan 8 kali percobaan diperoleh nilai objektif seperti Tabel 5

Tabel 5 Pengaruh Parameter Jumlah Iterasi

Jumlah Iterasi	Nilai Objektif (*B*)	Waktu Proses (ms)	Jumlah Iterasi	Nilai Objektif (*B*)	Waktu Proses (ms)
50	664,692	535	400	662,911	3264
	667,569	552		662,830	3113
	675,553	449		662,901	3963
	681,370	412		662,860	3888
	670,824	474		662,849	3246
100	667,04	1107	500	662,839	4934
	663,46	1111		662,852	4068
	664,091	1064		**662,817**	**4065**
	665,651	1079		662,85	4474
	663,920	1095		662,833	4808
200	663,213	1735	800	662,830	7331
	663,240	1943		662,827	6230
	663,490	1909		**662,818**	**6505**
	663,831	1928		**662,817**	**7200**
	663,197	1934		**662,819**	**7730**
300	662,970	2936	1000	662,830	7819
	662,890	2840		**662,818**	**8941**
	662,860	2808		662,820	7556
	662,88	2559		**662,818**	**7467**
	662,95	2728		**662,816**	**7556**

Dari Tabel 6.5 diketahui bahwa nilai objektif terbaik (yang lebih kecil daripada FCM) diperoleh mulai dari jumlah iterasi 500 dan seterusnya, dimana nilai objektif yang diperoleh yaitu 662,817. Perbedaan nilai yang terjadi pada iterasi yang sama dikarenakan setiap kali perhitungan yang berbeda, proses seleksi, *crossover*, dan mutasi yang terjadi juga berbeda tergantung bilangan acak yang didapatkan. Hasil ini bisa lebih baik maupun lebih buruk.

3.5 Pengaruh Kombinasi Pc dan Pm

Percobaan pertama adalah dengan mengambil nilai Pc = 0.6, dan Pm = 0,001 ukuran populasi = 50, jumlah iterasi = 500. Percobaan selanjutnya adalah dengan menambah nilai Pm dan tetap menggunakan nilai parameter lain yang sama seperti pada percobaan pertama kemudian dilanjutkan mengubah Pc = 0,1 dan seterusnya. Satu parameter dihitung sebanyak 3 kali

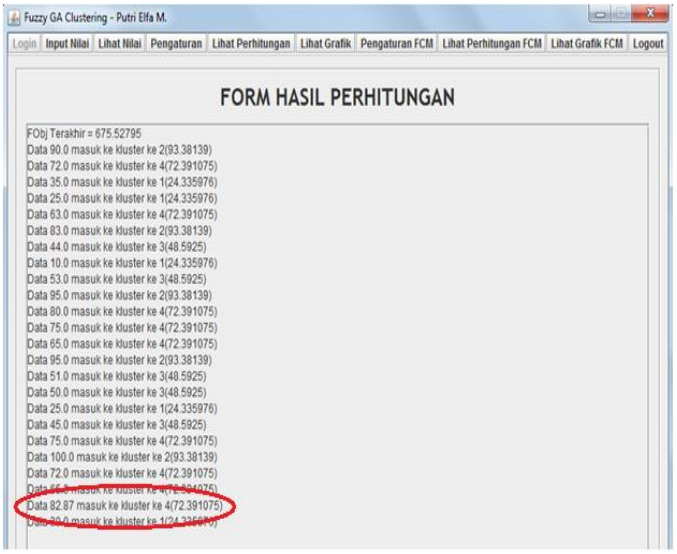
Hasil pengujian menunjukkan bahwa perubahan nilai Pm sangat berpengaruh pada hasil fungsi objektif dimana nilai Pm yang dapat menghasilkan fungsi objektif terbaik adalah antara 0,7- 0,9. Nilai parameter Pm yang cenderung besar ini menyebabkan jumlah kromosom baru hasil mutasi lebih banyak untuk generasi berikutnya sehingga nilai objektif terbaik akan mudah dicapai.

Untuk perubahan nilai Pc tidak terlalu berpengaruh secara signifikan pada hasil fungsi objektif, namun ketika kombinasi Pc = 0,3 dan Pm = 0,9 setelah 10 kali perhitungan semua hasil proses menghasilkan nilai objektif yang lebih rendah dari FCM. Hal ini dikarenakan parameter Pc=0,3 tidak terlalu kecil tapi juga tidak besar.

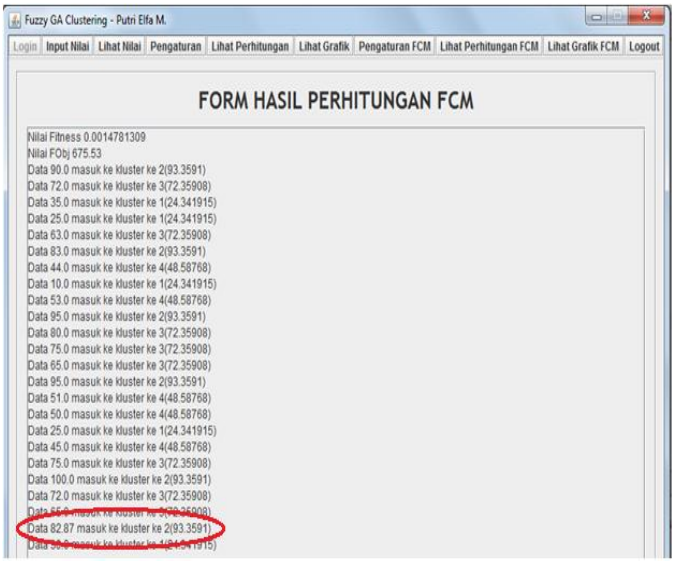
3.6 Perbandingan FCM dan GFS dengan Data Terperinci

Percobaan selanjutnya dilakukan dengan mengubah-ubah data yang lebih presisi untuk menguji apakah fungsi objektif yang berbeda dengan selisih yang sangat kecil berpengaruh terhadap anggota *cluster*. Hal ini terbukti pada hasil yang diperoleh pada data 82,87 dimana fungsi objektif yang dihasilkan dari proses GA dan GFS berbeda dengan selisih 0,01. Fungsi objektif hasil proses FCM adalah 675,53 dan fungsi objektif hasil proses GFS adalah 675,527. Hasil percobaan diatas menunjukkan bahwa dengan selisih fungsi objektif yang kecil yaitu 0,01 dapat berpengaruh terhadap anggota *cluster* jika datanya adalah data yang lebih presisi (dengan beberapa angka dibelakang koma). Hal ini dikarenakan walaupun fungsi objektif hanya selisih 0,01 tapi hal tersebut berpengaruh terhadap titik pusat *cluster*. Data yang lebih terinci, akan membutuhkan ketelitian lebih untuk memilih anggota pada *cluster*.

Pada hasil proses FCM_GA data nilai 82,87 masuk ke *cluster* 4 dengan titik pusat *cluster* adalah 72,391 sedangkan hasil proses FCM, data nilai 82,87 masuk ke *cluster* 2 dengan titik pusat *cluster* adalah 93,359.



Gambar 3. Hasil Perhitungan GFS



Gambar 4. Hasil Perhitungan FCM

4. KESIMPULAN

1. Algoritma genetika (GFS) dapat digunakan untuk mengelompokkan data nilai mahasiswa dengan nilai fungsi objektif yang lebih kecil daripada menggunakan FCM.
2. Meskipun selisih nilai objektif hasil proses GFS dan FCM tidak terlalu besar tetapi hal tersebut berpengaruh terhadap keanggotaan pada tiap-tiap *cluster* karena data yang lebih terperinci (beberapa dibelakang koma) membutuhkan ketelitian tinggi untuk memilih anggota *cluster*.
3. Setiap dilakukan beberapa kali perhitungan pada parameter genetika yang sama, hasilnya bisa berubah-ubah karena setiap kali perhitungan, proses seleksi, crossover dan mutasi yang terjadi juga berbeda. Namun pada kombinasi Pc 0,3 dan Pm 0,9 perubahan yang terjadi sangat kecil dan relative stabil.
4. Jika ditinjau dari waktu komputasi, metode GFS masih relative membutuhkan waktu yang lebih lama dibanding menggunakan FCM karena untuk menemukan individu terbaik menggunakan proses evolusi (seleksi, crossover, mutasi) yang membutuhkan waktu lebih besar daripada FCM yang hanya meminimumkan jarak dengan menggunakan iterasi. Namun dengan perkembangan prosesor yang semakin baik kemungkinan hal tersebut dapat diatasi.
5. Semakin besar ukuran populasi dan jumlah iterasi maka hasil yang didapatkan akan semakin optimal dan stabil, stabil dalam artian nilai objektif relatif sama meskipun dilakukan beberapa kali perhitungan. karena semakin banyak jumlah generasi maka memberikan eksplorasi terhadap ruang pencarian yang lebih besar walaupun harus dibayar dengan waktu yang lebih lama.

5. SARAN

Berdasarkan pada pengujian yang telah dilakukan pada sistem yang dibuat, masih banyak kekurangan dan kelemahan sehingga perlu dikembangkan lagi agar kinerjanya lebih baik, oleh karena itu saran yang diberikan adalah sebagai berikut:

1. Metode GFS ini lebih baik digunakan untuk studi kasus yang membutuhkan data yang lebih presisi, misalnya pengukuran yang membutuhkan tingkat ketelitian yang tinggi.
2. Pada penelitian lebih lanjut diharapkan melakukan riset algoritma untuk mengoptimalkan nilai parameter Pc dan Pm sehingga didapatkan nilai Pc dan Pm yang terbaik yang dapat digunakan pada berbagai jenis data yang berbeda.

DAFTAR PUSTAKA

[1] Cox, E. 2005. *Fuzzy Modeling and Genetic Algorithm for Data Mining and Exploration*, Morgan Kaufmann Publisher, San Francisco.

[2] Klir G. J., Yuan B., 1995. *Fuzzy Sets and Fuzzy Logic, Theory and Application*. Prentice Hall International, Inc.

[3] Widyastuti N., Hamzah A., 2007, Penggunaan Algoritma Genetika Dalam Peningkatan Kinerja *Fuzzy Clustering* untuk Pengenalan Pola, *Seminar Penggunaan Algoritma Genetika*, Yogyakarta

[4] Gen, M. and Cheng, R., 2000, *Genetic Algorithms and Engineering Optimization*, John Wiley & Sons, New York.

Kumpulan jurnaL Ilmu Komputer (KLIK)
Volume 01, No.01 September 2014
ISSN: 2406-7857

IMPLEMENTASI METODE FUZZY C-MEANS PADA SISTEM CLUSTERINGDATA VARIETAS PADI

Nurjanah[1], Andi Farmadi[2], Fatma Indriani[2]
[1,2,3] Prodi Ilmu Komputer Fakultas MIPA Universitas Lambung Mangkurat

Jl. A. Yani Km 36 Banjarbaru, Kalimantan Selatan
Email: nurjanah_abs@yahoo.co.id

Abstract

Mutations with gamma rays conducted on five local rice varieties tidal South Kalimantan produce a lot of data availability. In order for these data not only become a graveyard of useless data required a method that could be used to probe the hidden information from the data. The method known as data mining. Data mining is a technique to gain knowledge from the data by looking for certain patterns or rules of a number of large amounts of data. One method of data mining is clustering, where clustering is usually used to group objects that are similar in the same class or segment. By utilizing the data of local rice varieties tidal South Kalimantan mutated by gamma rays, data mining process is done by grouping the data based on the harvest age, productive tillers, and weight of 1000 seeds into 4 groups using fuzzy c-means algorithm. From that cluster information, carried ranking using the Simple Additive Weighting method and acquired knowledge about improved varieties by harvest age, productive tillers, and a weight of 1000 is kuatek with a dose of 30 krad.

Keywords *: Data Mining, Cluster, Fuzzy C-Means, Local Rice Varieties*

Abstrak

Mutasi dengan sinar gamma yang dilakukan terhadap lima varietas padi lokal pasang surut kalimantan selatan menghasilkan tersedianya banyak data. Agar data-data tersebut tidak hanya menjadi kuburan data yang tidak berguna dibutuhkan sebuah metode yang bisa digunakan untuk menggali informasi–informasi tersembunyi dari data tersebut. Metode tersebut dikenal dengan data mining. Data mining merupakan suatu teknik untuk menggali pengetahuan dari data dengan mencari pola atau aturan tertentu dari sejumlah data dalam jumlah besar. Salah satu metode data mining adalah klastering, dimana klastering biasanya digunakan untuk mengelompokan objek-objek yang memiliki kemiripan dalam kelas atau segmen yang sama. Dengan memanfaatkan data varietas padi hasil mutasi dengan sinar gamma dilakukan proses penggalian data dengan cara mengelompokkan data umur panen, anakan produktif, dan bobot 1000 menjadi 4 klaster menggunakan algoritma fuzzy c-means. Dari informasi 4 klaster tersebut, dilakukan perangkingan klaster menggunakan metode Simple Additive Weighting dan diperoleh pengetahuan tentang varietas unggul berdasarkan umur panen, anakan produktif, dan bobot 1000 yaitu kuatek dengan dosis 30 krad.

Kata kunci *: Data Mining, Klaster, Fuzzy C-Means, Varietas Padi Lokal*

1. PENDAHULUAN

Varietas padi lokal Kalimantan Selatan masih menjadi pilihan para petani untuk ditanam, yang dalam penanamannya tidak terlalu bergantung pada input pertanian. Selain itu, varietas padi lokal juga bersifat adaptif dengan lingkungan Kalimantan Selatan yang didominasi oleh lahan rawa, meskipun berumur panjang dan produktivitas relatif rendah daripada varietas padi unggul. Akan tetapi sifat "pera" nasi yang dihasilkan membuat harga jual varietas padi lokal lebih tinggi dibandingkan dengan varietas padi unggul [3].

Perjuangan dalam peningkatan produksi pangan untuk memenuhi kebutuhan masyarakat Kalimantan Selatan merupakan tantangan bagi masyarakat pertanian yang dapat melahirkan peluang berupa pengembangan teknologi alternatif untuk menerobos kebuntuan dan stagnasi produksi beras lokal. Alternatif itu dapat diarahkan pada pengembangan varietas lokal yang dapat beradaptasi pada lingkungan produksi yang buruk dengan produktivitas yang tinggi dan umur panen yang pendek. Peningkatan keragaman populasi dapat dilakukan salah satunya melalui induksi mutasi secara fisik dengan iradiasi sinar gamma.

Salah satu upaya untuk meningkatkan produksi pangan varietas padi lokal pasang surut kalimantan selatan yaitu dengan cara meningkatkan keragaman populasi, dilakukan radiasi dengan sinar gamma terhadap lima varietas padi. Dari upaya tersebut dihasilkan banyak data salah satunya yaitu data varietas padi hasil mutasi dengan sinar gammamaka dibutuhkan sebuah metode yang bisa digunakan untuk menggali informasi–informasi tersembunyi dari data tersebut. Metode tersebut dikenal dengan *data mining*.

Data mining adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar. Salah satu tugas dalam data mining adalah klastering. Tujuan utama dari klastering adalah pengelompokan sejumlah data/obyek ke dalam klaster sehingga dalam setiap klaster akan berisi data yang semirip mungkin [1].

Penggalian suatu data menjadi pengetahuan adalah kegiatan yang sangat berguna, karena pengetahuan tersebut bisa digunakan untuk memecahkan suatu masalah yang ada. *Data mining* merupakan salah satu teknik untuk menggali pengetahuan dari data. Agar data-data tersebut tidak menjadi kuburan data yang tidak bergunamaka dibuat sistem yang mampu menggali data varietas padi yang dimutasi dengan sinar gamma yang dapat mengorganisir data varietas padi yang dimutasi dengan sinar gamma dalam suatu database serta dapat menghasilkan pengelompokan sejumlah data/obyek ke dalam klaster yang memberikan pengetahuan mengenai varietas unggulmenggunakan metode *Fuzzy C-Means* (FCM).

2. METODOLOGI PENELITIAN

2.1. Analisa Data

Pada tahap ini akan dilakukan analisis data dengan KDD (*knowledge Discovery in Databases*). Keseluruhan proses KDD dari mulai memasukan data mengalami 3 tahap secara global sebelum menjadi hasil yang berupa informasi, yaitu:

(a) *Preprocessing*

Pada tahap ini struktur basis data akan dipersiapkan sehingga mempermudah proses mining. Proses *preprocessing* ini mencakup tiga hal utama yaitu :

1) seleksi data
   untuk memilih data yang akan digunakan dalam proses data mining. Dalam penelitian ini dipilih 3 data varietas padi yaitu umur panen, anakan produktif, dan bobot 1000 biji;
2) *Cleaning*
   Memastikan kualitas data yang telah dipilih pada tahap seleksi data, pada tahap ini masalah yang harus dihadapi adalah noisy data dan missing values. Proses cleaning data perlu dilakukan agar data bersih dari duplikasi data, data yang tidak konsisten, atau kesalahan cetak. Sehingga data yang telah melewati proses ini siap untuk di proses di data mining. Pada penelitian ini proses pembersihan data hanya dilakukan pada data yang missing value;
3) integrasi data
   Setelah melakukan pembersihan data yang menyeluruh, maka akan dilakukan tahapan selanjutnya yaitu integrasi data, karena data lebih dari satu tabel maka dilakukan tahap penggabungan data dari berbagai tabel. Kemudian data padi diberi kode agar tidak ada data yang memiliki nama yang sama. Kode padi terdiri dari nomor padi, baris padi, dosis radiasi, dan varietas padi.

(b) Data Mining

Proses data mining yang dilakukan pada penelitian ini adalah menerapkan teknik *clustering* dengan algoritma *Fuzzy C-Means* untuk mendapatkan karakteristik dari data. *Cluster* yang digunakan pada penelitian ini sebanyak empat klaster.

Konsep dasar FCM (*Fuzzy C-Means*) pertama kali adalah menentukan pusat *cluster* yang akan menandai lokasi rata-rata untuk tiap *cluster*. Pada kondisi awal, pusat *cluster* ini masih belum akurat. Tiap-tiap titik data memiliki derajat keanggotan untuk tiap-tiap *cluster*. Dengan cara memperbaiki pusat *cluster* dan derajat keanggotaan tiap-tiap titik data secara berulang, maka akan dapat dilihat bahwa pusat *cluster* akan bergerak menuju lokasi yang tepat. Perulangan ini didasarkan pada derajat keanggotaan yang menggambarkan jarak dari titik data yang diberikan ke pusat *cluster* yang terbobot oleh derajat keanggotaan titik data tersebut [6].

Pada perhitungan menggunakan FCM dengan menggunakan data yang sama tetapi diolah dengan jumlah *cluster* yang berbeda, maka hasil pengelompokannya akan sedikit berbeda, karena data tidak diolah dengan satu variabel saja tetapi dengan semua variabel. Perbedaan hasil pengelompokan itu dikarenakan data pada kelompok tertentu kemungkinan akan berpindah pada kelompok lain jika diolah dengan jumlah *cluster* yang berbeda, ini menunjukkan bahwa sistem aplikasi sudah berjalan dengan benar [6]

Output dari FCM bukan merupakan fuzzy inference sistem, namun merupakan derajat pusat *cluster* dan beberapa derajat keanggotaan untuk

tiap-tiap titik data. Algoritma *Fuzzy C-Means* disusun dengan langkah sebagai [6]:

(1) Input data
input data yang akan di*cluster*yaitu data varietas padi (X) berupa matrik berukuran n × m (n = jumlah data, m = atribut setiap data). X_{ij} = data ke- i (i = 1, 2, …, n), atribut ke- j (j = 1, 2, …, m).

(2) Batasan
(a) Jumlah *cluster* = c = 4
(b) Pangkat = w = 2
(c) Maksimum iterasi = Maxit = 100
(d) Error terkecil yang diharapkan = ξ = 10^{-3}
(e) Fungsi obyektif awal = Po = 0
(f) Iterasi awal = t = 1

(3) Membangkitkan bilangan random μ_{ik}, i = 1, 2, …, n; k = 1, 2, …,c; sebagai elemen-elemen matrik partisi awal U, dengan jumlah setiap nilai elemen kolom dalam satu baris adalah 1 (satu)

$$\sum_{i=1}^{c} \mu_{ci} = 1 \qquad \dots (1)$$

(4) Menghitung pusat *cluster* ke- k : V_{kj} dengan k = 1, 2, …, c; dan j = 1, 2, …,m.

$$V_{kj} = \frac{\sum_{i=1}^{n}\left((\mu_{ik})^{w} * X_{ij}\right)}{\sum_{i-1}^{n}(\mu_{ik})^{w}} \qquad \dots (2)$$

(5) Menghitung fungsi obyektif pada iterasi ke- t, P_t :

$$P_t = \sum_{i=1}^{n}\sum_{k=1}^{c}\left(\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right](\mu_{ik})^{w}\right) \qquad \dots (3)$$

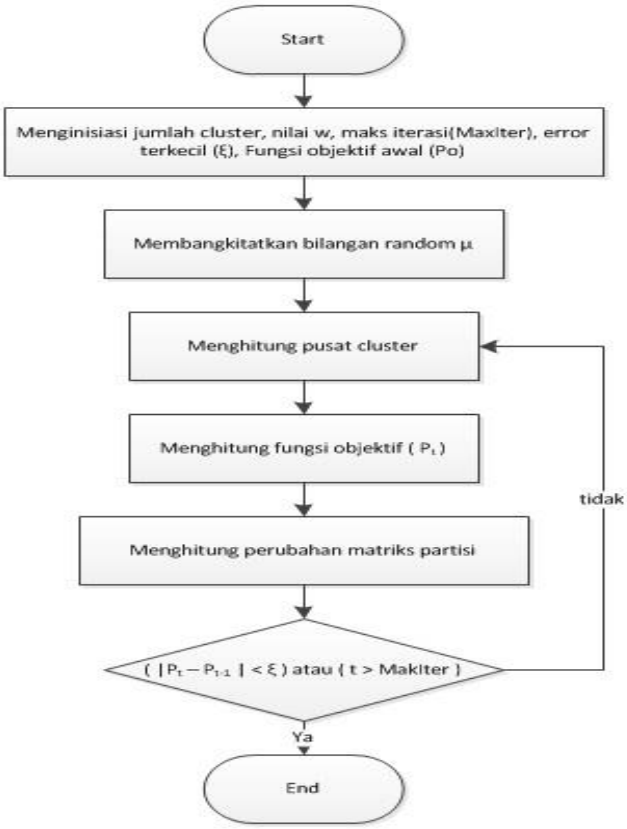
(6) Menghitung perubahan matrik partisi:

$$\mu_{ik} = \frac{\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right]^{\frac{-1}{w-1}}} \dots (4)$$

dengan i = 1, 2, …, n; dan k = 1, 2, …, c

(7) Mengecek kondisi berhenti:
1. Jika : ($|P_t - P_{t-1}| < \xi$) atau (t > MakIter) maka berhenti;
2. Jika tidak : t = t + 1, ulangi langkah ke- d (menghitung V_{kj}).

Bagan alir untuk metode *Fuzzy C-Means* yaitu sebagai berikut :



Gambar 1. Bagan Alir Metode *Fuzzy C-Means*

(c) ***Postprocessing***

Postprocessing bertujuan untuk membantu pengguna dalam memahami informasi, Karena pusat *cluster* masih tidak bisa dienterpretasikan secara langsung, maka diperlukan metode *Simple Additive Weighting* yang bisa melakukan perangkingan nilai pusat *cluster* terakhir tersebut.

Metode SAW (*Simple Additive Weighting*) sering juga dikenal istilah metode penjumlahan terbobot. Konsep dasar metode SAW adalah mencari penjumlahan terbobot dari rating kinerja pada setiap alternatif pada semua atribut. Langkah-langkah metode dalam metode SAW adalah [7] :

1. Membuat matriks keputusan Z berukuran m x n, dimana m = alternatif yang akan dipilih dan n = kriteria .
2. Memberikan nilai x setiap alternatif (i) pada setiap kriteria (j) yang sudah ditentukan, dimana, i=1,2,…m dan j=1,2,…n pada matriks keputusan Z,

$$\mathbf{Z} = \begin{vmatrix} x_{11} & x_{y12} & \dots & x_{1j} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ x_{i1} & x_{i2} & & x_{ij} \end{vmatrix} \dots (5)$$

3. Memberikan nilai bobot preferensi (W) oleh pengambil keputusan untuk masing-masing kriteria yang sudah ditentukan.

$$W = [W_1\ W_2 W_3 \dots\ W_j] \dots(6)$$

4. Melakukan normalisasi matriks keputusan Z dengan cara menghitung nilai rating kinerja ternormalisasi (r_{ij}) dari alternatif A_i pada atribut C_j

$$r_{ij} = \begin{cases} \frac{x_{ij}}{\underset{i}{MAX}(x_{ij})} & \text{jika j adalah atribut keuntungan} \\ \frac{\underset{i}{MIN}(x_{ij})}{x_{ij}} & \text{jika j adalah atribut biaya} \end{cases} \dots(7)$$

Dengan ketentuan :

a. Dikatakan atribut keuntungan apabila atribut banyak memberikan keuntungan bagi pengambil keputusan, sedangkan atribut biaya merupakan atribut yang banyak memberikan pengeluaran jika nilainya semakin besar bagi pengambil keputusan.
b. Apabila berupa atribut keuntungan maka nilai (x_{ij}) dari setiap kolom atribut dibagi dengan nilai (MAX x_{ij}) dari tiap kolom, sedangkan untuk atribut biaya, nilai (MIN x_{ij}) dari tiap kolom atribut dibagi dengan nilai (x_{ij}) setiap kolom.

5. Hasil dari nilai rating kinerja ternormalisasi (r_{ij}) membentuk matriks ternormalisasi (N)

$$\mathbf{N} = \begin{vmatrix} r_{11} & r_{y12} & \dots & r_{1j} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ r_{i1} & r_{i2} & & r_{ij} \end{vmatrix} \dots (8)$$

6. Melakukan proses perankingan dengan cara mengalikan matriks ternormalisasi (N) dengan nilai bobot preferensi (W).
7. Menentukan nilai preferensi untuk setiap alternatif (V_i) dengan cara menjumlahkan hasil kali antara matriks ternormalisasi (N) dengan nilai bobot preferensi (W).

$$V_i = \sum_{j=1}^{n} w_j r_{ij} \dots (9)$$

Nilai Vi yang lebih besar mengindikasikan bahwa alternatif Ai merupakan alternatif terbaik.

2.3. Implementasi

Implementasi dilakukan dengan metode Fuzzy C-Means (FCM) pada sistem clustering data varietas padilokal pasang surut kalimantan selatan yang dimutasi dengan sinar gamma menggunakan delphi XE2

2.4. Pengujian

Untuk memastikan apakah perhitungan dengan metode fuzzy c-means sudah benar dilakukan uji coba perhitungan secara manual meggunakan MS. Excel dan Matlab.Pengujian dilakukan untuk menghindari kesalahan dari sistem yang dibuat. Apabila terjadi kesalahan, maka sistem akan diperbaiki kembali sampai hasil proses sesuai dengan apa yang diharapkan.

3. HASIL DAN PEMBAHASAN

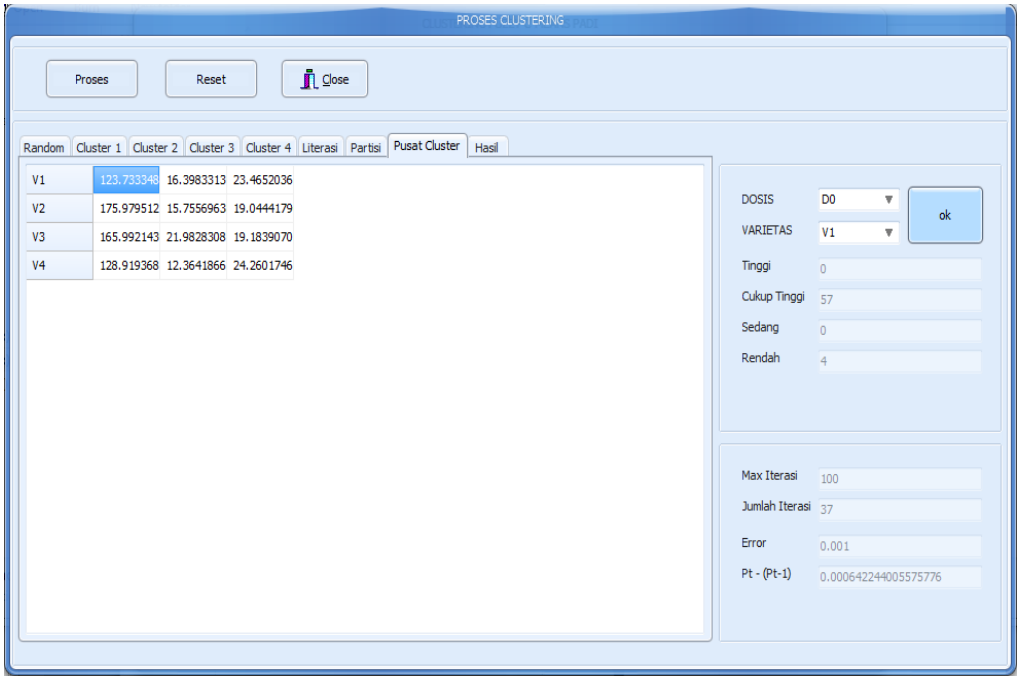
Dariproses clustering Fuzzy C-Means dengan batasan jumlah cluster sebanyak 4 cluster, pangkat 2, masimal iterasi 100, error 0,001 proses berhenti pada iterasi ke-37 dengan error 0,000642 dan diperoleh pusat cluster iterasi terakhir yaitu pada Tabel 1 atau pada Gambar 2 dengan menggunakan rumus dibawah ini (persamaan 2):

$$V_{kj} = \frac{\sum_{i=1}^{2420}\left((\mu_{ik})^2 * X_{ij}\right)}{\sum_{i=1}^{2420}(\mu_{ik})^2}$$

Keterangan :
V_{kj}: pusat cluster dengan k=1,2,...,c dan j = 1,2,...,n
X_{ij} : data ke-i (i=1,2,...m), atribut ke-j (j=1,2,...n)
μ_{ik} : bilangan acak dengan i=1,2,...m dan k=1,2,...c

Tabel 1. Pusat Cluster Delphi

	Umur Panen	Anakan Produktif	Bobot 1000
Klaster 1	165.992143	21.9828308	19.1839070
Klaster 2	123.733348	16.3983313	23.4652036
Klaster 3	128.919368	12.3641866	24.2601746
Klaster 4	175.979512	15.7556963	19.0444179



Gambar 2. Form Proses Fuzzy C-Means

Dari Tabel 1 dapat dilihat bahwa klaster 1 memiliki umur panen terendah ketiga, anakan produktif terbanyak pertama, dan bobot 1000 terberat ketiga. Klaster 2 memiliki umur panen terendah pertama, anakan produktif terbanyak kedua, dan bobot 1000 terberat kedua. Klaster 3 memiliki umur panen terendah kedua, anakan produktif terbanyak keempat, dan bobot 1000 terberat pertama.

Klaster 4 memiliki umur panen terendah keempat, anakan produktif terbanyak ketiga, dan bobot 1000 terberat keempat.

Dilihat dari umur panen klaster 2 lebih unggul karena memiliki umur panen terendah dibandingkan klaster 1, klaster 3, dan klaster 4. Untuk anakan produktif klaster 1 lebih unggul karena memiliki anakan produktif terbanyak. Dan untuk bobot 1000 klaster 3 lebih unggul karena memiliki bobot seribu terberat apabila dibandingkan dengan klaster lainnya.

Dari hasil pusat klaster (Tabel 1) tersebut dilakukan perangkingan dengan metode *Simple Additive Weighting* (SAW) untuk melihat klaster mana yang memiliki umur panen, anakan produktif dan bobot 1000 terbaik. Perhitungan nilai prefrensi dilakukan menggunakan rumus dibawah ini (persamaam 9) :

$$V_i = \sum_{j=1}^{n} w_j r_{ij}$$

Keterangan :
V_i : nilaipreferensiuntuksetiapalternatif
w_j : nilai bobot preferensi
r_{ij} : matriks ternormalisasi

Berikut tabel nilai prefrensi dan tabel hasil pengelompokan per varietas :

Tabel 2. Nilai Prefensi Klaster

Klaster	Nila Prefensi	Kategori
Klaster 1	2.535898966	Cukup Tinggi
Klaster 2	2.712760587	Tinggi
Klaster 3	2.53314209	Sedang
Klaster 4	2.209208036	Rendah

Tabel 3. Hasil Pengelompokan Varietas Siam Harli

Kategori	Varietas Siam Harli (V1)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	-	-	-	46
Cukup Tinggi	57	77	32	-
Sedang	-	-	-	44
Rendah	4	2	3	-

Tabel 4. Hasil Pengelompokan Varietas Siam Unus

Kategori	Varietas Siam Unus (V2)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	-	1	-	-
Cukup Tinggi	48	205	48	86
Sedang	-	1	4	-
Rendah	4	-	5	8

Tabel 5. Hasil Pengelompokan Varietas Kuatek

Kategori	Varietas Kuatek (V3)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	-	-	-	371
Cukup Tinggi	7	152	18	176
Sedang	-	-	-	-
Rendah	41	5	132	-

Tabel 6. Hasil Pengelompokan Varietas Siam Sebelas

Kategori	Varietas Siam Sebelas (V4)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	5	-	-	25
Cukup Tinggi	1	28	9	-
Sedang	1	-	-	22
Rendah	25	158	78	-

Tabel 7. Hasil Pengelompokan Varietas Siam Gumpal

Kategori	Varietas Siam Gumpal (V5)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	-	-	-	13
Cukup Tinggi	-	-	-	-
Sedang	-	-	-	116
Rendah	54	173	105	-

Pengetahuan yang didapat dari hasil penambangan data/klastering ini berupa varietas unggul yaitu varietas yang masuk kategori tinggi. Varietas yang masuk kategori tinggi terdiri atas rata-rata umur panen sebesar 123,7665737; rata-rata anakan produktif sebesar 16,2834365; dan rata-rata bobot 1000 sebesar 23,46984343 dengan nilai preferensi sebesar 2,712760587. Diantara kelima varietas tersebut yang paling banyak masuk kategori tinggi yaitu varietas kuatek dengan dosis 30 krad yakni sebanyak 371 data padi atau sebesar 67,82 % data padi masuk kategori tinggi

Tabel 8. Presentase Data Padi Yang Termasuk Kategori Tinggi

Varietas	Dosis			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Siam Harli (V1)	-	-	-	51,11%
Siam Unus (V2)	-	0,93%	-	-
Kuatek (V3)	-	-	-	67,82%
Siam Sebelas (V4)	15,61%	-	-	53,19%
Siam Gumpal (V5)	-	-	-	10,08%

4. KESIMPULAN

Dari hasil penelitian dan pengamatan dari sistem yang telah dibuat, maka dapat diperoleh kesimpulan sebagai berikut :

1. Metode Fuzzy C-Means berhasil diimplemantasikan pada clustering data varietas padi lokal pasang surut kalimantan selatan yang dimutasi dengan sinar gamma.
2. Dilihat dari umur panen klaster 2 lebih unggul karena memiliki umur panen terendah, untuk anakan produktif klaster 1 lebih unggul karena memiliki anakan produktif terbanyak, dan untuk bobot 1000 klaster 3 lebih unggul karena memiliki bobot seribu terberat apabila dibandingkan dengan klaster lainnya.
3. Pengetahuan yang didapat mengenai varietas unggul berdasarkan umur panen, anakan produktif dan bobot 1000 terbaik menggunakan SAW memiliki nilai preferensi sebesar 23,46984343 yaitu kategori tinggi.
4. Diantara kelima varietas yang paling banyak masuk kategori tinggi yaitu varietas kuatek dengan dosis 30 krad yakni sebanyak 371 data padi atau sebesar 67,82 % data padi masuk kategori tinggi.

DAFTAR PUSTAKA

[1]. Budi Santosa, "**Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis***",* Graha Ilmu, Yogyakarta, 2007.

[2]. Iko Pramudiono. "***Pengantar Data Mining : Menambang Permata Pengetahuan di Gunung Data***" 2003.

[3]. Noor, M., Y. Rina dan Noorginayuwati, "**Persepsi petani tentang lahan gambut dan pengelolaannya",** Laporan Penelitian, Balai Penelitian Pertanian Lahan Rawa, Banjarbaru, 2007.

[4]. Nurjanah, "**Clustering Varietas Padi Lokal Pasang Surut Kalimantan Selatan Yang Dimutasi Dengan Sinar Gamma Menggunakan Metode Fuzzy C-Means**" Naskah Skripsi Universitas Lambung Mangkurat, Banjarbaru, 2014.

[5]. Nuqson Masykur Huda, "***Aplikasi Data Mining Untuk Menampilkan Informasi Tingkat Kelulusan Mahasiswa (Studi Kasus Di Fakultas MIPA Universitas Diponegoro)",*** Naskah Skripsi Universitas Diponegoro, Semarang, 2010.

[6]. Sri Kusumadewi and Hari Purnomo, "**Aplikasi Logika Fuzzy Untuk Pendukung Keputusan***",* Yogyakarta, Graha Ilmu, 2010.

[7]. Wibowo S, Henry., Amalia, Riska., Fadlun M, Andi., Arivanty, Kurnia, "**Sistem Pendukung Keputusan Untuk MenentukanPenerima Beasiswa Bank BRI Menggunakan FMADM (Studi Kasus: Mahasiswa Fakultas Teknologi Industri Unversitas Islam Indonesia)***",* Seminar Nasional Aplikasi Teknologi Informasi, Yogyakarta, 2008.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 4, Issue 1 , January 2017

Microscopic Image Segmentation Using Fuzzy C Means For Leukemia Diagnosis

Dr.T.Karthikeyan, N.Poornima

Associate Professor, Department of Computer Science, PSG college of Arts and Science, Coimbatore, India.

Research Scholar, Department of Computer Science, PSG college of Arts and Science, Coimbatore, India.

ABSTRACT: Leukemias are cancers of the blood-forming tissues. White blood cells may be produced in excessive amounts and are unable to work properly which weakens the immune system. Careful microscopic examination of blood smear or bone marrow aspirate is the only way to effective diagnosis of leukemia. The need for automation of leukemia detection arises since the above specific tests are time consuming and costly. Morphological analysis of blood slides are influenced by factors such as hematologists experience and tiredness, resulting in non standardized reports. A low cost and efficient solution is to use image analysis for quantitative examination of stained blood microscopic images for leukemia detection. In this paper, Fuzzy c means is compared with k means for image segmentation, in which Fuzzy c means gives higher accuracy than k means, Gabor Texture Extraction method is used to extract color features from images and finally extracted features are used for classification. Fuzzy c means gives 90% accuracy whereas k means gives 83% accuracy. Support Vector Machine is used for classification. The whole work has been developed using MATLAB 7 environment.

KEYWORDS: Leukemia, microscopic images, k means, Fuzzy c means, Gabor texture extraction, SVM

I.INTRODUCTION

Leukemia is a type of cancer that affects the bone marrow causing increased production of white blood cells which influx the blood stream [1]. Leukemias are cancers of the blood-forming tissues. White blood cells may be produced in excessive amounts and are unable to work properly which weakens the immune system. The blood is made up of fluid called plasma and three types of cells and each type has special functions. Leukocytes which play a major role in diagonosis of different diseases [13]. White blood cells help the body fight infections and other diseases. Red blood cells carry oxygen from the lungs to the body's tissues and take carbon dioxide from the tissues back to the lungs. The red blood cells give blood its color. Platelets help form blood clots that control bleeding. Blood cells are formed in the bone marrow, the soft, spongy center of bones. New blood cells are called blasts. Some blasts stay in the marrow to mature. Some travel to other parts of the body to mature. Normally, blood cells are produced in an orderly, controlled way, as the body needs them. This process helps keep us healthy. In a leukemic person, the bone marrow produces abnormal white blood cells that are called leukemic cells and leukemic blast cells [19]. When leukemia develops, the body produces large numbers of abnormal blood cells. In most types of leukemia, the abnormal cells are white blood cells. The leukemia cells usually look different from normal blood cells, and they do not function properly.

A.TYPES OF LEUKEMIA

There are several types of leukemia. They are grouped in two ways. One way is by how quickly the disease develops and gets worse. The other way is by the type of blood cell that is affected. Leukemia is either acute or chronic. In acute leukemia, the abnormal blood cells remain very immature and cannot carry out their normal functions. The number of blasts increases rapidly, and the disease gets worse quickly. In chronic leukemia, some blast cells are present, but in general, these cells are more mature and can carry out some of their normal functions. Also, the number of blasts increases less rapidly than in acute leukemia. As a result, chronic leukemia gets worse gradually.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 4, Issue 1 , January 2017

Leukemia can arise in either of the two main types of white blood cells — lymphoid cells or myeloid cells. When leukemia affects lymphoid cells, it is called lymphocytic leukemia. When myeloid cells are affected, the disease is called myeloid or myelogenous leukemia.

The disease appears in one of four major forms [19]:

- **Acute lymphocytic leukemia** (ALL) is the most common type of leukemia in young children. This disease also affects adults, especially above 65 years and older.
- **Acute myeloid leukemia** (AML) occurs in both adults and children. This type of leukemia is sometimes called acute nonlymphocytic leukemia (ANLL) .
- **Chronic lymphocytic leukemia** (CLL) most often affects adults over the age of 55. It sometimes occurs in younger adults, but it almost never affects children.
- **Chronic myeloid leukemia** (CML) occurs mainly in adults. A very small number of children also develop this disease.

Leukemia can develop at any point in cell differentiation. The illustration below shows the development of the formed elements of the blood.

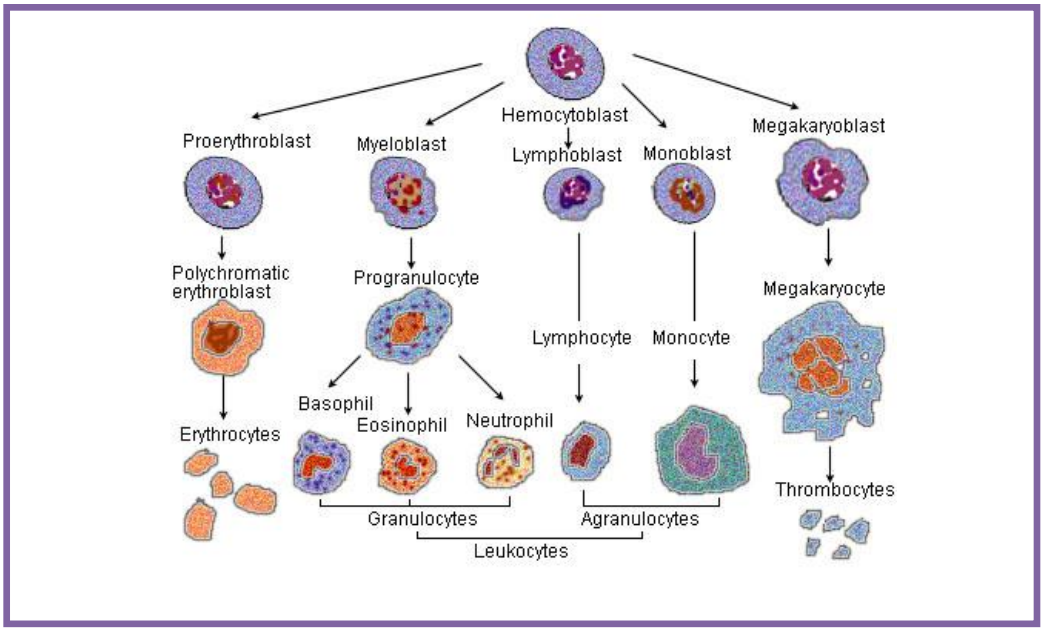


Fig 1: development of the formed elements of the blood

B. MAIN COMPONENTS OF IMAGE PROCESSING

Step1: Image Acquisition - capturing images.
Step 2. Image Enhancement - improvement of quality of images.
Step 3. Image Representation - image can be represented mathematically, graphically, and statistically.
Step 4. Image Transformation - transform the input image from one domain into another e.g an image in spatial domain can be converted into frequency domain by using Fourier transform.
Step 5. Image Restoration - analysis and modeling of different types of noise mixed in images.
Step 6. Color image processing - Various color spaces and formats are converted in it.
Step 7. Image compression - reduce the size of image or reduce redundancy without any significant change in the inherent content of the image.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 4, Issue 1 , January 2017

Step 8. Morphological image processing - these operations are generally used with image segmentation.
Step 9. Image segmentation, representation and description - the selected region of interests can be extracted and various boundaries, edges and other similar information could be obtained.
Step 10. Object recognition - the pattern recognition and matching [1]

II.LITERATURE REVIEW

S.Jagadeesh *et.al* proposed segmentation of bone marrow image aspirate using watershed algorithm[16]. T. Markiewicz *et.al* use Support Vector Machine (SVM) classifier[17] and exploit features in blood cell images to classify leukemic cells[3]. Khot S.T *et.al* used Support Vector Machine classifier to detect leukemic cells.[4]. Niranjan Chatap., *et.al* proposed Support Vector Machine and kNN for detecting leukemic cells[5]. Ms. Minal D. Joshi *et.al* proposed a method to detect acute leukemia. By histogram equalization method ,[7] enhance contrast of the grayscale image. K means is used for blood smear image segmentation and features are extracted and extracted features are applied to SVM classifier [13].

III.STEPS FOR PROPOSED METHODOLOGY

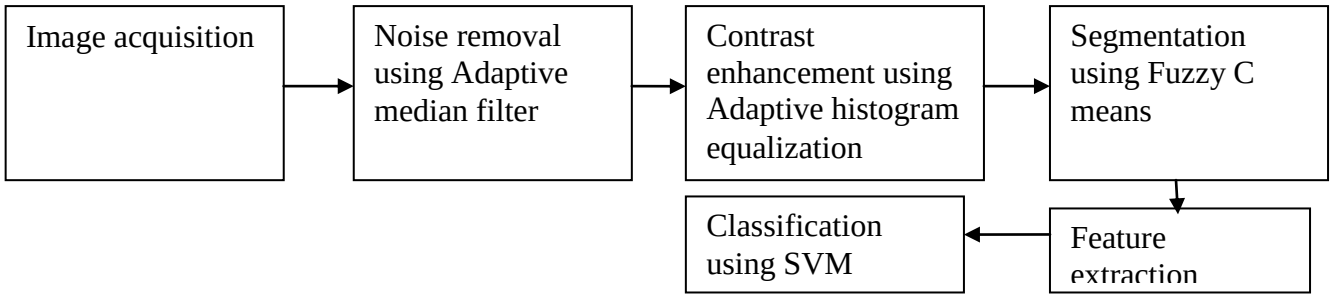


Fig 2: Proposed Framework for detecting leukemia cells.

A. Image Acquisition:

For detecting Leukemia, microscopic images are acquired. These images are taken from net. This stage includes image pre processing

B. Image Pre processing:

Noise is the undesirable effects produced in the image. During image acquisition or transmission, several factors are responsible for introducing noise in the image. This stage includes noise removal and contrast enhancement of image. Adaptive median Filter is used for noise removal and Adaptive Histogram Equalization is used for contrast enhancement.

C. Image segmentation:

The next stage deals with image segmentation. Segmentation partitions an input image into its constituent parts or objects[1]. On one hand rugged segmentation procedure brings the process a long way towards successful solution of an imaging problem. Goal of image segmentation is to extract information from input image [13].In this paper, segmentation is done using Fuzzy c means clustering algorithm.

Image Segmentation stage aims to separate and detect white blood cells (WBC) and red blood cell (RBC). The first stage of image segmentation is to detect WBC. Image segmentation is the process of dividing the image into parts. The process of segmentation involves the severance of areas that hold similar characteristics. In the field of image processing, color plays two important roles. First, color can easily be used for simplifying the process of object identification and object extraction from a scene. Second, human can recognize thousands of color shades and intensities. Two algorithms are compared in this paper such as k means and Fuzzy c means clustering algorithm.

K means clustering:

It is a partitioning algorithm and it use an optimization function to partition the cells into successive clusters [2]. k-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other [17]. The next step is to take each point belonging to a given data set and associate it to the nearest

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 4, Issue 1 , January 2017

center. When no point is pending, the first step is completed and an early group age is done. At this point re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more

Fuzzy C means for image segmentation:

FCM generalizes the hard c-means algorithm to allow a point to partially belong to multiple clusters. It produces a soft partition for a given dataset [8]. The FCM clustering algorithm is a soft segmentation method that has been widely used for microscopic image segmentation. FCM clustering approach could retain information from the original image than hard segmentation methods known as K-Means clustering. The clustering methods are considerable benefits especially for images which contain huge sets of pixel data.

In particular, the FCM algorithm, assign pixels to fuzzy clusters without labels. Unlike the K-Means clustering which force pixels to belong exclusively to one class, Fuzzy C-means allows pixels to belong to more than one cluster based on degrees of membership. Thus, points on the edge of a cluster may be in the cluster to a lesser degree than points in the centre of cluster. However, its main drawbacks include its computational complexity and the fact that the performance degrades significantly with increased noise.FCM is a clustering method which allows single data belong to more than one clusters.

D. Feature Extraction:

To detect leukemia, features like geometrical, texture, color, and statistical are extracted from segmented image. Gabor texture extraction method is used. In color feature, RGB color spaces will be transformed.

E. Image Classification:

Classification of cells is more important in medical image. Based on classification results we can identify whether the human is affected with leukemia or not. It is an optimization theory based classifier that finds a maximal margin surface, which separates the data into distinct classes [2]. Another name for SVM is Kernel Machines (as nonlinear SVM uses Kernel mapping). Kernel method is a popular technique to transform nonlinearly separable data to linearly separable form. Proposed framework built a system to detect leukemia cells of images of bone marrow. Using Support Vector Machine (SVM) classifier and blood cell images features that are related to geometry, texture, and statistical analysis, the system was built. The pressure is on selection and generation of features for getting out the best recognition. Textural parameters such as entropy, contrast, mean value and angular second momentum have been used .Geometrical parameters are compactness, perimeter, concavity points and symmetry radius, area and filled area. For statistical analysis parameters are mean value and for nucleus standard deviation and for gradient matrix cytoplasm, mean and standard deviation are considered. Classifier is trained with features of sample images. When the features of input image are fed into classifier, the classifier classifies output.

IV.RESULTS AND DISCUSSION

The proposed technique is applied to 19 images of blood smear. Images are taken from Google. The images are in JPG format.The results obtained are shown below. In the result, input image is added with salt and pepper noise. And noisy image is preprocessed for noise removal. Then image is used for contrast enhancement. Then the image is segmented by k means and Fuzzy c means and both are compared. Then next color and shape features are extracted from segmented image. Then classification is done.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 4, Issue 1 , January 2017

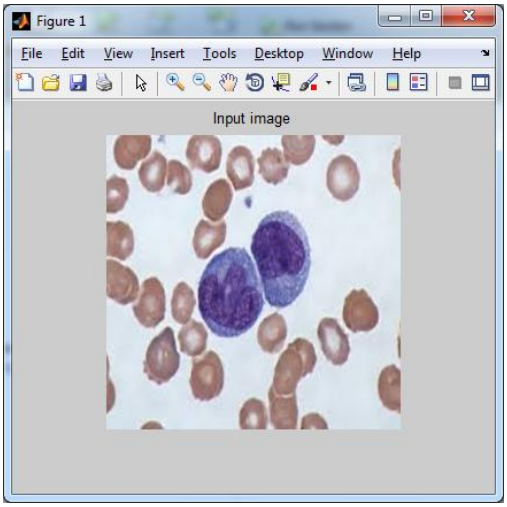


Fig 1:Input image

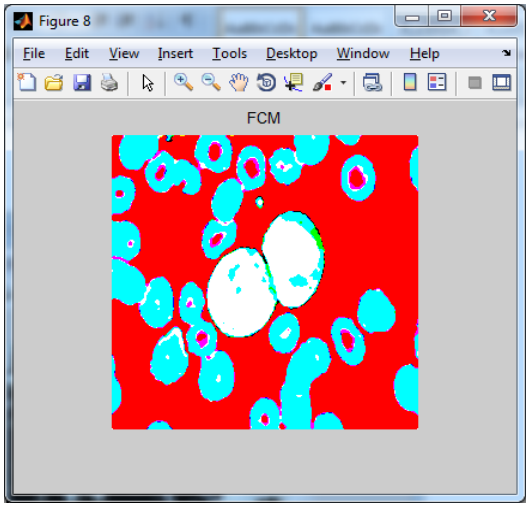


Fig 2: Fuzzy c means segmentation

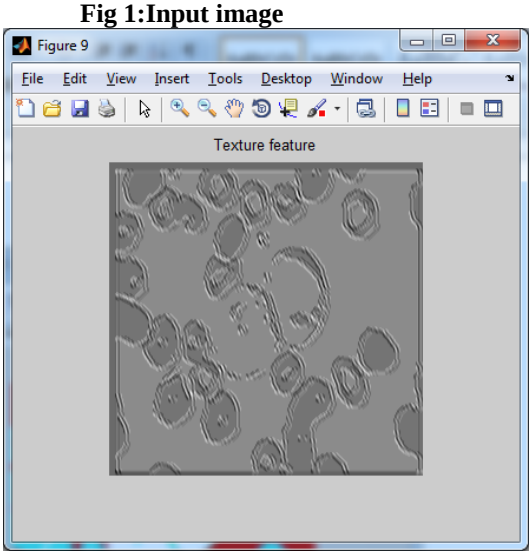


Fig 11: Feature extracted image

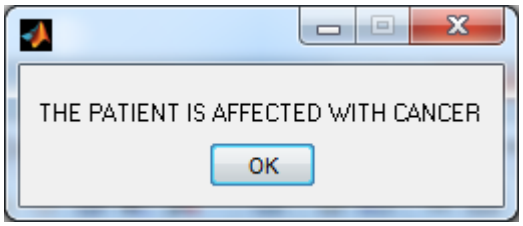


Fig 12: Leukemia Classification

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 4, Issue 1 , January 2017

PERFORMANCE CHART:

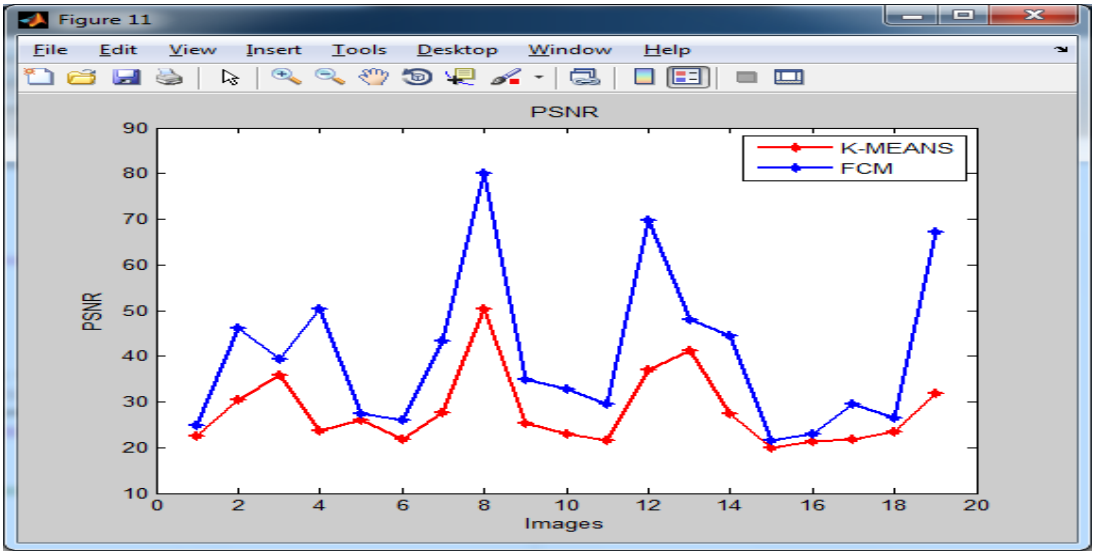


Fig 13: Performance chart for k means and fuzzy c means

V.CONCLUSION

Leukemia is a severe disease which affects blood cells of bone marrow. It leads death if left untreated. A low cost and efficient solution is to use image analysis for quantitative examination of stained blood microscopic images for leukemia detection. In this paper, Fuzzy c means is compared with k means for image segmentation, in which Fuzzy c means gives higher accuracy than k means, Gabor Texture Extraction method is used to extract color features from images and finally extracted features are used for classification. Fuzzy c means gives 90% accuracy whereas k means gives 83% accuracy. Based on segmentation result, color and shape features are extracted. Then the extracted values are fed into SVM classifier. Finally the classifier classifies the segmented image based on its features and produces the results.

VI.REFERENCES

[1] N.Poornima, T.Karthikeyan .," A survey on Image Processing Techniques used for Detection of leukemic Cells" International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 4, April 2016
[2] T. Karthikeyan, B. Ragavan , N.Poornima.," A comparative study of algorithms used for leukemia detection" International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 5, Issue 4, April 2016
[3] T. Markiewicz, S. Osowski, B., Marianska,L., Moszczynski, "Automatic Recognition of the Blood Cells of Myelogenous Leukemia Using SVM", *Proceedings of International Joint Conference on Neural Networks,Montreal* , Canada, July 31 – August 4, 2005,pp. 2496-2501.
[4] Khot S.T, Sneha Bhalekar, Divya Jaggi and Dolly Rani., " An innovative approach in myelogenous leukemia detection using attribute analysis" (IJAEEE) ISSN (Print): 2278-8948, Volume-2, Issue-5, 2013
[5] S.Jagadeesh 1, Dr.E.Nagabhooshanam 2, Dr.S.Venkatachalam., "Image processing based approach to cancer cell prediction in blood samples" International Journal of Technology and Engineering Sciences Vol.1 (1), ISSN: 2320-8007
[6] "Data Mining Methods" written by Rajan Chattamvelli.
[7] Ms. Minal D. Joshi, Prof. Atul H. Karode, Prof. S.R.Suralkar "White Blood Cells Segmentation and Classification to Detect Acute Leukemia" (IJETTCS) Volume 2, Issue 3, May – June 2013.
[8]" Fuzzy Logic: Intelligence, control, and information" written by John Yen and Reza Langari.
[9] "Medical Image Processing concepts and applications" G.R Sinha , Bhagwati Charan Patel.
[10]udel.edu./kschieff/leukemiadiagnosis.html.
[11] Himali P. Vaghela, Hardik Modi, Manoj Pandya, Manoj Pandya.," Leukemia Detection using Digital Image Processing Techniques" *(IJAIS) – ISSN : 2249-0868 Volume 10 – No.1, November 2015*
[12] Mrs. Trupti A. Kulkarni-Joshi, Prof. Dilip S. Bhosale., **"**A Fast Segmentation Scheme for Acute Lymphoblastic Leukemia Detection"(IJAREEIE) Vol. 3, Issue 2, February 2014 ISSN (Print) : 2320 – 3765 ISSN (Online): 2278 – 8875
[13]C.Vidhya, P.Saravana kumar, K.Keerthika, C.Nagalakshmi, B.Medona devi.," Classification of acute lymphoblastic leukemia in blood microscopic images using svm**" (**ICETSH-2015**)** ISSN: 2348 – 8549
[14] Tejashri G. Patil, V. B. Raskar.," Blood microscopic image segmentation & acute leukemia detection"(IJERMT) *ISSN: 2278-9359 (Volume-4, Issue-9)SEP 2015.*
[15] A. Arputha Regina .," Detection of Leukemia with Blood Microscopic Images"(IJIRCCE) ISSN(Online): 2320-9801 ISSN (Print): 2320-9798 Vol.3, Special Issue 3, April 2015

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 4, Issue 1 , January 2017

[16] https://sites.google.com/site/dataclusteringalgorithms/fuzzy-c-means-clustering-algorithm
[17] https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm
[18] Suman Shrestha.," IMAGE DENOISING USING NEW ADAPTIVE BASED MEDIAN FILTER" Signal & Image Processing : An International Journal (SIPIJ) Vol.5, No.4, August 2014
[19]Subhan, Ms. Parminder Kaur., "Significant Analysis of Leukemic Cells Extraction and Detection Using KNN and Hough Transform Algorithm" (IJCST) volume 3 issue 1, jan-feb 2015.

AUTHOR'S BIOGRAPHY

Dr. Thirunavukarasu Karthikeyan received his doctorate in Computer Science from Bharathiyar University in 2009. Presently he is working as an Associate Professor in Computer Science Department of P.S.G. College of Arts and Science, Coimbatore. His research interests are Image Coding, Medical Image Processing, Data Mining and Software Engineering.He has contributed as a program committee member for a number of international conferences. He is the review board member of various reputed journals. He is board of studies member for various autonomous institutions and universities.

Mrs N.Poornima has completed Msc degree at Bharathiar University. Now currently she is doing Mphil inss computer science. Her area of interests are Image Processing and Data Mining.

ANALISIS PERBANDINGAN METODE *FUZZY C-MEANS* DAN *SUBTRACTIVE FUZZY C-MEANS*

Baiq Nurul Haqiqi[1], Robert Kurniawan[2]
[1,2]Jurusan Komputasi Statistik, Sekolah Tinggi Ilmu Statistik (STIS)
Email: [1] qiqi0693@naver.com, [2] robertk@stis.ac.id

Abstract

Fuzzy C-Means (FCM) is one of the most frequently used clustering method. However FCM has some disadvantages such as number of clusters to be prespecified and partition matrix to be randomly initiated which makes clustering result becomes inconsistent. Subtractive Clustering (SC) is an alternative method that can be used when number of clusters are unknown. Moreover, SC produces consistent clustering result. A hybrid method of FCM and SC called Subtractive Fuzzy C-Means (SFCM) is proposed to overcome FCM's disadvantages using SC. Both SFCM and FCM are implemented to cluster generated data and the result of the two methods are compared. The experiment shows that generally SFCM produces better clustering result than FCM based on six validity indices.

Keywords : Clustering, Fuzzy C-Means, Subtractive Clustering, Subractive Fuzzy C-Means

1. Pendahuluan

Pengelompokan (*clustering*) merupakan salah satu teknik yang paling penting dalam *data mining*[8]. Salah satu metode pengelompokan yang paling sering digunakan adalah *Fuzzy C-Means* (FCM). Metode FCM memiliki beberapa kelemahan, antara lain membutuhkan banyaknya kelompok dan matriks keanggotaan kelompok yang ditetapkan sebelumnya[6]. Biasanya, matriks keanggotaan kelompok awal diinisialisasikan secara acak yang menyebabkan metode FCM memiliki masalah inkonsistensi.

Alternatif metode pengelompokan lainnya yang dapat digunakan jika jumlah kelompok tidak diketahui sebelumnya adalah metode *Substractive Clustering* (SC). SC memperoleh hasil yang lebih konsisten dibandingkan dengan FCM[2]. Selain itu, SC memiliki kecepatan yang lebih baik dibandingkan FCM, namun SC memiliki akurasi yang lebih rendah dibandingkan dengan FCM[4].

Untuk menjembatani kekurangan dan kelebihan kedua metode tersebut diusulkan sebuah metode baru yang merupakan penggabungan (*hybrid*) dari keduanya yang disebut *Subtractive Fuzzy C-Means* (SFCM). Metode ini digunakan oleh Liu, Xiao, Wang, Shi, dan Fang[7] dalam penelitiannya dan menyimpulkan bahwa SFCM secara umum memberikan solusi yang lebih baik dibandingkan dengan FCM serta memberikan tingkat kecepatan yang lebih tinggi dalam hal konvergensi fungsi objektif. Kelebihan metode SFCM lainnya ditunjukkan oleh Hossen, Rahman, Sayeed, Samsuddin, dan Rokhani[4] dengan menggunakan indeks validitas *Partition Coefficient* (PC) yang menyimpulkan bahwa SFCM dapat meningkatkan kecepatan, mengurangi jumlah iterasi, menghasilkan partisi data yang lebih stabil, dan lebih akurat.

Penelitian ini bertujuan untuk membandingkan kehandalan dari hasil pengelompokan menggunakan metode FCM dan SFCM, yang berdasarkan studi literatur, bahwa belum pernah ada penelitian lain yang membandingkan antara kedua metode tersebut. Pada penelitian ini digunakan data bangkitan dengan distribusi Uniform (0,1). Hasil pengelompokan terbaik dipilih berdasarkan berdasarkan indeks validitas *Partition*

Coefficent (PC), *Modified Partition Coefficient* (MPC), *Classification Entropy* (CE), *Partition Index* (PI), *Fukuyama Sugeno Index* (FS), dan *Xie Beni Index* (XB).

2. Tinjauan Pustaka

2.1. *Fuzzy C-Means* (FCM)

Metode ini ditemukan pertama kali oleh Dunn pada tahun 1973 kemudian dikembangkan lagi oleh Bezdek pada tahun 1981. Ide dasar dari metode ini mirip dengan metode *K-Means*. FCM didasarkan pada logika *fuzzy*, setiap titik data dimasukkan ke suatu kelompok berdasarkan nilai keanggotaannya pada kelompok tersebut.

Algoritma FCM adalah sebagai berikut[5]:

1) Menentukan banyak kelompok (*c*), *fuzzifier* (*m*), maksimum iterasi (*MaxIter*), perubahan nilai fungsi objektif terkecil yang diharapkan (ε), fungsi objektif awal ($P_0 = 0$), dan iterasi awal ($t = 1$)
2) Membangkitkan bilangan random u_{ik} dengan *i* merupakan banyak data dan *k* merupakan banyak kelompok sebagai elemen-elemen awal matriks keanggotaan awal U.
3) Menghitung pusat kelompok ke-*i* dengan persamaan:

$$p_i = \frac{\sum_{k=1}^{N} (u_{ik})^m x_k}{\sum_{k=1}^{N} (u_{ik})^m} \tag{1}$$

dengan u_{ik} adalah nilai keanggotaan objek ke-*k* dengan pusat kelompok ke-*i*, X_k adalah objek data ke-k, *N* adalah banyaknya objek penelitian, dan *m* adalah *fuzzifier*.

4) Menghitung fungsi objektif pada iterasi ke-*t* dengan persamaan:

$$J(P,U,X,c,m) = \sum_{i=1}^{c} \sum_{k=1}^{N} (u_{ik})^m d_{ik}^2 (x_k, p_i) \tag{2}$$

dengan *c* adalah banyak kelompok yang diinginkan, *N* adalah banyak objek penelitian, u_{ik} adalah nilai keanggotaan objek ke-*k* pada kelompok ke-*i* yang merupakan bagian dari matriks U, *m* adalah *fuzzifier*, dan $d_{ik}^2(x_k, p_i)$ adalah jarak antara vektor pengamatan ke-k dengan pusat kelompok ke-i.

5) Menghitung perubahan matriks keanggotaan dengan persamaan:

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}^2}{d_{jk}^2} \right)^{\frac{1}{m-1}}} \tag{3}$$

dengan u_{ik} adalah nilai keanggotaan objek ke-*k* dengan pusat kelompok ke-*i*, d_{ik}^2 adalah jarak antara objek ke-*k* dengan pusat kelompok ke-*i*, d_{jk}^2 adalah jarak antara objek ke-*k* dengan pusat kelompok ke-*j*, dan *m* adalah *fuzzifier*.

6) Cek kondisi berhenti
   - Jika $|J_t - J_{t-1}| < \varepsilon$ atau $t > MaxIter$ maka berhenti;
   - Jika tidak : $t = t + 1$, ulangi langkah ke-3

2.2. *Subtractive Fuzzy C-Means* (SFCM)

Subtractive Clustering (SC) merupakan metode pengelompokan yang diperkenalkan oleh Stephen L. Chiu pada tahun 1994. Metode ini merupakan modifikasi dari *Mountain Method* (MM) yang diperkenalkan oleh Yager dan Filev pada tahun 1992.

Dimisalkan terdapat n buah titik data $\{x_1, x_2,.., x_n\}$ dalam sebuah ruang berdimensi M. Setiap titik data dipertimbangkan sebagai kandidat pusat kelompok yang nilai potensinya dihitung dengan persamaan[3]:

$$P_i = \sum_{j=1}^{n} e^{-\frac{4}{r_\alpha^2}\cdot\left\|x_i - x_j\right\|^2} \tag{4}$$

P_i adalah nilai potensi untuk titik data ke-i atau potensi dari nilai x_i, $\|.\|$ menotasikan jarak Euclidean, dan r_α adalah sebuah konstanta positif. Konstanta r_α merupakan radius atau jari-jari yang mendefinisikan sebuah lingkungan tetangga. Data yang memiliki potensi yang tinggi adalah data yang memiliki jumlah tetangga paling banyak.

Setelah nilai potensi dari semua titik data dihitung, titik data dengan nilai potensi paling tinggi dipilih sebagai pusat kelompok pertama. Misalkan x_1^* adalah data yang terpilih sebagai pusat kelompok pertama dan P_1^* sebagai nilai potensinya. Nilai potensi untuk setiap titik data diperbaharui dengan persamaan[7]:

$$P_i = P_i - P_1^* e^{-\frac{4}{r_\beta^2}\cdot\left\|x_i - x_1^*\right\|^2} \tag{5}$$

dimana r_β merupakan konstanta positif.

Setelah nilai potensi tiap titik data diperbaharui, titik data dengan nilai potensi terbesar dipilih sebagai pusat kelompok kedua. Selanjutnya nilai potensi tiap titik data kembali diperbaharui. Proses ini akan terus berlanjut sampai diperoleh jumlah kelompok yang cukup.

Subtractive Fuzzy C-Means (SFCM) merupakan penggabungan antara metode pengelompokan *Subtractive Clustering* (SC) dan *Fuzzy C-Means* (FCM). SC digunakan untuk menentukan jumlah kelompok dan matriks keanggotaan awal FCM sehingga inisialisasi secara random tidak perlu dilakukan. Secara garis besar algoritma metode SFCM adalah sebagai berikut:

1) Menghitung potensi awal tiap titik data dengan Persamaan (4)
2) Menetapkan titik data dengan potensi paling tinggi sebagai pusat kelompok pertama dan P_1^* sebagai nilai potensinya.
3) Perbaharui potensi tiap titik data dengan Persamaan (5)
4) Jika rasio potensi titik data tertinggi terbaru dengan potensi titik data tertinggi sebelumnya lebih besar dari *accept ratio* maka titik data dengan potensi tertinggi terbaru ditetapkan menjadi pusat kelompok dan tahap 3 diulangi kembali. Jika rasio kurang dari *reject ratio* maka iterasi dihentikan berlanjut ke tahap 5.
5) Menghitung matriks keanggotaan awal FCM berdasarkan pusat kelompok yang didapatkan melalui metode SC menggunakan Persamaan (3). Jika terdapat jarak d_{ik}^2 yang bernilai nol maka u_{ik} (nilai keanggotaan data k pada kelompok ke-i) akan bernilai 1 dan u_{jk} (nilai keanggotaan data k pada kelompok lainnya) akan bernilai 0[7].
6) Melakukan tahapan 3 sampai dengan 6 pada algoritma FCM hingga ditemukannya matriks keanggotaan dan pusat kelompok terakhir.

2.3. Indeks Validitas

Beberapa indeks validitas yang sering digunakan dalam penelitian-penelitian adalah :

1) *Partition Coefficient* (PC)

Indeks ini mengukur jumlah *overlapping* antarkelompok. Persamaan indeks PC oleh Bezdek sebagai berikut[9] :

$$PC(c) = \frac{1}{N}\sum_{i=1}^{c}\sum_{k=1}^{N} u_{ik}^2 \tag{6}$$

dengan *N* adalah banyak objek penelitian, *c* adalah banyak kelompok, dan u_{ik} adalah nilai keanggotaan objek ke-k dengan pusat kelompok ke-i. Indeks ini memiliki rentang 1/c sampai 1. Jumlah kelompok yang optimal ditunjukkan oleh nilai PC yang paling besar.

2) *Classification Entropy* (CE)

CE hanya mengukur kekaburan (*fuzziness*) dari partisi kelompok. Persamaan indeks dapat dituliskan sebagai berikut[9]:

$$CE(c) = -\frac{1}{N}\sum_{i=1}^{c}\sum_{k=1}^{N} u_{ik}\ln(u_{ik}) \tag{7}$$

dengan *N* adalah banyak objek penelitian, *c* adalah banyak kelompok, dan u_{ik} adalah nilai keanggotaan objek ke-k dengan pusat kelompok ke-i. Indeks ini memiliki rentang 0 sampai ln(c). Indeks CE yang semakin kecil menunjukkan pengelompokan yang lebih baik.

3) *Partition Index (PI)/Separation and Compactness* (SC)

PI merupakan rasio antara jumlah kepadatan dan pemisahan kelompok-kelompok. Indeks ini ditulis menjadi persamaan sebagai berikut[9]:

$$PI(c) = \sum_{i=1}^{c}\frac{\sum_{k=1}^{N}(u_{ik})^m\|x_k - v_i\|^2}{N_i\sum_{j=1}^{c}\|v_j - v_i\|^2} \tag{8}$$

dengan *N* adalah banyak objek penelitian, N_i adalah banyak objek penelitian kelompok ke-i, *c* adalah banyak kelompok, u_{ik} adalah nilai keanggotaan objek ke-k dengan pusat kelompok ke-i, *m* adalah *fuzzifier*, $\|x_k - v_i\|$ adalah jarak euclidean titik data (x_k) dengan pusat kelompok v_i, dan $\|v_j - v_i\|$ jarak euclidean antar pusat kelompok. Nilai SC yang rendah mengindikasikan partisi kelompok yang lebih baik.

4) *Fukuyama Sugeno Index* (FS)

Persamaan indeks oleh Fukuyama dan Sugeno sebagai berikut[9]:

$$FS(c) = \sum_{i=1}^{c}\sum_{k=1}^{N}(u_{ik})^m\|x_k - v_i\|^2 - \sum_{i=1}^{c}\sum_{k=1}^{N}(u_{ik})^m\|v_i - \bar{v}\|^2 \tag{9}$$

dengan *N* adalah banyak objek penelitian, *c* adalah banyak kelompok, u_{ik} adalah nilai keanggotaan objek ke-k dengan pusat kelompok ke-i, *m* adalah *fuzzifier*, $\|x_k - v_i\|$ adalah jarak euclidean titik data (x_k) dengan pusat kelompok v_i, $\|v_i - \bar{v}\|$ adalah jarak euclidean pusat kelompok v_i dengan rata-rata pusat kelompok. Nilai FS yang rendah mengindikasikan partisi kelompok yang lebih baik.

5) *Xie and Beni's Index (XB)*

XB bertujuan untuk menghitung rasio total variasi di dalam kelompok dan pemisahan kelompok. Indeks ini dituliskan sebagai berikut[9] :

$$XB(c) = \frac{\sum_{i=1}^{c}\sum_{k=1}^{N}(u_{ik})^m\|x_k - v_i\|^2}{N\min_{i,k}\|v_k - v_i\|^2} \tag{10}$$

dengan *N* adalah banyak objek penelitian, *c* banyak kelompok, u_{ik} adalah nilai keanggotaan objek ke-k dengan pusat kelompok ke-i. *m* adalah *fuzzifier*, $\|x_k - v_i\|$ adalah

jarak euclidean titik data (x_k) dengan pusat kelompok v_i,dan $\|\mathrm{v_k} - \mathrm{v_i}\|$ adalah jarak euclidean antar pusat kelompok. Nilai XB yang rendah mengindikasikan partisi kelompok yang lebih baik.

6) *Modified Partition Coefficient (MPC).*

Indeks ini diajukan oleh Dave (1996) untuk mengatasi kekurangan PC dan CE. Nilai PC dan CE memiliki kecenderungan berubah secara monoton seiring dengan berubahnya nilai c (Wang dan Zhang, 2007). Persamaan indeks ini dituliskan sebagai berikut[9]:

$$MPC(c) = 1 - \frac{c}{c-1}\left(1 - PC\right) \tag{11}$$

dengan *c* adalah banyak kelompok dan PC adalah indeks PC.

3. Metodologi Penelitian

3.1. Sumber Data

Data yang digunakan pada penelitian ini merupakan data bangkitan berdistribusi Uniform dengan rentang 0-1 yang diperoleh melalui fungsi *runif()* pada *software* R. Jumlah data bangkitan yang digunakan adalah sebanyak 20 dan 100 data yang masing-masing terdiri dari tiga variabel

3.2. Metode Penelitian

Metode analisis penelitian ini adalah sebagai berikut:

1) Melakukan pengelompokan menggunakan metode FCM dan SFCM pada data bangkitan 20 dan 100 data.
2) Menghitung enam indeks validitas hasil pengelompokan metode FCM dan SFCM
3) Menentukan jumlah kelompok optimal FCM dan SFCM berdasarkan enam indeks validitas
4) Membandingkan hasil pengelompokan FCM dan SFCM pada jumlah kelompok optimal.

4. Hasil dan Pembahasan

4.1. Data Bangkitan 20 Data

Pengelompokan metode FCM dilakukan pada jumlah kelompok sebanyak dua sampai dengan sepuluh. Parameter FCM lainnya ditetapkan sama untuk tiap jumlah kelompok yaitu *fuzzifier* sebesar 2, iterasi maksimum sebanyak 1000 kali, dan ε sebesar 10^{-5}. Indeks validitas yang didapatkan untuk tiap jumlah kelompok seperti tercantum pada Tabel 1.

Tabel 1. Indeks Validitas Hasil Pengelompokan FCM Data Bangkitan 20

c	PC	MPC	CE	FS	XB	PI
2	0,6635	0,3271	0,5115	1,2777	0,4598	0,9412
3	0,5860	0,3789	0,7346	0,0304	0,2412	0,3088
4	0,5818	0,4424	0,8195	-1,0955	0,1603	0,1210
5	0,5540	0,4425	0,9174	-1,4625	0,2047	0,0761
6	0,5438	0,4526	0,9842	-1,6484	0,1528	0,0624
7	0,5596	0,4861	0,9988	-1,8248	0,1936	0,0495
8	0,5514	0,4873	1,0456	-1,8670	0,1734	0,0418
9	0,5751	0,5220	1,0245	-1,9909	0,1371	0,0376
10	0,6010	0,5567	0,9881	-2,1677	0,1306	0,0328

Dari Tabel 1 dijelaskan bahwa jumlah kelompok sebanyak sepuluh mempunyai indeks MPC, FS, XB, dan PI terbaik sehingga dapat disimpulkan bahwa berdasarkan metode FCM jumlah kelompok optimal pada data tersebut sebanyak sepuluh kelompok.

Hal yang sama dilakukan menggunakan metode SFCM. Untuk metode SFCM radius diubah-ubah sampai didapatkan jumlah kelompok sebanyak dua sampai dengan sepuluh. Parameter SFCM lainnya ditetapkan sama untuk tiap radius yaitu *squash factor* sebesar 1.5, *accept ratio* sebesar 0.5, *reject ratio* sebesar 0.15, *fuzzifier* sebesar 2, iterasi maksimum sebanyak 1000 kali, dan ε sebesar 10^{-5}. Indeks validitas yang didapatkan untuk tiap jumlah kelompok seperti pada Tabel 2.

Tabel 2. Indeks Validitas Hasil Pengelompokan SFCM Data Bangkitan 20

r	c	PC	MPC	CE	FS	XB	PI
0,68	2	0,6635	0,3271	0,5115	1,2777	0,4598	0,9412
0,65	3	0,5860	0,3789	0,7346	0,0304	0,2412	0,3088
0,6	4	0,5818	0,4424	0,8195	-1,0955	0,1603	0,1210
0,55	5	0,5540	0,4425	0,9174	-1,4625	0,2047	0,0761
0,45	6	0,5533	0,4639	0,9643	-1,8919	0,2072	0,0535
0,44	7	0,5725	0,5013	0,9596	-2,3679	0,1624	0,0366
0,41	8	0,5898	0,5312	0,9607	-2,5842	0,1836	0,0308
0,4	9	0,6096	0,5608	0,9393	-2,7831	0,0970	0,0217
0,35	10	0,6183	0,5759	0,9369	-2,9008	0,1013	0,0168

Berdasarkan Tabel 2 terlihat bahwa jumlah kelompok sebanyak sepuluh memberikan hasil terbaik pada indeks MPC, FS, dan PI. Dapat disimpulkan bahwa berdasarkan metode SFCM jumlah kelompok optimal pada data tersebut sebanyak sepuluh kelompok. Perbandingan hasil pengelompokan optimal kedua metode tersebut seperti yang tercantum pada Tabel 3.

Tabel 3. Perbandingan Hasil FCM dan SFCM Data Bangkitan 20

Metode	Waktu Komputasi (detik)	Fungsi Objektif Terakhir	PC	MPC	CE	FS	XB	PI
FCM	14,8819	0,20634	0,6010	0,5567	0,988	-2,1677	0,1306	0,0328
SFCM	14,1240	0,16256	0,6183	0,5759	0,937	-2,9008	0,1013	0,0168

Tabel 3 menginformasikan bahwa metode SFCM memberikan hasil pengelompokan yang lebih baik berdasarkan keenam indeks. Selain itu dengan iterasi sebanyak 1000 kali, metode SFCM memberikan nilai fungsi objektif terakhir lebih kecil dibandingkan FCM. Dari segi waktu komputasi kecepatan metode SFCM memberikan hasil yang lebih cepat 0,7579 detik. Selain itu, berdasarkan dua tabel hasil perhitungan sebelumnya terlihat bahwa secara umum metode SFCM memberikan hasil pengelompokan yang lebih baik dibandingkan metode FCM. Jika dikaitkan dengan inkonsistensi FCM, pada jumlah kelompok dua sampai dengan lima FCM memberikan hasil yang stabil. Pada saat itulah nilai yang dihasilkan oleh FCM dan SFCM persis sama. Namun pada jumlah kelompok di atas lima FCM memberikan hasil yang bervariasi setiap kali pengelompokan dilakukan.

4.2. Data Bangkitan 100 Data

Tahapan yang sama seperti pada pengelompokan data bangkitan 20 data dilakukan pada data bangkitan 100 data. Indeks validitas yang didapatkan untuk tiap jumlah kelompok menggunakan metode FCM seperti tercantum pada Tabel 4.

Tabel 4. Indeks Validitas Hasil Pengelompokan FCM Data Bangkitan 100

c	PC	MPC	CE	FS	XB	PI
2	0,6354	0,2708	0,5443	7,8340	0,4708	0,9433
3	0,5275	0,2913	0,8139	2,2282	0,3243	0,3807
4	0,4590	0,2787	1,0213	-0,3445	0,2823	0,2319
5	0,4214	0,2768	1,1645	-2,0994	0,2690	0,1405
6	0,3999	0,2799	1,2738	-3,1684	0,2065	0,1069
7	0,3862	0,2839	1,3571	-4,1055	0,3287	0,0821
8	0,3840	0,2960	1,4133	-4,9074	0,3823	0,0689
9	0,3716	0,2930	1,4876	-5,1729	0,2512	0,0576
10	0,3783	0,3093	1,5124	-6,0103	0,2148	0,0473

Berdasarkan Tabel 4 terlihat bahwa jumlah kelompok sebanyak sepuluh memberikan hasil terbaik pada indeks MPC, FS, dan PI sehingga dapat disimpulkan bahwa berdasarkan metode FCM jumlah kelompok optimal pada data tersebut sebanyak sepuluh kelompok. Sedangkan indeks validitas pengelompokan menggunakan metode SFCM untuk tiap jumlah kelompok didapatkan seperti tercantum pada Tabel 5.

Tabel 5. Indeks Validitas Hasil Pengelompokan SFCM Data Bangkitan 100

r	c	PC	MPC	CE	FS	XB	PI
0,79	2	0,6354	0,2708	0,5443	7,8340	0,4708	0,9433
0,75	3	0,5275	0,2913	0,8139	2,2282	0,3243	0,3807
0,7	3	0,5275	0,2913	0,8139	2,2282	0,3243	0,3807
0,65	3	0,5275	0,2913	0,8139	2,2282	0,3243	0,3807
0,6	3	0,5275	0,2913	0,8139	2,2282	0,3243	0,3807
0,55	4	0,4590	0,2787	1,0213	-0,3445	0,2823	0,2319
0,53	5	0,4214	0,2768	1,1645	-2,0994	0,2690	0,1405
0,5	6	0,3999	0,2799	1,2738	-3,1684	0,2065	0,1069
0,465	7	0,3916	0,2902	1,3504	-4,3800	0,1790	0,0831
0,45	8	0,3819	0,2937	1,4235	-5,0889	0,1800	0,0679
0,4	9	0,3756	0,2975	1,4758	-5,3989	0,2965	0,0555
0,37	10	0,3759	0,3066	1,5154	-6,0077	0,2478	0,0463

Pada Tabel 5 menjelaskan bahwa jumlah kelompok sebanyak sepuluh memberikan hasil terbaik pada indeks MPC, FS, dan PI sehingga disimpulkan bahwa berdasarkan metode SFCM jumlah kelompok optimal pada data tersebut sebanyak sepuluh kelompok. Perbandingan hasil pengelompokan optimal kedua metode tersebut tercantum pada Tabel 6.

Berdasarkan Tabel 6 dapat disimpulkan bahwa metode FCM memberikan hasil pengelompokan yang lebih baik berdasarkan lima indeks kecuali PI. Selain itu dengan iterasi sebanyak 1000 kali, metode FCM memberikan nilai fungsi objektif lebih kecil dibandingkan SFCM. Dari segi waktu komputasi, kecepatan metode FCM memberikan hasil yang lebih cepat 14,6904 detik. Namun, jika dilihat secara keseluruhan, metode

SFCM memberikan hasil pengelompokan yang lebih baik dibandingkan metode FCM. Hasil pengelompokan kedua metode sama pada jumlah kelompok dua sampai dengan enam. Namun, pada jumlah kelompok tujuh dan sembilan SFCM memberikan hasil yang lebih baik dibandingkan FCM. Pada jumlah kelompok delapan kedua metode memberikan hasil seri dimana masing-masing metode dinilai baik oleh tiga indeks berbeda.

Tabel 6. Perbandingan Hasil Optimum FCM dan SFCM Data Bangkitan 100

Metode	Waktu Komputasi (detik)	Fungsi Objektif Terakhir	PC	MPC	CE	FS	XB	PI
FCM	52,9887	1,8584	0,3783	0,3093	1,512	-6,0103	0,2148	0,0473
SFCM	67,9491	1,8659	0,3759	0,3066	1,515	-6,0077	0,2478	0,0463

5. Kesimpulan

Sesuai dengan tujuan dari penelitian, berdasarkan hasil uji coba dengan menggunakan data bangkitan menghasilkan kesimpulan bahwa pengelompokan menggunakan metode SFCM memberikan hasil lebih baik dibandingkan metode FCM pada data sebanyak 20, sedangkan FCM memberikan hasil lebih baik dibandingkan metode SFCM pada data sebanyak 100 dimana jumlah kelompok optimal yang diberikan oleh kedua metode pada kedua data bangkitan sama. Ketika hasil yang diberikan FCM pada tiap kali perulangan adalah konsisten, hasil yang diberikan kedua metode persis sama. Jika dilihat secara keseluruhan pada umumnya SFCM memberikan hasil pengelompokan yang lebih baik dibandingkan FCM pada jumlah kelompok sama. Untuk penelitian selanjutnya disarankan untuk membandingkan jarak yang dipakai dalam kedua metode tersebut, misalnya dengan jarak manhattan dan mahalanobis. Diharapkan dengan perbedaan jarak yang digunakan akan menghasilkan kehandalan yang lebih baik lagi.

DAFTAR PUSTAKA

1. Balasko, B., Abonyi, J. dan Feil, B., *Fuzzy Clustering and Data Analysis Toolbox (for Use With Matlab)*, [online] Available: http://www.fmt.vein.hu/softcomp/fclusttoolbox/.
2. Bataineh, K.M., Naji, M., dan Saqer,M., A Comparison Study between Various Fuzzy Clustering Algorithms, *Jordan Journal of Mechanical and Industrial Engineering*, 2011, Vol.5, No.4: 335-343.
3. Chiu, S.L., Fuzzy Model Indentification Based on Cluster Estimation, *Journal of Intelligent and Fuzzy Systems*, 1994, Vol. 2: 267-278.
4. Hossen, J., Rahman, A., Sayeed, S., Samsuddin, K., dan Rokhani, F, A Modified Hybrid Fuzzy Clustering Algorithm for Data Partitions, *Australian Journal of Basic and Applied Sciences*, 2011, Vol.5, No.8: 674-681.
5. Kusumadewi, S. dan Purnomo, H., *Aplikasi Logika Fuzzy untuk Pendukung Keputusan*, Graha Ilmu, Yogyakarta, 2010.
6. Le, T. dan Altman, T., A new initialization method for the Fuzzy C-Means Algorithm using Fuzzy Subtractive Clustering, *Proc. International Conference on Information and Knowledge Engineering*, Las Vegas USA, 2011, Vol. 1: 144-150.
7. Liu, W.Y., Xiao, C.J., Wang, B.W., Shi, Y., dan Fang, S.F., Study on Combining Subtractive Clustering with Fuzzy C-Means Clustering, *International Conference on Machine Learning and Cybernetics*, 2003, Vol. 5: 2659 – 2662.

8. Suman dan Mittal, P., Comparison and Analysis of Various Clustering Methods in Data mining On Education Data Set Using the WEAK Tool, *International Journal of Emerging Trends & Technology in Computer Science*, 2014, Vol. 3, Issue 2.
9. Wang, W. dan Zhang, Y., On Fuzzy Cluster Validity Indices, *Fuzzy Sets System*, 2007, Vol. 158, No. 19: 2095-2117.

Kumpulan jurnaL Ilmu Komputer (KLIK)
Volume 01, No.01 September 2014
ISSN: 2406-7857

IMPLEMENTASI METODE FUZZY C-MEANS PADA SISTEM CLUSTERINGDATA VARIETAS PADI

Nurjanah[1], Andi Farmadi[2], Fatma Indriani[2]
[1,2,3] Prodi Ilmu Komputer Fakultas MIPA Universitas Lambung Mangkurat

Jl. A. Yani Km 36 Banjarbaru, Kalimantan Selatan
Email: nurjanah_abs@yahoo.co.id

Abstract

Mutations with gamma rays conducted on five local rice varieties tidal South Kalimantan produce a lot of data availability. In order for these data not only become a graveyard of useless data required a method that could be used to probe the hidden information from the data. The method known as data mining. Data mining is a technique to gain knowledge from the data by looking for certain patterns or rules of a number of large amounts of data. One method of data mining is clustering, where clustering is usually used to group objects that are similar in the same class or segment. By utilizing the data of local rice varieties tidal South Kalimantan mutated by gamma rays, data mining process is done by grouping the data based on the harvest age, productive tillers, and weight of 1000 seeds into 4 groups using fuzzy c-means algorithm. From that cluster information, carried ranking using the Simple Additive Weighting method and acquired knowledge about improved varieties by harvest age, productive tillers, and a weight of 1000 is kuatek with a dose of 30 krad.

Keywords *: Data Mining, Cluster, Fuzzy C-Means, Local Rice Varieties*

Abstrak

Mutasi dengan sinar gamma yang dilakukan terhadap lima varietas padi lokal pasang surut kalimantan selatan menghasilkan tersedianya banyak data. Agar data-data tersebut tidak hanya menjadi kuburan data yang tidak berguna dibutuhkan sebuah metode yang bisa digunakan untuk menggali informasi–informasi tersembunyi dari data tersebut. Metode tersebut dikenal dengan data mining. Data mining merupakan suatu teknik untuk menggali pengetahuan dari data dengan mencari pola atau aturan tertentu dari sejumlah data dalam jumlah besar. Salah satu metode data mining adalah klastering, dimana klastering biasanya digunakan untuk mengelompokan objek-objek yang memiliki kemiripan dalam kelas atau segmen yang sama. Dengan memanfaatkan data varietas padi hasil mutasi dengan sinar gamma dilakukan proses penggalian data dengan cara mengelompokkan data umur panen, anakan produktif, dan bobot 1000 menjadi 4 klaster menggunakan algoritma fuzzy c-means. Dari informasi 4 klaster tersebut, dilakukan perangkingan klaster menggunakan metode Simple Additive Weighting dan diperoleh pengetahuan tentang varietas unggul berdasarkan umur panen, anakan produktif, dan bobot 1000 yaitu kuatek dengan dosis 30 krad.

Kata kunci *: Data Mining, Klaster, Fuzzy C-Means, Varietas Padi Lokal*

1. PENDAHULUAN

Varietas padi lokal Kalimantan Selatan masih menjadi pilihan para petani untuk ditanam, yang dalam penanamannya tidak terlalu bergantung pada input pertanian. Selain itu, varietas padi lokal juga bersifat adaptif dengan lingkungan Kalimantan Selatan yang didominasi oleh lahan rawa, meskipun berumur panjang dan produktivitas relatif rendah daripada varietas padi unggul. Akan tetapi sifat "pera" nasi yang dihasilkan membuat harga jual varietas padi lokal lebih tinggi dibandingkan dengan varietas padi unggul [3].

Perjuangan dalam peningkatan produksi pangan untuk memenuhi kebutuhan masyarakat Kalimantan Selatan merupakan tantangan bagi masyarakat pertanian yang dapat melahirkan peluang berupa pengembangan teknologi alternatif untuk menerobos kebuntuan dan stagnasi produksi beras lokal. Alternatif itu dapat diarahkan pada pengembangan varietas lokal yang dapat beradaptasi pada lingkungan produksi yang buruk dengan produktivitas yang tinggi dan umur panen yang pendek. Peningkatan keragaman populasi dapat dilakukan salah satunya melalui induksi mutasi secara fisik dengan iradiasi sinar gamma.

Salah satu upaya untuk meningkatkan produksi pangan varietas padi lokal pasang surut kalimantan selatan yaitu dengan cara meningkatkan keragaman populasi, dilakukan radiasi dengan sinar gamma terhadap lima varietas padi. Dari upaya tersebut dihasilkan banyak data salah satunya yaitu data varietas padi hasil mutasi dengan sinar gammamaka dibutuhkan sebuah metode yang bisa digunakan untuk menggali informasi–informasi tersembunyi dari data tersebut. Metode tersebut dikenal dengan *data mining*.

Data mining adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar. Salah satu tugas dalam data mining adalah klastering. Tujuan utama dari klastering adalah pengelompokan sejumlah data/obyek ke dalam klaster sehingga dalam setiap klaster akan berisi data yang semirip mungkin [1].

Penggalian suatu data menjadi pengetahuan adalah kegiatan yang sangat berguna, karena pengetahuan tersebut bisa digunakan untuk memecahkan suatu masalah yang ada. *Data mining* merupakan salah satu teknik untuk menggali pengetahuan dari data. Agar data-data tersebut tidak menjadi kuburan data yang tidak bergunamaka dibuat sistem yang mampu menggali data varietas padi yang dimutasi dengan sinar gamma yang dapat mengorganisir data varietas padi yang dimutasi dengan sinar gamma dalam suatu database serta dapat menghasilkan pengelompokan sejumlah data/obyek ke dalam klaster yang memberikan pengetahuan mengenai varietas unggulmenggunakan metode *Fuzzy C-Means* (FCM).

2. METODOLOGI PENELITIAN

2.1. Analisa Data

Pada tahap ini akan dilakukan analisis data dengan KDD (*knowledge Discovery in Databases*). Keseluruhan proses KDD dari mulai memasukan data mengalami 3 tahap secara global sebelum menjadi hasil yang berupa informasi, yaitu:

(a) *Preprocessing*

Pada tahap ini struktur basis data akan dipersiapkan sehingga mempermudah proses mining. Proses *preprocessing* ini mencakup tiga hal utama yaitu :

1) seleksi data
   untuk memilih data yang akan digunakan dalam proses data mining. Dalam penelitian ini dipilih 3 data varietas padi yaitu umur panen, anakan produktif, dan bobot 1000 biji;
2) *Cleaning*
   Memastikan kualitas data yang telah dipilih pada tahap seleksi data, pada tahap ini masalah yang harus dihadapi adalah noisy data dan missing values. Proses cleaning data perlu dilakukan agar data bersih dari duplikasi data, data yang tidak konsisten, atau kesalahan cetak. Sehingga data yang telah melewati proses ini siap untuk di proses di data mining. Pada penelitian ini proses pembersihan data hanya dilakukan pada data yang missing value;
3) integrasi data
   Setelah melakukan pembersihan data yang menyeluruh, maka akan dilakukan tahapan selanjutnya yaitu integrasi data, karena data lebih dari satu tabel maka dilakukan tahap penggabungan data dari berbagai tabel. Kemudian data padi diberi kode agar tidak ada data yang memiliki nama yang sama. Kode padi terdiri dari nomor padi, baris padi, dosis radiasi, dan varietas padi.

(b) Data Mining

Proses data mining yang dilakukan pada penelitian ini adalah menerapkan teknik *clustering* dengan algoritma *Fuzzy C-Means* untuk mendapatkan karakteristik dari data. *Cluster* yang digunakan pada penelitian ini sebanyak empat klaster.

Konsep dasar FCM (*Fuzzy C-Means*) pertama kali adalah menentukan pusat *cluster* yang akan menandai lokasi rata-rata untuk tiap *cluster*. Pada kondisi awal, pusat *cluster* ini masih belum akurat. Tiap-tiap titik data memiliki derajat keanggotan untuk tiap-tiap *cluster*. Dengan cara memperbaiki pusat *cluster* dan derajat keanggotaan tiap-tiap titik data secara berulang, maka akan dapat dilihat bahwa pusat *cluster* akan bergerak menuju lokasi yang tepat. Perulangan ini didasarkan pada derajat keanggotaan yang menggambarkan jarak dari titik data yang diberikan ke pusat *cluster* yang terbobot oleh derajat keanggotaan titik data tersebut [6].

Pada perhitungan menggunakan FCM dengan menggunakan data yang sama tetapi diolah dengan jumlah *cluster* yang berbeda, maka hasil pengelompokannya akan sedikit berbeda, karena data tidak diolah dengan satu variabel saja tetapi dengan semua variabel. Perbedaan hasil pengelompokan itu dikarenakan data pada kelompok tertentu kemungkinan akan berpindah pada kelompok lain jika diolah dengan jumlah *cluster* yang berbeda, ini menunjukkan bahwa sistem aplikasi sudah berjalan dengan benar [6]

Output dari FCM bukan merupakan fuzzy inference sistem, namun merupakan derajat pusat *cluster* dan beberapa derajat keanggotaan untuk

tiap-tiap titik data. Algoritma *Fuzzy C-Means* disusun dengan langkah sebagai [6]:

(1) Input data
input data yang akan di*cluster*yaitu data varietas padi (X) berupa matrik berukuran n × m (n = jumlah data, m = atribut setiap data). X_{ij} = data ke- i (i = 1, 2, …, n), atribut ke- j (j = 1, 2, …, m).

(2) Batasan
(a) Jumlah *cluster* = c = 4
(b) Pangkat = w = 2
(c) Maksimum iterasi = Maxit = 100
(d) Error terkecil yang diharapkan = ξ = 10^{-3}
(e) Fungsi obyektif awal = Po = 0
(f) Iterasi awal = t = 1

(3) Membangkitkan bilangan random μ_{ik}, i = 1, 2, …, n; k = 1, 2, …,c; sebagai elemen-elemen matrik partisi awal U, dengan jumlah setiap nilai elemen kolom dalam satu baris adalah 1 (satu)

$$\sum_{i=1}^{c} \mu_{ci} = 1 \qquad \dots (1)$$

(4) Menghitung pusat *cluster* ke- k : V_{kj} dengan k = 1, 2, …, c; dan j = 1, 2, …,m.

$$V_{kj} = \frac{\sum_{i=1}^{n}\left((\mu_{ik})^{w} * X_{ij}\right)}{\sum_{i-1}^{n}(\mu_{ik})^{w}} \qquad \dots (2)$$

(5) Menghitung fungsi obyektif pada iterasi ke- t, P_t :

$$P_t = \sum_{i=1}^{n}\sum_{k=1}^{c}\left(\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right](\mu_{ik})^{w}\right) \qquad \dots (3)$$

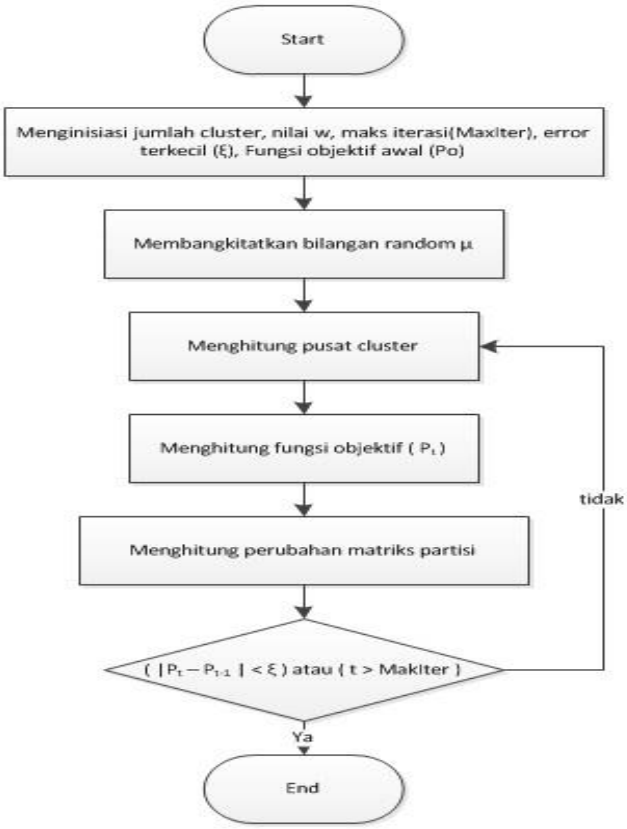
(6) Menghitung perubahan matrik partisi:

$$\mu_{ik} = \frac{\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right]^{\frac{-1}{w-1}}} \dots (4)$$

dengan i = 1, 2, …, n; dan k = 1, 2, …, c

(7) Mengecek kondisi berhenti:
1. Jika : ($|P_t - P_{t-1}| < \xi$) atau (t > MakIter) maka berhenti;
2. Jika tidak : t = t + 1, ulangi langkah ke- d (menghitung V_{kj}).

Bagan alir untuk metode *Fuzzy C-Means* yaitu sebagai berikut :



Gambar 1. Bagan Alir Metode *Fuzzy C-Means*

(c) ***Postprocessing***

Postprocessing bertujuan untuk membantu pengguna dalam memahami informasi, Karena pusat *cluster* masih tidak bisa dienterpretasikan secara langsung, maka diperlukan metode *Simple Additive Weighting* yang bisa melakukan perangkingan nilai pusat *cluster* terakhir tersebut.

Metode SAW (*Simple Additive Weighting*) sering juga dikenal istilah metode penjumlahan terbobot. Konsep dasar metode SAW adalah mencari penjumlahan terbobot dari rating kinerja pada setiap alternatif pada semua atribut. Langkah-langkah metode dalam metode SAW adalah [7] :

1. Membuat matriks keputusan Z berukuran m x n, dimana m = alternatif yang akan dipilih dan n = kriteria .
2. Memberikan nilai x setiap alternatif (i) pada setiap kriteria (j) yang sudah ditentukan, dimana, i=1,2,…m dan j=1,2,…n pada matriks keputusan Z,

$$\mathbf{Z} = \begin{vmatrix} x_{11} & x_{y12} & \dots & x_{1j} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ x_{i1} & x_{i2} & & x_{ij} \end{vmatrix} \dots (5)$$

3. Memberikan nilai bobot preferensi (W) oleh pengambil keputusan untuk masing-masing kriteria yang sudah ditentukan.

$$W = [W_1\ W_2 W_3 \dots\ W_j] \dots(6)$$

4. Melakukan normalisasi matriks keputusan Z dengan cara menghitung nilai rating kinerja ternormalisasi (r_{ij}) dari alternatif A_i pada atribut C_j

$$r_{ij} = \begin{cases} \frac{x_{ij}}{\underset{i}{MAX}(x_{ij})} & \text{jika j adalah atribut keuntungan} \\ \frac{\underset{i}{MIN}(x_{ij})}{x_{ij}} & \text{jika j adalah atribut biaya} \end{cases} \dots(7)$$

Dengan ketentuan :

a. Dikatakan atribut keuntungan apabila atribut banyak memberikan keuntungan bagi pengambil keputusan, sedangkan atribut biaya merupakan atribut yang banyak memberikan pengeluaran jika nilainya semakin besar bagi pengambil keputusan.
b. Apabila berupa atribut keuntungan maka nilai (x_{ij}) dari setiap kolom atribut dibagi dengan nilai (MAX x_{ij}) dari tiap kolom, sedangkan untuk atribut biaya, nilai (MIN x_{ij}) dari tiap kolom atribut dibagi dengan nilai (x_{ij}) setiap kolom.

5. Hasil dari nilai rating kinerja ternormalisasi (r_{ij}) membentuk matriks ternormalisasi (N)

$$\mathbf{N} = \begin{vmatrix} r_{11} & r_{y12} & \dots & r_{1j} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ r_{i1} & r_{i2} & & r_{ij} \end{vmatrix} \dots (8)$$

6. Melakukan proses perankingan dengan cara mengalikan matriks ternormalisasi (N) dengan nilai bobot preferensi (W).
7. Menentukan nilai preferensi untuk setiap alternatif (V_i) dengan cara menjumlahkan hasil kali antara matriks ternormalisasi (N) dengan nilai bobot preferensi (W).

$$V_i = \sum_{j=1}^{n} w_j r_{ij} \dots (9)$$

Nilai Vi yang lebih besar mengindikasikan bahwa alternatif Ai merupakan alternatif terbaik.

2.3. Implementasi

Implementasi dilakukan dengan metode Fuzzy C-Means (FCM) pada sistem clustering data varietas padilokal pasang surut kalimantan selatan yang dimutasi dengan sinar gamma menggunakan delphi XE2

2.4. Pengujian

Untuk memastikan apakah perhitungan dengan metode fuzzy c-means sudah benar dilakukan uji coba perhitungan secara manual meggunakan MS. Excel dan Matlab.Pengujian dilakukan untuk menghindari kesalahan dari sistem yang dibuat. Apabila terjadi kesalahan, maka sistem akan diperbaiki kembali sampai hasil proses sesuai dengan apa yang diharapkan.

3. HASIL DAN PEMBAHASAN

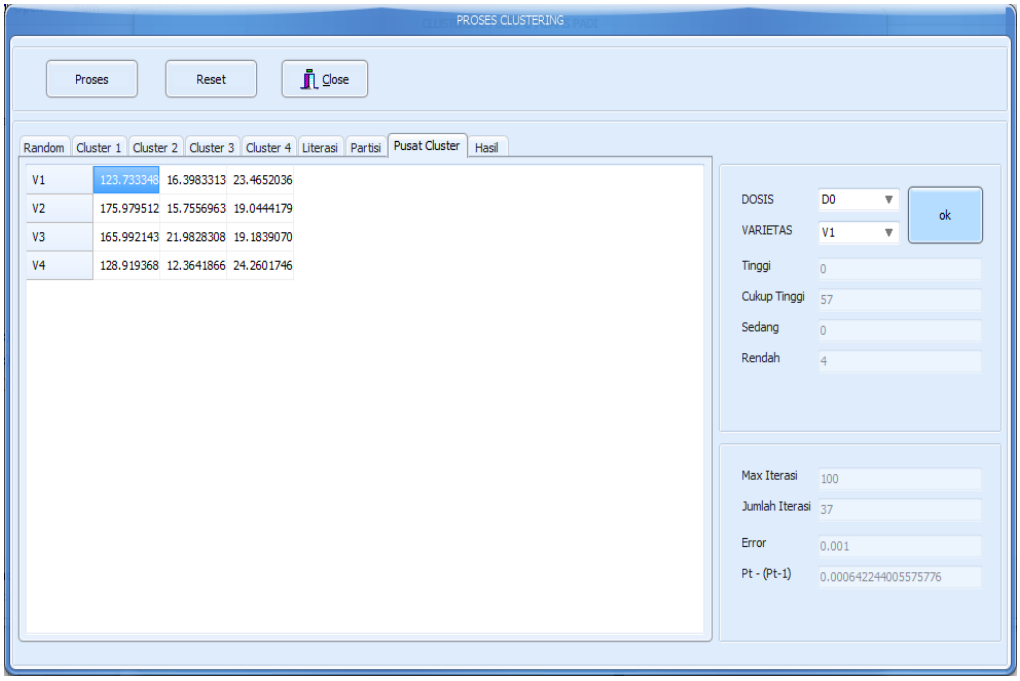
Dariproses clustering Fuzzy C-Means dengan batasan jumlah cluster sebanyak 4 cluster, pangkat 2, masimal iterasi 100, error 0,001 proses berhenti pada iterasi ke-37 dengan error 0,000642 dan diperoleh pusat cluster iterasi terakhir yaitu pada Tabel 1 atau pada Gambar 2 dengan menggunakan rumus dibawah ini (persamaan 2):

$$V_{kj} = \frac{\sum_{i=1}^{2420}\left((\mu_{ik})^2 * X_{ij}\right)}{\sum_{i=1}^{2420}(\mu_{ik})^2}$$

Keterangan :
V_{kj}: pusat cluster dengan k=1,2,...,c dan j = 1,2,...,n
X_{ij} : data ke-i (i=1,2,...m), atribut ke-j (j=1,2,...n)
μ_{ik} : bilangan acak dengan i=1,2,...m dan k=1,2,...c

Tabel 1. Pusat Cluster Delphi

	Umur Panen	Anakan Produktif	Bobot 1000
Klaster 1	165.992143	21.9828308	19.1839070
Klaster 2	123.733348	16.3983313	23.4652036
Klaster 3	128.919368	12.3641866	24.2601746
Klaster 4	175.979512	15.7556963	19.0444179



Gambar 2. Form Proses Fuzzy C-Means

Dari Tabel 1 dapat dilihat bahwa klaster 1 memiliki umur panen terendah ketiga, anakan produktif terbanyak pertama, dan bobot 1000 terberat ketiga. Klaster 2 memiliki umur panen terendah pertama, anakan produktif terbanyak kedua, dan bobot 1000 terberat kedua. Klaster 3 memiliki umur panen terendah kedua, anakan produktif terbanyak keempat, dan bobot 1000 terberat pertama.

Klaster 4 memiliki umur panen terendah keempat, anakan produktif terbanyak ketiga, dan bobot 1000 terberat keempat.

Dilihat dari umur panen klaster 2 lebih unggul karena memiliki umur panen terendah dibandingkan klaster 1, klaster 3, dan klaster 4. Untuk anakan produktif klaster 1 lebih unggul karena memiliki anakan produktif terbanyak. Dan untuk bobot 1000 klaster 3 lebih unggul karena memiliki bobot seribu terberat apabila dibandingkan dengan klaster lainnya.

Dari hasil pusat klaster (Tabel 1) tersebut dilakukan perangkingan dengan metode *Simple Additive Weighting* (SAW) untuk melihat klaster mana yang memiliki umur panen, anakan produktif dan bobot 1000 terbaik. Perhitungan nilai prefrensi dilakukan menggunakan rumus dibawah ini (persamaam 9) :

$$V_i = \sum_{j=1}^{n} w_j r_{ij}$$

Keterangan :
V_i : nilaipreferensiuntuksetiapalternatif
w_j : nilai bobot preferensi
r_{ij} : matriks ternormalisasi

Berikut tabel nilai prefrensi dan tabel hasil pengelompokan per varietas :

Tabel 2. Nilai Prefensi Klaster

Klaster	Nila Prefensi	Kategori
Klaster 1	2.535898966	Cukup Tinggi
Klaster 2	2.712760587	Tinggi
Klaster 3	2.53314209	Sedang
Klaster 4	2.209208036	Rendah

Tabel 3. Hasil Pengelompokan Varietas Siam Harli

Kategori	Varietas Siam Harli (V1)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	-	-	-	46
Cukup Tinggi	57	77	32	-
Sedang	-	-	-	44
Rendah	4	2	3	-

Tabel 4. Hasil Pengelompokan Varietas Siam Unus

Kategori	Varietas Siam Unus (V2)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	-	1	-	-
Cukup Tinggi	48	205	48	86
Sedang	-	1	4	-
Rendah	4	-	5	8

Tabel 5. Hasil Pengelompokan Varietas Kuatek

Kategori	Varietas Kuatek (V3)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	-	-	-	371
Cukup Tinggi	7	152	18	176
Sedang	-	-	-	-
Rendah	41	5	132	-

Tabel 6. Hasil Pengelompokan Varietas Siam Sebelas

Kategori	Varietas Siam Sebelas (V4)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	5	-	-	25
Cukup Tinggi	1	28	9	-
Sedang	1	-	-	22
Rendah	25	158	78	-

Tabel 7. Hasil Pengelompokan Varietas Siam Gumpal

Kategori	Varietas Siam Gumpal (V5)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	-	-	-	13
Cukup Tinggi	-	-	-	-
Sedang	-	-	-	116
Rendah	54	173	105	-

Pengetahuan yang didapat dari hasil penambangan data/klastering ini berupa varietas unggul yaitu varietas yang masuk kategori tinggi. Varietas yang masuk kategori tinggi terdiri atas rata-rata umur panen sebesar 123,7665737; rata-rata anakan produktif sebesar 16,2834365; dan rata-rata bobot 1000 sebesar 23,46984343 dengan nilai preferensi sebesar 2,712760587. Diantara kelima varietas tersebut yang paling banyak masuk kategori tinggi yaitu varietas kuatek dengan dosis 30 krad yakni sebanyak 371 data padi atau sebesar 67,82 % data padi masuk kategori tinggi

Tabel 8. Presentase Data Padi Yang Termasuk Kategori Tinggi

Varietas	Dosis			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Siam Harli (V1)	-	-	-	51,11%
Siam Unus (V2)	-	0,93%	-	-
Kuatek (V3)	-	-	-	67,82%
Siam Sebelas (V4)	15,61%	-	-	53,19%
Siam Gumpal (V5)	-	-	-	10,08%

4. KESIMPULAN

Dari hasil penelitian dan pengamatan dari sistem yang telah dibuat, maka dapat diperoleh kesimpulan sebagai berikut :

1. Metode Fuzzy C-Means berhasil diimplemantasikan pada clustering data varietas padi lokal pasang surut kalimantan selatan yang dimutasi dengan sinar gamma.
2. Dilihat dari umur panen klaster 2 lebih unggul karena memiliki umur panen terendah, untuk anakan produktif klaster 1 lebih unggul karena memiliki anakan produktif terbanyak, dan untuk bobot 1000 klaster 3 lebih unggul karena memiliki bobot seribu terberat apabila dibandingkan dengan klaster lainnya.
3. Pengetahuan yang didapat mengenai varietas unggul berdasarkan umur panen, anakan produktif dan bobot 1000 terbaik menggunakan SAW memiliki nilai preferensi sebesar 23,46984343 yaitu kategori tinggi.
4. Diantara kelima varietas yang paling banyak masuk kategori tinggi yaitu varietas kuatek dengan dosis 30 krad yakni sebanyak 371 data padi atau sebesar 67,82 % data padi masuk kategori tinggi.

DAFTAR PUSTAKA

[1]. Budi Santosa, "**Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis***",* Graha Ilmu, Yogyakarta, 2007.

[2]. Iko Pramudiono. "***Pengantar Data Mining : Menambang Permata Pengetahuan di Gunung Data***" 2003.

[3]. Noor, M., Y. Rina dan Noorginayuwati, "**Persepsi petani tentang lahan gambut dan pengelolaannya",** Laporan Penelitian, Balai Penelitian Pertanian Lahan Rawa, Banjarbaru, 2007.

[4]. Nurjanah, "**Clustering Varietas Padi Lokal Pasang Surut Kalimantan Selatan Yang Dimutasi Dengan Sinar Gamma Menggunakan Metode Fuzzy C-Means**" Naskah Skripsi Universitas Lambung Mangkurat, Banjarbaru, 2014.

[5]. Nuqson Masykur Huda, "***Aplikasi Data Mining Untuk Menampilkan Informasi Tingkat Kelulusan Mahasiswa (Studi Kasus Di Fakultas MIPA Universitas Diponegoro)",*** Naskah Skripsi Universitas Diponegoro, Semarang, 2010.

[6]. Sri Kusumadewi and Hari Purnomo, "**Aplikasi Logika Fuzzy Untuk Pendukung Keputusan***",* Yogyakarta, Graha Ilmu, 2010.

[7]. Wibowo S, Henry., Amalia, Riska., Fadlun M, Andi., Arivanty, Kurnia, "**Sistem Pendukung Keputusan Untuk MenentukanPenerima Beasiswa Bank BRI Menggunakan FMADM (Studi Kasus: Mahasiswa Fakultas Teknologi Industri Unversitas Islam Indonesia)***",* Seminar Nasional Aplikasi Teknologi Informasi, Yogyakarta, 2008.

A Robust Fuzzy Local Information C-Means Clustering Algorithm

Stelios Krinidis and Vassilios Chatzis

***Abstract*—This paper presents a variation of fuzzy c-means (FCM) algorithm that provides image clustering. The proposed algorithm incorporates the local spatial information and gray level information in a novel fuzzy way. The new algorithm is called fuzzy local information C-Means (FLICM). FLICM can overcome the disadvantages of the known fuzzy c-means algorithms and at the same time enhances the clustering performance. The major characteristic of FLICM is the use of a fuzzy local (both spatial and gray level) similarity measure, aiming to guarantee noise insensitiveness and image detail preservation. Furthermore, the proposed algorithm is fully free of the empirically adjusted parameters (a, λ_g, λ_s, etc.) incorporated into all other fuzzy c-means algorithms proposed in the literature. Experiments performed on synthetic and real-world images show that FLICM algorithm is effective and efficient, providing robustness to noisy images.**

***Index Terms*—Clustering, fuzzy c-means, fuzzy constraints, gray level constraints, image segmentation, spatial constraints.**

I. Introduction

IMAGE segmentation is one of the first and most important tasks in image analysis and computer vision. In the literature, various methods have been proposed for object segmentation and feature extraction, described in [1] and [2]. However, the design of robust and efficient segmentation algorithms is still a very challenging research topic, due to the variety and complexity of images. Image segmentation is defined as the partitioning of an image into nonoverlapped, consistent regions which are homogeneous in respect to some characteristics such as intensity, color, tone, texture, etc. The image segmentation can be divided into four categories: thresholding, clustering, edge detection and region extraction. In this paper, a clustering method for image segmentation will be considered.

Clustering is a process for classifying objects or patterns in such a way that samples of the same cluster are more similar to one another than samples belonging to different clusters. There are two main clustering strategies: the hard clustering scheme and the fuzzy clustering scheme. The conventional hard clustering methods classify each point of the data set just to one cluster. As a consequence, the results are often very crisp, i.e., in image clustering each pixel of the image belongs just to one cluster. However, in many real situations, issues such as limited spatial resolution, poor contrast, overlapping intensities, noise and intensity inhomogeneities reduce the effectiveness of hard (crisp) clustering methods. Fuzzy set theory [3] has introduced the idea of partial membership, described by a membership function. Fuzzy clustering, as a soft segmentation method, has been widely studied and successfully applied in image clustering and segmentation [4]–[9]. Among the fuzzy clustering methods, fuzzy c-means (FCM) algorithm [10] is the most popular method used in image segmentation because it has robust characteristics for ambiguity and can retain much more information than hard segmentation methods [11]. Although the conventional FCM algorithm works well on most noise-free images, it is very sensitive to noise and other imaging artifacts, since it does not consider any information about spatial context.

To compensate this drawback of FCM, a preprocessing image smoothing step has been proposed in [9], [12], and [13]. However, by using smoothing filters important image details can be lost, especially boundaries or edges. Moreover, there is no way to control the trade-off between smoothing and clustering. Thus, many researchers have incorporated local spatial information into the original FCM algorithm to improve the performance of image segmentation [5], [11], [14].

Tolias and Panas [5] developed a fuzzy rule-based scheme called the ruled-based neighborhood enhancement system to impose spatial constraints by postprocessing the FCM clustering results.

Noordam *et al.* [6] proposed a geometrically guided FCM (GG-FCM) algorithm, a semi-supervised FCM technique, where a geometrical condition is used determined by taking into account the local neighborhood of each pixel.

Pham [15] modified the FCM objective function by including a spatial penalty on the membership functions. The penalty term leads to an iterative algorithm, which is very similar to the original FCM and allows the estimation of spatially smooth membership functions.

Ahmed *et al.* [9] proposed FCM_S where the objective function of the classical FCM is modified in order to compensate the intensity inhomogeneity and allow the labelling of a pixel to be influenced by the labels in its immediate neighborhood. One disadvantage of FCM_S is that the neighborhood labelling is computed in each iteration step, something that is very time-consuming.

Chen and Zhang [12] proposed FCM_S1 and FCM_S2, two variants of FCM_S algorithm in order to reduce the computational time. These two algorithms introduced the extra mean and median-filtered image, respectively, which can be computed in

Manuscript received May 08, 2009; revised December 15, 2009. First published January 19, 2010; current version published April 16, 2010. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Patrick J. Flynn.

The authors are with the Department of Information Management, Technological Institute of Kavala, 65404, Kavala, Greece (e-mail: stelios.krinidis@mycosmos.gr; chatzis@teikav.edu.gr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2010.2040763

1057-7149/$26.00 © 2010 IEEE

advance, to replace the neighborhood term of FCM_S. Thus, the execution times of both FCM_S1 and FCM_S2 are considerably reduced.

Szilagyi *et al.* [13] proposed the enhanced FCM (EnFCM) algorithm to accelerate the image segmentation process. The structure of the EnFCM is different from that of FCM_S and its variants. First, a linearly-weighted sum image is formed from both original image and each pixel's local neighborhood average gray level. Then clustering is performed on the basis of the gray level histogram instead of pixels of the summed image. Since, the number of gray levels in an image is generally much smaller than the number of its pixels, the computational time of EnFCM algorithm is reduced, while the quality of the segmented image is comparable to that of FCM_S [13].

More recently, Cai *et al.* [16] proposed the fast generalized FCM algorithm (FGFCM) which incorporates the spatial information, the intensity of the local pixel neighborhood and the number of gray levels in an image. This algorithm forms a nonlinearly-weighted sum image from both original image and its local spatial and gray level neighborhood. The computational time of FGFCM is very small, since clustering is performed on the basis of the gray level histogram. The quality of the segmented image is well enhanced [16].

However, EnFCM as well as FGFCM, share a common crucial parameter a (or λ). This parameter is used to control the tradeoff between the original image and its corresponding mean- or median-filtered image. It has a crucial impact on the performance of those methods, but its selection is generally difficult because it should keep a balance between robustness to noise and effectiveness of preserving the details. In other words, the value of a has to be chosen large enough to tolerate the noise, and, on the other hand, it has to be chosen small enough to preserve the image sharpness and details [13]. Thus, we can conclude that the determination of a is in fact noise-dependent to some degree. Since the kind of image noise is generally *a priori* unknown, the selection of a (or λ) is, in practice, experimentally made, usually using trial-and-error experiments [9], [12], [13]. Moreover, the value of a (or λ) is fixed for all pixel neighborhoods over the image.

This paper deals with the above mentioned problems. It presents FLICM, a novel robust fuzzy local information c-means clustering algorithm, which can handle the defect of the selection of parameter a (or λ), as well as promoting the image segmentation performance. In FLICM, a novel fuzzy factor is defined to replace the parameter a used in EnFCM and FCM_S and its variants, and the parameter λ used in FGFCM and its variants. The new fuzzy local neighborhood factor can automatically determine the spatial and gray level relationship and is fully free of any parameter selection. Thus, FLICM has the following attractive characteristics: 1) it is relatively independent of the types of noise, and as a consequence, it is a better choice for clustering in the absence of prior knowledge of the noise; 2) the fuzzy local constraints incorporate simultaneously both the local spatial and the local gray level relationship in a fuzzy way; 3) the fuzzy local constraints can automatically be determined, so there is no need of any parameter determination; 4) the balance among image details and noise is automatically achieved by the fuzzy local constraints, enhancing concurrently the clustering performance. All these characteristics make FLICM a more general and suitable for image clustering algorithm.

The remainder of the paper is organized as follows. Section II briefly describes fuzzy c-means clustering algorithms with spatial constraints (FCM_S, FCM_S1 and FCM_S2), followed by the EnFCM and the FGFCM algorithms. The FLICM algorithm is introduced in Section III. Experimental results are presented in Section IV and conclusions are drawn in Section V.

II. Preliminary Theory

A. Fuzzy C-Means (FCM) Algorithm

The fuzzy c-means (FCM) clustering algorithm was first introduced by Dunn [17] and later extended by Bezdek [10]. The algorithm is an iterative clustering method that produces an optimal c partition by minimizing the weighted within group sum of squared error objective function J_m [10]

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{c} u_{ji}^m d^2(x_i, v_j) \tag{1}$$

where $\mathbf{X} = \{x_1, x_2, \ldots, x_N\} \subseteq \mathbb{R}^m$ is the data set in the m-dimensional vector space, N is the number of data items, c is the number of clusters with $2 \le c < N$, u_{ji} is the degree of membership of x_i in the jth cluster, m is the weighting exponent on each fuzzy membership, v_j is the prototype of the center of cluster j, $d^2(x_i, v_j)$ is a distance measure between object x_i and cluster center v_j. A solution of the object function J_m can be obtained through an iterative process, which is carried as follows.

1) Set values for c, m, and ε.
2) Initialize the fuzzy partition matrix $U^{(0)}$.
3) Set the loop counter $b = 0$.
4) Calculate the c cluster centers $v_j^{(b)}$ with $U^{(b)}$

$$v_j^{(b)} = \frac{\sum_{i=1}^{N}\left(u_{ji}^{(b)}\right)^m x_i}{\sum_{i=1}^{N}\left(u_{ji}^{(b)}\right)^m}. \tag{2}$$

5) Calculate the membership matrix $U^{(b+1)}$

$$u_{ji}^{(b+1)} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ji}}{d_{ki}}\right)^{2/m-1}}. \tag{3}$$

6) If $\max\left\{U^{(b)} - U^{(b+1)}\right\} < \varepsilon$ then stop, otherwise, set $b = b+1$ and go to step 4.

B. Fuzzy Clustering With Constraints (FCM_S) and Its Variants

Ahmed *et al.* [9] proposed a modification of the standard FCM by introducing a term that allows the labelling of a pixel to be influenced by labels in its immediate neighborhood. The neighborhood effect acts as a regularizer and biases the solution

toward piecewise-homogeneous labelling. The modified objective function of FCM_S is defined as follows:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{c} u_{ji}^m \|x_i - v_j\|^2 + \frac{a}{N_R}\sum_{i=1}^{N}\sum_{j=1}^{c} u_{ji}^m \sum_{r\in N_i} \|x_r - v_j\|^2 \quad (4)$$

where x_i is the gray level value of the ith pixel, N is the total number of pixels, v_j is the prototype value of the jth center, u_{ji} represents the fuzzy membership of the ith pixel with respect to cluster j, N_R is its cardinality, x_r represents the neighbor of x_i, and N_i stands for the set of neighbors falling into a window around pixel x_i. The parameter a is used to control the effect of the neighbors term. By definition, each sample point x_i satisfies the constraint that $\sum_{j=1}^{c} u_{ji} = 1$. The calculation of membership partition matrix and the cluster centers is performed as follows:

$$u_{ji} = \frac{\left(\|x_i - v_j\|^2 + \frac{a}{N_R}\sum_{r\in N_i}\|x_r - v_j\|^2\right)^{1/m-1}}{\sum_{k=1}^{c}\left(\|x_i - v_k\|^2 + \frac{a}{N_R}\sum_{r\in N_i}\|x_r - v_k\|^2\right)^{1/m-1}} \quad (5)$$

$$v_j = \frac{\sum_{i=1}^{N} u_{ji}^m \left(x_i + \frac{a}{N_R}\sum_{r\in N_i} x_r\right)}{(1+a)\sum_{i=1}^{N} u_{ji}^m}. \quad (6)$$

The second term $1/N_R \sum_{r\in N_i} x_r$ in the numerator of (6) is in fact a neighbor average gray level value around x_i within a window. The image composed by all the neighbor average values around the image pixels forms a so-called mean-filtered image.

Chen and Zhang [12] proposed FCM_S1, a variant of FCM_S, where the neighborhood term is presented much more simplified. The objective function is written as follows:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{c} u_{ji}^m \|x_i - v_j\|^2 + a\sum_{j=1}^{c}\sum_{r\in N_i} u_{jr}^m \|\overline{x_r} - v_j\|^2 \quad (7)$$

where $\overline{x_r}$ is the average of neighboring pixels lying within a window around x_r. Unlike (4), $\overline{x_r}$ can be computed in advance, reducing the whole computation time since $1/N_R \sum_{r\in N_i} \|x_r - v_j\|^2$ in (4) is replaced by $\|\overline{x_r} - v_j\|^2$. An iterative algorithm for the minimization of (7) with respect to u_{ji} and v_j can be similarly to FCM_S derived as follows:

$$u_{ji} = \frac{\left(\|x_i - v_j\|^2 + a\|\overline{x_i} - v_j\|^2\right)^{1/m-1}}{\sum_{k=1}^{c}\left(\|x_i - v_k\|^2 + a\|\overline{x_i} - v_k\|^2\right)^{1/m-1}} \quad (8)$$

$$v_j = \frac{\sum_{i=1}^{N} u_{ji}^m \left(x_i + a\overline{x_i}\right)}{(1+a)\sum_{i=1}^{N} u_{ji}^m}. \quad (9)$$

The essence of FCM_S1 is to make both the original and the corresponding mean-filtered image, have the same prototypes or clustering results. Furthermore, FCM_S1 not only considerably reduces the execution time, but also improves the robustness to Gaussian noise [12].

However, FCM_S1 is unsuitable for images corrupted by impulse noise [12]. In order to overcome this problem, Chen and Zhang [12] designed the FCM_S2, a variant of FCM_S1 in which the mean-filtered image is replaced by the median-filtered one.

The iterative algorithm that solves the objective functions of FCM_S, FCM_S1 and FCM_S2 is more or less the same used in the classical FCM algorithm. Just the corresponding functions should be replaced and the mean and median-filtered images (only for FCM_S1 and FCM_S2) should be calculated before the first iteration.

C. Enhanced Fuzzy C-Means Clustering (EnFCM)

Szilagyi *et al.* [13] proposed the EnFCM algorithm to speed up the clustering process for gray level images. In order to accelerate FCM_S, a linearly-weighted sum image ξ is in advance formed from the original image and its local neighbor average image in terms of

$$\xi_i = \frac{1}{1+a}\left(x_i + \frac{a}{N_R}\sum_{j\in N_i} x_j\right) \quad (10)$$

where ξ_i denotes the gray level value of the ith pixel of the image ξ and N_i stands for the set of neighbors (x_j) falling into a window around x_i. The parameter a plays the same role as before, that is to control the effect of the neighbors term. Then, the clustering method [13] is performed on the gray level histogram of the newly generated image ξ. As a consequence, the objective function in this case is defined as

$$J_m = \sum_{i=1}^{M}\sum_{j=1}^{c} \gamma_i u_{ji}^m (\xi_i - v_j)^2 \quad (11)$$

where v_j represents the prototype of the jth cluster, u_{ji} represents the fuzzy membership of gray level value i with respect to cluster j, M denotes the number of gray levels of image ξ, which is generally much smaller than N, and γ_i is the number of pixels having gray level value equal to i. Thus, $\sum_{k=1}^{M} \gamma_k = N$ and under the constraint that $\sum_{j=1}^{c} u_{ji} = 1$ for any i, the J_m (11) is minimized, using the following equations for calculation of the membership partition matrix and the cluster centers:

$$u_{ji} = \frac{(\xi_i - v_j)^{2/m-1}}{\sum_{k=1}^{c}(\xi_i - v_k)^{2/m-1}} \quad (12)$$

$$v_j = \frac{\sum_{i=1}^{M} \gamma_i u_{ji}^m \xi_i}{\sum_{i=1}^{M} \gamma_i u_{ji}^m}. \quad (13)$$

The iterative process of the EnFCM algorithm is similar to FCM, but it is applied to the new image ξ (10) by using (12) and (13).

Due to fact that the gray level value of the pixels is generally encoded with 8 bit resolution (256 gray levels), the number M of gray levels is generally much smaller than the size N of the image. Thus, the execution time is significantly reduced.

EnFCM provides comparable segmenting results to FCM_S, but the segmenting quality depends on the chosen window size, the parameter a and the filtering method. If the parameter a is chosen large enough, then the method is resistant to noise, but, on the other hand, when a is chosen small enough the segmented image maintain its sharpness and details. However, when there is no prior knowledge about the image noise, the selection of parameter a is not an easy task and has to be made by experience or by using the trial-and-error method.

D. Fast Generalized Fuzzy C-Means Clustering (FGFCM)

Cai *et al.* [16] proposed the fast generalized fuzzy c-means (FGFCM) algorithm to improve the clustering results, as well as to facilitate the choice of the neighboring control parameter. In order to improve the clustering results FGFCM exploits a local similarity measure that combine both spatial and gray level image information, in terms of

$$S_{ij} = \begin{cases} e^{-\max(|p_i-p_j|,|q_i-q_j|)/\lambda_s - \|x_i-x_j\|^2/\lambda_g\sigma_i^2}, & i \neq j \\ 0, & i = j, \end{cases} \tag{14}$$

where the ith pixel is the center of the local window and the jth pixel represents the set of the neighbors falling into the window around the ith pixel. (p_i, q_i) are the coordinates of pixel i and x_i is its gray level value. λ_s and λ_g are two scale factors playing a role similar to factor a in EnFCM, and σ_i is defined as

$$\sigma_i = \sqrt{\frac{\sum\limits_{j\in N_i} \|x_i - x_j\|^2}{N_R}}. \tag{15}$$

FGFCM incorporates local and gray level information (14) into its objective function generating in advance a new image ξ as follows:

$$\xi_i = \frac{\sum\limits_{j\in N_i} S_{ij}x_j}{\sum\limits_{j\in N_i} S_{ij}} \tag{16}$$

where ξ_i denotes the gray level value of the ith pixel of the image ξ, x_j represents the gray level value of the neighbors of x_i (window center), N_i is the set of neighbors falling in the local window and S_{ij} is the local similarity measure between the ith and the jth pixel.

Furthermore, Cai *et al.* [16] proposed two variants of the FGFCM algorithm, the FGFCM_S1 and the FGFCM_S2, which incorporate modified similarity measures in their objective functions. FGFCM_S1 uses $S_{ij} = 1$ for all i and j, and naturally ξ_i is equal to the mean of the neighbors within a specified window, including the ith pixel. On the other hand, FGFCM_S2, uses as similarity measure the $S_{ij} = \text{median}\{x_j\}$, that is, ξ_i is the median of the neighbors within a specified window, including the ith pixel.

The FGFCM algorithm can be summarized as follows.

1) Set values for c, m, and ε.
2) Compute the new image ξ in terms of (16).
3) Initialize the fuzzy partition matrix $U^{(0)}$, and set the loop counter $b = 0$.
4) Update the cluster centers $v_j^{(b)}$ using (13).
5) Update the membership matrix $U^{(b+1)}$ using (12).
6) If $\max\left\{U^{(b)} - U^{(b+1)}\right\} < \varepsilon$ then stop, otherwise, set $b = b + 1$ and go to step 4.

The significant reduction of execution time attributes to taking into account a range of distribution of gray levels of the given image, as in EnFCM algorithm.

FGFCM and its variants, provide good enough segmenting results, but the segmenting quality depends on the chosen window size and the parameters λ_s and λ_g. Parameter λ_s could be fixed to 3 [16]. Thus, parameter λ_g controls the algorithm. If λ_g is chosen large enough, then the method is resistant to noise, but on the other hand, when λ_g is chosen small enough the segmented image maintain its sharpness and details. However, because there is no prior knowledge about the image noise, the selection of parameter λ_g is not an easy task, and has to be made by experience or by using the trial-and-error method.

III. Fuzzy Local Information C-Means (FLICM) Clustering Algorithm

Motivated by individual strengths of FCM_S1, FCM_S2, EnFCM, and FGFCM and its variations, we propose, in this paper, a novel and robust FCM framework for image clustering called Fuzzy Local Information C-means (FLICM) clustering algorithm.

A. Introducing the Fuzzy Factor G

All the methods described in the previous Section have yielded effective clustering results for images [12], [13], and [16], but still have some disadvantages. 1) Although the introduction of local spatial information enhances their insensitiveness to noise to some extent, they still lack enough robustness [18]–[20] to noise and outliers, especially in absence of prior knowledge of the noise. 2) There is a crucial parameter a (or λ) in their objective functions, used to balance between robustness to noise and effectiveness of preserving the details of the image. Generally, its selection has to be made by experience or trial and error experiments. 3) They are all applied on a static image, which has to be computed in advance. Details of the original image could be lost depending on the method used to generate the new image.

In order to overcome the above mentioned disadvantages a new factor in FCM objective function is needed. The new factor should have some special characteristics:

- to incorporate local spatial and local gray level information in a fuzzy way in order to preserve robustness and noise insensitiveness;
- to control the influence of the neighborhood pixels depending on their distance from the central pixel;
- to use the original image avoiding preprocessing steps that could cause detail missing;
- to be free of any parameter selection.

So, we introduce the novel fuzzy factor G_{ki} defined as

$$G_{ki} = \sum_{\substack{j\in N_i \\ i\neq j}} \frac{1}{d_{ij}+1}(1-u_{kj})^m \|x_j - v_k\|^2 \tag{17}$$

where the ith pixel is the center of the local window (for example, 3×3), k is the reference cluster and the jth pixel belongs in the set of the neighbors falling into a window around the ith pixel (N_i). $d_{i,j}$ is the spatial Euclidean distance between pixels i and j, u_{kj} is the degree of membership of the jth pixel in the kth cluster, m is the weighting exponent on each fuzzy membership, and v_k is the prototype of the center of cluster k.

It is easy to see that the factor G_{ki} is completely free of using any parameter that controls the balance between the image noise and the image details. The control of this balance is automatically achieved by the definition of the fuzziness of each image pixel (both spatial and gray level). Also, by using d_{ij}, the factor G_{ki} makes the influence of the pixels within the local window, to change flexibly according to their distance from the central pixel. Thus, more local spatial information can be used. It is worth indicating that the shape of the local window used in our experiments is square, but also, windows with other shapes such as diamond or circle can easily be adopted to the algorithm. As a whole, G_{ki} reflects the damping extent of the neighbors with the spatial distances from the central pixel. In contrast, the parameter a (or λ) in FCM_S, EnFCM, FGFCM, and their variants, is globally taken as a constant and, thus, it is relatively difficult to vary adaptively with different spatial locations or distances from the central pixel. Moreover, there is no need of preprocessing steps to apply the algorithm, as it will be shown in the following. The important role of G_{ki} during the application of the algorithm will also be shown in the following subsection.

B. General Framework of FLICM

By using the definition of G_{ki}, we now propose a robust FCM framework for image clustering, named Fuzzy Local Information C-Means (FLICM) clustering algorithm. It incorporates local spatial and gray level information into its objective function, defined in terms of

$$J_m = \sum_{i=1}^{N} \sum_{k=1}^{c} \left[u_{ki}^m \|x_i - v_k\|^2 + G_{ki} \right]. \tag{18}$$

The two necessary conditions for J_m to be at its local minimal extreme, with respect to u_{ki} and v_k is obtained as follows:

$$u_{ki} = \frac{1}{\sum_{j=1}^{c} \left(\frac{\|x_i - v_k\|^2 + G_{ki}}{\|x_i - v_j\|^2 + G_{ji}} \right)^{1/m-1}} \tag{19}$$

$$v_k = \frac{\sum_{i=1}^{N} u_{ki}^m x_i}{\sum_{i=1}^{N} u_{ki}^m}. \tag{20}$$

Thus, the FLICM algorithm is given as follows.

Step 1. Set the number c of the cluster prototypes, fuzzification parameter m and the stopping condition ε.
Step 2. Initialize randomly the fuzzy partition matrix.
Step 3. Set the loop counter $b = 0$.
Step 4. Calculate the cluster prototypes using (20).
Step 5. Compute membership values using (19).
Step 6. $\max\left\{U^{(b)} - U^{(b+1)}\right\} < \varepsilon$ then stop, otherwise, set $b = b + 1$ and go to step 4.

When the algorithm has converged, a defuzzification process takes place in order to convert the fuzzy partition matrix U to a crisp partition. The maximum membership procedure is the most important method that has been developed to defuzzify the partition matrix U. This procedure assigns the pixel i to the class C with the highest membership

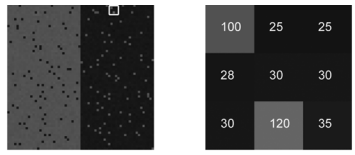
$$C_i = \arg_k \{\max\{u_{ki}\}\}, \quad k = 1, 2, \ldots, c. \tag{21}$$

It is used to convert the fuzzy image achieved by the proposed algorithm to the crisp segmented image.

The measure used in the FLICM objective function (18) is still the Euclidean metric as in FCM, which is computationally simple. Moreover, differently from FCM, FLICM is robust because of the introduction of the factor G_{ki}, which can be analyzed as follows.

The noise tolerance and outliers resistance property, completely relies on the definition of G_{ki}, as it is seen in (18). G_{ki} is automatically determined rather than artificially set, even in the absence of any prior noise knowledge. Two basic cases which describe the performance of the algorithm when outliers are present in the window will be presented in the following. As it will be shown the G_{ki} of the noise-corrupted pixels within a window will be kept to similar value to the central pixel ignoring the influence of the noise. The G_{ki} value will adaptively change in every iteration, converging to the central pixel's value and thus preserving the insensitiveness to noise and outliers.

- Case 1: The central pixel is not a noise and some pixels within its local window may be corrupted by noise. An example illustrated in Fig. 1 depicts this situation, in which a 3×3 window was used. This window was extracted from the noisy image (marked with a rectangle) shown on the left of the top row of the Fig. 1. It is clearly shown that after five iterations the algorithm converges and the corresponding membership values of the noisy, as well as of the no-noisy pixels converge to a similar value, ignoring the noisy pixels [Fig. 1(a)–1(d)]. The neighboring pixels, where their corresponding windows are intercovered, are examined as well. Generally, in such cases, the gray level values of the noisy pixels are far different from the other pixels within the window, and thus the factor G_{ki} balances their membership values. Therefore, the combination of the spatial and the gray level constraints incorporated in the G_{ki} suppress the influence of the noisy pixels, and, hence, the algorithm becomes more robust to outliers.
- Case 2: The central pixel is corrupted by noise, while the other pixels within its local window are homogenous, not corrupted by noise. Such an example is clearly demonstrated in Fig. 2. Again a 3×3 window was used, which was extracted from the noisy image (marked with a rectangle) shown on the left of the top row of the Fig. 2. It is shown that after five iterations the membership value of the noisy (central) pixel converges to a similar to neighboring pixels membership value, ignoring in this way the potential influence by noise, as shown in Fig. 2(d). Generally, in such cases the factor G_{ki} balances the membership value of the central pixel taking into account the spatial, as well as the gray level of the no-noisy neighboring pixels in a



0.51	0.97	0.02
0.01	0.85	0.66
0.58	0.38	0.44

$v_L = 65.60$
$v_R = 63.94$
(a)

0.58	0.48	0.62
0.58	0.46	0.52
0.47	0.49	0.44

$v_L = 68.06$
$v_R = 61.91$
(b)

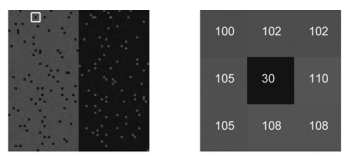
0.56	0.79	0.83
0.72	0.73	0.79
0.73	0.43	0.76

$v_L = 95.88$
$v_R = 33.87$
(c)

0.77	0.99	0.99
0.95	0.94	0.97
0.96	0.65	0.96

$v_L = 98.05$
$v_R = 31.77$
(d)

Fig. 1. 3×3 window with noise (marked with a rectangle in the initial image), their corresponding membership values and the cluster centers (v_L and v_R). (a) The initial membership values, (b) after one iteration, (c) after three iterations, and (d) after five iterations.



0.61	0.93	0.20
0.78	0.95	0.69
0.64	0.86	0.35

$v_L = 65.60$
$v_R = 63.94$
(a)

0.55	0.57	0.53
0.57	0.64	0.52
0.56	0.56	0.58

$v_L = 68.06$
$v_R = 61.91$
(b)

0.19	0.21	0.19
0.22	0.44	0.23
0.21	0.23	0.21

$v_L = 95.88$
$v_R = 33.87$
(c)

0.01	0.01	0.01
0.01	0.20	0.01
0.01	0.01	0.01

$v_L = 98.05$
$v_R = 31.77$
(d)

Fig. 2. 3×3 window with noise (marked with a rectangle in the initial image), their corresponding membership values and the cluster centers (v_L and v_R). (a) The initial membership values, (b) after one iteration, (c) after three iterations, and (d) after five iterations.

fuzzy manner. Thus, the proposed method becomes more robust to outliers, since the membership value of the central pixel is not influenced by noise.

The above two examples just give some intuitive illustrations about the robustness of our algorithm. The enhancement of its robustness to noise and outliers is based on the incorporation of the G_{ki} with fuzzy spatial and gray level localities constraints (18).

Another issue that is worth to point out, is the denoising potential of the proposed method in comparison to other methods. FCM_S1 uses a mean-type filtering, so it is relatively suitable for noisy images corrupted by Gaussian noise, whilst FCM_S2 uses a median-type filtering, and as a consequence, it is relatively suitable for images corrupted by impulsive noise. Also, in both cases, the final effectiveness of ignoring the noise in clustering, relies on the value of the parameter a. Thus, it is generally hard to choose the proper method (FCM_S1 or FCM_S2) and the optimal parameter a for good clustering results. On the other hand, FGFCM is independent of the noise type, but its clustering results depend also on parameter λ. Parameter λ has a relatively small value range, and one can reach more easily than FCM_S1 and FCM_S2 the correct parameter. But its selection still needs experience or usage of the trial-and-error method. Compared with FCM_S1, FCM_S2 or FGFCM, FLICM is independent of the type of the noise and completely free of any parameter selection or use. Its denoising performance is shown in the experimental results section.

The factor G_{ki} also seems to preserve more image information. It incorporates both local spatial and gray level relationship. The local spatial relationship changes adaptively according to spatial distances from the central pixel. The local gray level relationship not only varies automatically according to different gray level difference between the pixels over an image, but also is dependent on their fuzzy membership values. Thus, the value of G_{ki} varies from pixel to pixel, as well as from iteration to iteration within a neighborhood window, which likely preserves more information than using the same values for each pixel and iteration. Therefore, FLICM adopting G_{ki} seems able to preserve more image details than the other methods.

The major characteristics of the FLICM are summarized below:

- it provides noise-immunity;
- it preserves image details;
- it is free of any parameter selection;
- it is applied on the original image.

IV. Experimental Results

In this section, we show the performance of the proposed method by presenting numerical results and examples on various synthetic and real images, with different types of noise and characteristics. Furthermore, we compare the efficiency and the robustness of FLICM with six fuzzy algorithms FCM_S1, FCM_S2, EnFCM, FGFCM_S1, FGFCM_S2, FGFCM, and two well-known nonfuzzy algorithms, k-means [21] and hierarchical clustering based on the SLINK algorithm [22]. The denoising performances of the above nine algorithms were compared with respect to the optimal segmentation accuracy

(SA), where SA is defined as the sum of the correctly classified pixels divided by the sum of the total number of pixels [9]

$$SA = \sum_{i=1}^{c} \frac{A_i \cap C_i}{\sum_{j=1}^{c} C_j} \tag{22}$$

where c is the number of clusters, A_i represents the set of pixels belonging to the ith class found by the algorithm, while C_i represents the set of pixels belonging to the ith class in the reference segmented image.

In our numerical experiments, we generally choose the parameters to be $\lambda_s = 3$, $\varepsilon = 0.00001$ and $N_R = 8$ (a 3×3 window centered around each pixel, except the central pixel itself) [16].

First, we apply these algorithms to a synthetic test image (Fig. 3(a): 128×128 pixels, two classes with two gray level values taken as 20 and 120) corrupted by different levels of Gaussian, Uniform and Salt & Pepper noise, respectively. The number of clusters is set equal to $c = 2$. Also, parameter λ_g is set equal to $\lambda_g = 6.0$ for FGFCM and its variants (obtained by searching the interval of $[0.5, 6]$ with respect to SA). The parameter a is chosen equal to $a = 4.2$ in all the algorithms FCM_S1, FCM_S2 and EnFCM [12], which is obtained by seeking the interval $[0.2, 8]$.

Fig. 3 illustrates the clustering results of a corrupted by gaussian noise (20%) image. FGFCM_S1 [Fig. 3(f)], FGFCM_S2 [Fig. 3(g)], FGFCM [Fig. 3(h)], k-means [Fig. 3(i)], and SLINK [Fig. 3(k)] are respectively affected by the noise to different extents, which indicates that these algorithms lack enough robustness to the gaussian noise. Visually, FCM_S1 [Fig. 3(c)], FCM_S2 [Fig. 3(d)], and EnFCM [Fig. 3(e)] remove most of the noise, but still their results are not satisfactory enough. On the other hand, FLICM [Fig. 3(k)] removes almost all the added noise achieving satisfactory results, fact that is verified by the segmentation accuracy (SA) results shown in Table I.

Table I gives the average segmentation accuracy results of the nine algorithms on the specific synthetic image corrupted respectively by different noises with different levels. Each experiment has been performed using five different random initializations and the typical one has been assumed as the result. It is clearly illustrated that the proposed FLICM algorithm gives rise to better denoising performance than FCM_S type, EnFCM, FGFCMs, and nonfuzzy algorithms, presenting robustness to all considered kind of noises against to the other eight algorithms.

Furthermore, we apply the nine clustering algorithms to the real image *eight* [23] [Fig. 4(a)], contaminated with salt & pepper (30%) [Fig. 4(b)]. The clustering results are shown in Fig. 4(c)–4(k). The parameters selected for this experiment are $c = 3$, $a = 1.8$ and $\lambda_g = 6.0$. It is clearly illustrated in Fig. 4(c)–4(j) that FCM_S1, FCM_S2, EnFCM, FGFCM, and its variants and the two nonfuzzy algorithms, are all influenced by the noise to different extents, which indicates that these algorithms lack enough robustness to the salt & pepper noise, while the proposed method FLICM [Fig. 4(k)] can basically eliminate the effect of the noise. It is also worth noting that

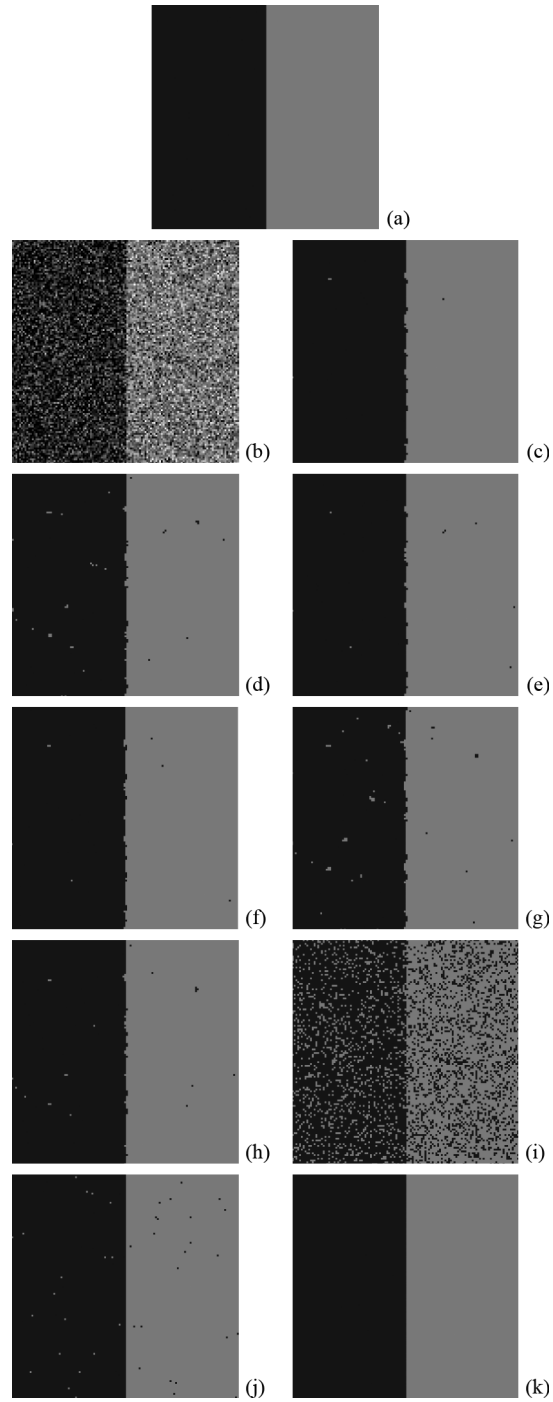


Fig. 3. Clustering of a synthetic image. (a) Original image, (b) the same image with gaussian noise (20%), (c) FCM_S1 result, (d) FCM_S2 result, (e) EnFCM result, (f) FGFCM_S1 result, (g) FGFCM_S2 result, (h) FGFCM result, (i) k-means result, (j) SLINK result, and (k) FLICM result.

the selection of the parameters a and λ has been performed by trial-and-error method selecting those with the smallest optimal segmentation accuracy (SA) error.

Besides, we also applied the same nine algorithms on the same image (*eight* [23]), as well as on a number of other images (e.g., Figs. 3–5). All the test images were corrupted by Gaussian, Uniform and Salt & Pepper noise at different levels: 3%, 5%, 8%, 10% to 35% with step 5%. Parameters a for FCM_S1, FCM_S2 and EnFCM, as well as parameter λ_g for FGFCM

TABLE I
SEGMENTATION ACCURACY (SA%) OF NINE ALGORITHMS ON SYNTHETIC IMAGES

	FCM_S1	FCM_S2	EnFCM	FGFCM_S1	FGFCM_S2	FGFCM	k-means	SLINK	FLICM
Gaussian 8%	99.28	99.71	99.30	97.58	99.81	98.62	92.08	99.77	**99.89**
Gaussian 10%	98.34	99.34	98.36	97.51	99.43	98.48	88.96	99.22	**99.64**
Gaussian 15%	97.21	98.04	97.25	97.04	97.72	97.60	79.93	98.49	**98.95**
Uniform 8%	99.98	99.98	**99.99**	97.65	99.92	98.68	82.88	81.92	99.98
Uniform 10%	99.49	99.76	99.52	97.61	99.78	98.56	81.55	68.05	**99.89**
Uniform 15%	96.54	95.67	96.26	96.41	94.25	95.92	80.01	57.92	**98.74**
Salt & Pepper 8%	99.60	99.62	99.60	97.66	74.90	88.12	92.82	99.29	**99.93**
Salt & Pepper 10%	99.45	99.53	99.45	97.51	74.89	86.69	92.78	99.17	**99.91**
Salt & Pepper 15%	99.10	99.46	99.10	97.34	74.88	83.33	91.31	98.68	**99.90**

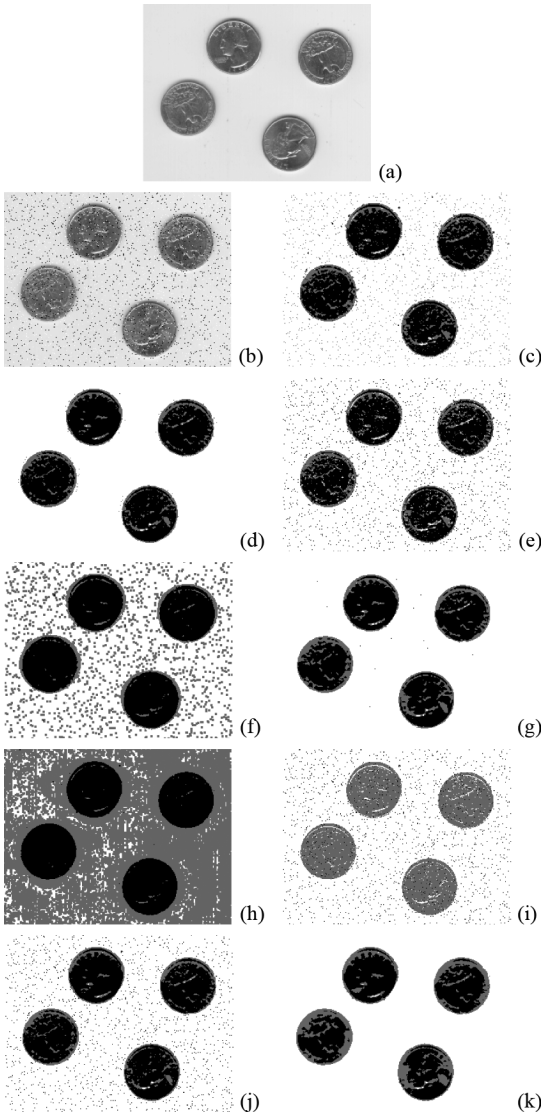


Fig. 4. Clustering results on eight image. (a) Original image, (b) the same image with salt & pepper noise (30%), (c) FCM_S1 result, (d) FCM_S2 result, (e) EnFCM result, (f) FGFCM_S1 result, (g) FGFCM_S2 result, (h) FGFCM result, (i) k-means result, (j) SLINK result, and (k) FLICM result.

and its variants were selected by performing the trial-and-error method and choosing those that maximizing the r quantitative index shown in (23). These results (Table II) can lead us to the conclusion drawn from the experimental results on the synthetic image. Each experiment has been performed using five different random initializations and the typical one has been assumed as the result.

TABLE II
COMPARISON SCORES (R%) OS NINE ALGORITHMS ON VARIOUS IMAGES

	Gaussian	Uniform	Salt & Pepper
FCM_S1	85.09	92.54	80.68
FCM_S2	86.50	93.12	80.05
EnFCM	86.72	92.37	80.77
FGFCM_S1	84.71	83.90	77.82
FGFCM_S2	85.19	83.33	73.09
FGFCM	86.01	63.07	64.32
k-means	76.56	79.08	67.96
SLINK	84.10	83.83	79.05
FLICM	**87.94**	**95.18**	**83.58**

Table II illustrates the clustering comparison of the nine algorithms calculating their scores using the following quantitative index [16], [24]:

$$r = \sum_{i=1}^{c} \frac{A_i \cap C_i}{A_i \cup C_i} \tag{23}$$

where c is the number of clusters, A_i represents the set of pixels belonging to the ith class found by the algorithm, while C_i represents the set of pixels belonging to the ith class in the reference segmented image. Index r is in fact a fuzzy similarity measure, indicating the degree of equality between A_i and C_i, and the larger the r is, the better the clustering is.

Furthermore, Fig. 6 shows some clustering results on real images [25]. The left column shows the initial images, while the right column depicts the clustering results as they were obtained by the proposed algorithm. The algorithm has been applied to each image using five different random initialization and every time the result was the same, that is, the one presented in Fig. 6. Figs. 3–6 and Table II illustrate that FLICM outperforms the other eight algorithms, which can attribute that the introduction of factor G_{ki} guarantees relative insensitivity both to noise and outliers.

Finally, Fig. 7 illustrates the average computational cost for each of the nine algorithms compared above. Each image size has been tested to five different images with random initializations and the cluster number varying from 2 to 5. The methods in the figure are presented in computational time descending order. The quickest method is FGFCM_S1, while the slowest is the SLINK. The proposed method is quite computational consuming, but this drawback is compensated for its very good performance as it was shown above. Furthermore, the proposed algorithm is easily programmed, since it does not contain any

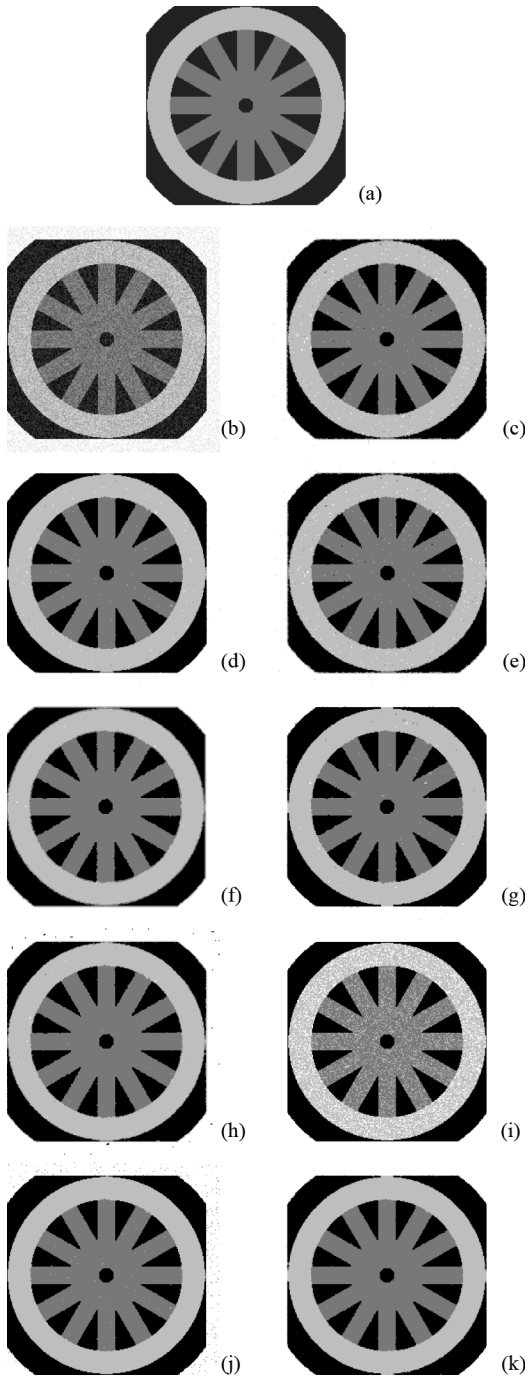


Fig. 5. Clustering results on wheel image. (a) Original image, (b) the same image with Gaussian noise (30%), (c) FCM_S1 result, (d) FCM_S2 result, (e) EnFCM result, (f) FGFCM_S1 result, (g) FGFCM_S2 result, (h) FGFCM result, (i) k-means result, (j) SLINK result, and (k) FLICM result.

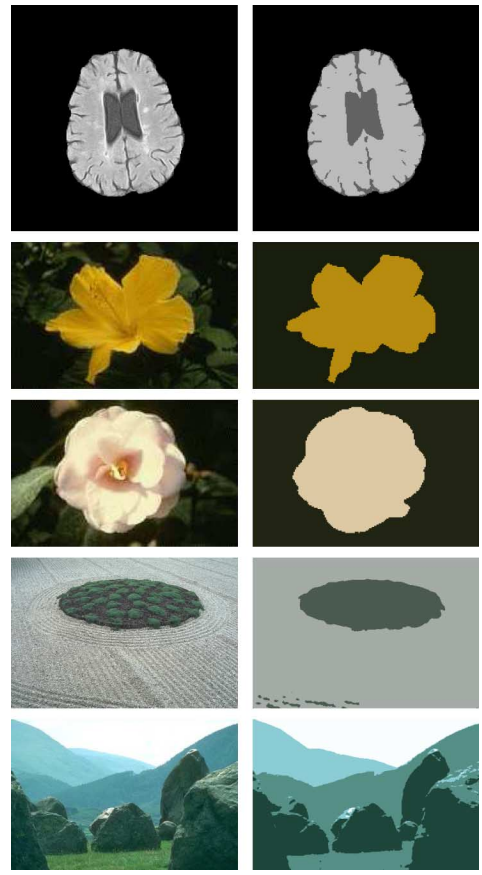


Fig. 6. Clustering results on real application image data by the proposed algorithm.

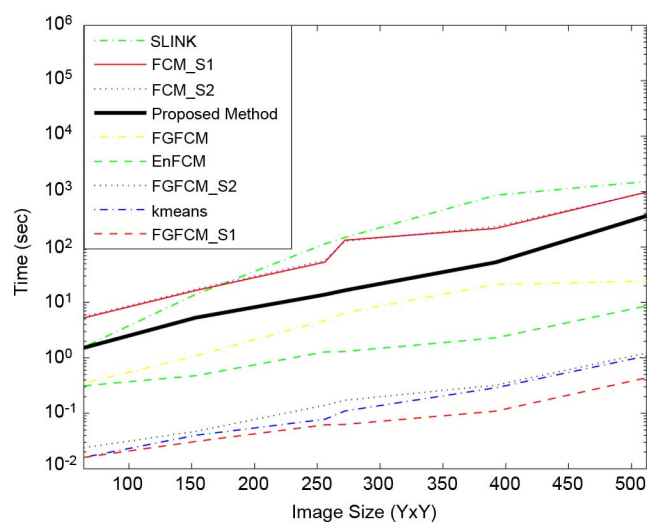


Fig. 7. Computational cost (in seconds) of the nine clustering algorithms.

complicated programming part. All experiments were performed on a Pentium IV (3 GHz) workstation under Windows XP Professional without any particular code optimization.

Since, the computational cost above presented is heavily influenced by the programming style, we also present a quick complexity analysis for all the tested algorithms. Taking into account the time complexity of the original FCM, which is $O(nc)$ (n is the histogram's length and c is the number of clusters), it is easily deduced that the total complexity of all FCM variations have the same complexity with a small variation depending on the preprocessing step each algorithm uses. On the other hand, the SLINK method has a complexity of $O(n^2)$, while the complexity for algorithms FCM_S1, FCM_S2 and the proposed one is $O(HWc)$, where H and W are the dimensions of the image under consideration.

V. Conclusion

In this paper, a novel robust fuzzy local information c-means (FLICM) algorithm for image clustering was introduced. The proposed algorithm can detect the clusters of an image overcoming the disadvantages of the known FCM algorithms and their variants. This is achieved by incorporating local spatial and gray level information. The FLICM introduces a new factor G_{ki} as a local (spatial and gray) similarity measure which aims to guarantee robustness both to noise and outliers. Also, the algorithm is relatively independent of the type of the added noise, and as a consequence, in the absence of prior knowledge of the noise, FLICM is the best choice for clustering. This is also enforced by the way that spatial and gray level image information are combined in the algorithm; the factor G_{ki} combines in a fuzzy manner the spatial and gray level information, rendering the algorithm more robust to all kind of noises, as well as to outliers. Furthermore, all the other fuzzy c-means algorithms for image clustering exploit, in their objective functions, a crucial parameter a (or λ), which is used to balance the robustness and effectiveness of ignoring the added noise. This parameter is mainly determined empirically or using the trial-and-error method. The FLICM is completely free of any parameter determination, while the balance between the noise and image details is automatically achieved by the fuzzy local constraints, enhancing concurrently the clustering performance. This is also enhanced, by the fact, that almost all the other methods perform the clustering on a precomputed image, while FLICM is applied on the original image.

References

[1] X. Munoz, J. Freixenet, X. Cufi, and J. Marti, "Strategies for image segmentation combining region and boundary information," *Pattern Recognit. Lett.*, vol. 24, no. 1, pp. 375–392, 2003.

[2] D. Pham, C. Xu, and J. Prince, "A survey of current methods in medical image segmentation," *Annu. Rev. Biomed. Eng.*, vol. 2, pp. 315–337, 2000.

[3] L. Zadeh, "Fuzzy sets," *Inf. Control*, vol. 8, pp. 338–353, 1965.

[4] J. Udupa and S. Samarasekera, "Fuzzy connectedness and object definition: Theory, algorithm and applications in image segmentation," *Graph. Models Image Process.*, vol. 58, no. 3, pp. 246–261, 1996.

[5] Y. Tolias and S. Panas, "Image segmentation by a fuzzy clustering algorithm using adaptive spatially constrained membership functions," *IEEE Trans. Syst., Man, Cybern.*, vol. 28, no. 3, pp. 359–369, Mar. 1998.

[6] J. Noordam, W. van den Broek, and L. Buydens, "Geometrically guided fuzzy C-means clustering for multivariate image segmentation," in *Proc. Int. Conf. Pattern Recognition*, 2000, vol. 1, pp. 462–465.

[7] M. Yang, Y. J. Hu, K. Lin, and C. C. Lin, "Segmentation techniques for tissue differentiation in MRI of ophthalmology using fuzzy clustering algorithms," *Magn. Res. Imag.*, vol. 20, no. 2, pp. 173–179, 2002.

[8] G. Karmakar and L. Dooley, "A generic fuzzy rule based image segmentation algorithm," *Pattern Recognit. Lett.*, vol. 23, no. 10, pp. 1215–1227, 2002.

[9] M. Ahmed, S. Yamany, N. Mohamed, A. Farag, and T. Moriarty, "A modified fuzzy C-means algorithm for bias field estimation and segmentation of MRI data," *IEEE Trans. Med. Imag.*, vol. 21, pp. 193–199, 2002.

[10] J. Bezdek, *Pattern Recognition With Fuzzy Objective Function Algorithms*. New York: Plenum, 1981.

[11] D. Pham, "An adaptive fuzzy C-means algorithm for image segmentation in the presence of intensity inhomogeneities," *Pattern Recognit. Lett.*, vol. 20, pp. 57–68, 1999.

[12] S. Chen and D. Zhang, "Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure," *IEEE Trans. Syst., Man, Cybern.*, vol. 34, pp. 1907–1916, 2004.

[13] L. Szilagyi, Z. Benyo, S. Szilagyii, and H. Adam, "MR brain image segmentation using an enhanced fuzzy C-means algorithm," in *Proc. 25th Annu. Int. Conf. IEEE EMBS*, 2003, pp. 17–21.

[14] M. Krinidis and I. Pitas, "Color texture segmentation based-on the modal energy of deformable surfaces," *IEEE Trans. Image Process.*, vol. 18, no. 7, pp. 1613–1622, Jul. 2009.

[15] D. Pham, "Fuzzy clustering with spatial constraints," in *Proc. Int. Conf. Image Processing*, New York, 2002, vol. II, pp. 65–68.

[16] W. Cai, S. Chen, and D. Zhang, "Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation," *Pattern Recognit.*, vol. 40, no. 3, pp. 825–838, Mar. 2007.

[17] J. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters," *J. Cybern.*, vol. 3, pp. 32–57, 1974.

[18] R. Hathaway, J. Bezdek, and Y. Hu, "Generalized fuzzy c-means clustering strategies using L norm distance," *IEEE Trans. Fuzzy Syst.*, vol. 8, pp. 576–582, Oct. 2000.

[19] K. Wu and M. Yang, "Alternative c-means clustering algorithms," *Pattern Recognit.*, vol. 35, no. 10, pp. 2267–2278, 2002.

[20] J. Leski, "Toward a robust fuzzy clustering," *Fuzzy Sets Syst.*, vol. 137, no. 2, pp. 215–233, 2003.

[21] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, 1967, vol. 1, pp. 281–297.

[22] R. Sibson, "SLINK: An optimally efficient algorithm for the single-link cluster method," *The Comput. J.*, vol. 16, no. 1, pp. 30–34, 1973.

[23] M. Mathworks and Natick, Image Processing Toolbox [Online]. Available: http://www.mathworks.com

[24] F. Masulli and A. Schenone, "A fuzzy clustering based segmentation system as support to diagnosis in medical imaging," *Artif. Intell. Med.*, vol. 16, no. 2, pp. 129–147, 1999.

[25] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th Int. Conf. Computer Vision*, Jul. 2001, vol. 2, pp. 416–423.

Stelios Krinidis was born in Kavala, Greece, in 1978. He received the B.Sc. degree in informatics in 1999 and the Ph.D. degree in informatics in 2004, both from the Aristotle University of Thessaloniki, Thessaloniki, Greece.

From 1999 to 2004, he was a researcher and teaching assistant in the Department of Informatics, University of Thessaloniki. From 2005 to 2008, he was a temporary Lecturer in the Department of Information Management, Technological Institute of Kavala, where he is currently a Senior Researcher. His current research interests include computational intelligence, pattern recognition, digital signal and 2-D and 3-D image processing and analysis, and computer vision.

Vassilios Chatzis received the Diploma of Electrical Engineering degree and the Ph.D. degree in informatics from the University of Thessaloniki, Thessaloniki, Greece, in 1991 and 1999, respectively.

Since 2002, he has been an Associate Professor with the Department of Information Management, Technological Institute of Kavala, Kavala, Greece. From 1993 to 1999, he served as a Researcher in the Artificial Intelligence and Information Analysis Laboratory, Department of Informatics, University of Thessaloniki. He also served as a Lecturer in the Electrical and Computer Engineering Department, University of Thrace, Thrace, Greece, from 2001 to 2002. His current research interests are in the areas of digital image processing, image retrieval, and fuzzy logic.

Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.201852605
Vol. 5, No. 2, Mei 2018, hlm. 171-176 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 51/E/KPT/2017 e-ISSN: 2528-6579

PENERAPAN ALGORITMA *FUZZY C MEANS* UNTUK ANALISIS PERMASALAHAN SIMPANAN WAJIB ANGGOTA KOPERASI

Risma Rustiyan R[1], Mustakim[2]

[1,2]Puzzle Research Data Technology, Fakultas Sains dan Teknologi
[1,2]Program Studi Sistem Informasi, Fakultas Sains dan Teknologi
Universitas Islam Negeri Sultan Syarif Kasim Riau
Email: [1]risma.rustiyan.r@students.uin-suska.ac.id , [2]Mustakim@uin-suska.ac.id

(Naskah masuk: 15 Januari 2018, diterima untuk diterbitkan: 3 Mei 2018)

Abstrak

Koperasi mempunyai peranan penting bagi perekonomian Indonesia**.** Perkembangan koperasi di Indonesia saat ini cukup pesat, pada data Badan Pusat Stastitik 3 tahun terakhir yang di-*update* pada tanggal 20 Juni 2016 sementara menyebutkan jumlah koperasi aktif di Indonesia pada tahun 2015 sebanyak 150.223. Pusat Koperasi Unit Desa (PUSKUD) Provinsi Riau merupakan salah satu jenis koperasi sekunder dalam bidang pertanian. Koperasi ini menjadi salah satu penunjang perekonomian rakyat. Proses yang berjalan antara PUSKUD dan anggota adalah pengelolaan sisa hasil usaha (SHU) dan dalam permodalan. Permasalahan yang terjadi adalah partisipasi aktif dari anggota koperasi untuk menunaikan kewajiban masih kurang dari yang diharapkan. Untuk membantu mengatasi permasalahan tersebut, perlu dibentuk suatu pengelompokan Anggota PUSKUD berdasarkan Kabupaten/ Kota domisili dalam pembayaran Simpanan Wajib. Metode yang digunakan adalah *Data Mining Clustering* dengan algoritma *Fuzzy C Means*. Dari hasil pengklusteran, pada tahap akhir analisis diketahui, Terdapat 75 anggota yang tersebar pada wilayah Kabupaten/ Kota Rokan Hulu, Kampar, Indragiri Hulu dan Indragiri Hilir serta terdaftar pada tahun 80-an yang perlu untuk ditinjau kembali. Hasil Pengujian Nilai Validitas PC, didapatkan sebesar 0,323732, dengan demikian kualitas *Cluster* masih jauh untuk mencapai kata optimal.

Kata kunci: *Clustering, Data Mining, Fuzzy C Means, Koperasi, Simpanan Wajib, Partition Coefficient*

IMPLEMENTATION OF FUZZY C MEANS ALGORITHM FOR ANALYSIS THE DEPOSIT PROBLEM OF COOPERATIVE MEMBERS

Abstract

Cooperatives have an important role for the Indonesian economy. The development of cooperatives in Indonesia is currently quite rapid, On the data from Badan Pusat Statistik (BPS) of the last 3 years updated on June 20, 2016 While mentioning the number of active cooperatives in Indonesia in 2015 as much as 150,223. Pusat Koperasi Unit Desa (PUSKUD) Provinsi Riau is one of the secondary cooperative in agriculture. This cooperative became one of the supporting people's economy. The process that runs between PUSKUD dan members is the management of the remaining results of the business and the capital. The problem that occurs is the use of members of the cooperative to fulfill the obligations is still less than expected . To help overcome these problems, it is necessary to form a grouping of PUSKUD Members by Regency / City domicile in the payment of Mandatory Deposits. The method used is Data Mining Clustering with Fuzzy C Means algorithm. From the results of the clustering, at the final stage of the analysis is known, There are 75 members scattered in the District / City of Rokan Hulu, Kampar, Indragiri Hulu dan Indragiri Hilir dan registered in the 80s that need to be reviewed. Test Result Validity Value of Partition Coefficient, obtained for 0,323732, thus the quality of Cluster is still far to achieve optimal.

Keywords: *Clustering, Cooperative, Data Mining, Fuzzy C Means, Mandatory Deposit, Partition Coefficient*

1. PENDAHULUAN

Koperasi mempunyai peranan penting bagi perekonomian Indonesia. Peranan itu ditunjukan saat adanya krisis ekonomi pada tahun 1998, yang mana Badan Usaha Milik Negara (BUMN), Badan Usaha Milik Daerah (BUMD) dan Badan Usaha Milik Swasta (BUMS) mengalami posisi yang sangat terpuruk, dan saat itulah Koperasi menjadi penopang ekonomi Indonesia (Santoso, 2004 dalam Ubaidillah, 2016). *International Cooperative Alliances* (ICA)

mengartikan Koperasi sebagai suatu perkumpulan yang terdiri dari orang-orang yang bergabung secara sukarela untuk mencapai tujuan bersama, yakni memenuhi kebutuhan dan aspirasi ekonomi, sosial, budaya dari anggota melalui perusahaan yang dikendalikan secara demokratis (Burhanuddin, 2005 dalam Ubaidillah, 2016).

Perkembangan Koperasi di Indonesia saat ini cukup pesat, dapat dibuktikan pada data Badan Pusat Statistik (BPS) 3 tahun terakhir yang di-*update* pada tanggal 20 Juni 2016 sementara menyebutkan pada tahun 2013 jumlah Koperasi aktif di Indonesia sebanyak 143.177, 147.249 pada tahun 2014 dan pada tahun 2015 sebanyak 150.223. Jumlah tersebut merupakan akumulasi dari 5 jenis Koperasi berdasarkan lapangan usaha yakni Koperasi Konsumsi, Koperasi Produksi, Koperasi Simpan Pinjam, Koperasi Serba Usaha dan Koperasi Jasa, kemudian 3 jenis Koperasi berdasarkan lingkungan usaha yakni Koperasi Fungsional, Koperasi Unit Desa (KUD) dan Koperasi Sekolah (Feryanto, 2011).

PUSKUD Provinsi Riau merupakan salah satu jenis Koperasi sekunder dalam bidang pertanian. Pada Provinsi Riau sektor pertanian menjadi salah satu penunjang perekonomian rakyat, khususnya pertanian kelapa sawit. Untuk anggota dari PUSKUD ini adalah KUD yang ada di seluruh Kabupaten Provinsi Riau, Kepulauan Riau (Kepri) dan beberapa berada di luar wilayah Riau dan Kepri, yang mendaftarkan diri sebagai anggota. Total KUD yang terdaftar menjadi anggota PUSKUD ada sebanyak 324 pada tahun 2014, 325 pada tahun 2015 dan 2016. Koperasi ini tersebar pada 12 Kabupaten/ Kota di Provinsi Riau dan Kepulauan Riau.

Proses bisnis yang dijalankan PUSKUD Riau adalah distribusi pupuk subsidi, pupuk non subsidi dan usaha-usaha terobosan atau dapat dikatakan sebagai kerjasama dengan mitra-mitra usaha. Selain itu proses yang berjalan antara PUSKUD dan KUD adalah pengelolaan sisa hasil usaha (SHU) dan dalam permodalan, yakni pengelolaan Simpanan Pokok dan Simpanan Wajib.

Simpanan Pokok adalah sejumlah uang yang harus disetorkan oleh anggota saat pertama kali terdaftar menjadi anggota Koperasi. Sedangkan Simpanan Wajib adalah simpanan yang harus dibayarkan oleh setiap anggota yang tempo waktunya mempunyai batasan (Rusdianto, 2010).

Menurut Data Simpanan Wajib tahun 2014-2016, partisipasi aktif dari anggota Koperasi seluruh Kabupaten di Provinsi Riau, Kepulauan Riau (Kepri) dan beberapa berada diluar wilayah Riau dan Kepri untuk menunaikan kewajiban masih kurang dari yang diharapkan. Berdasarkan penelitian yang dilakukan oleh Suskaniah dkk (2015) menyatakan bahwa Simpanan Pokok dan Simpanan Wajib berpengaruh tinggi terhadap Sisa Hasil Usaha (SHU), apabila Simpanan Pokok dan Simpanan Wajib dinaikkan maka akan meningkatkan nominal SHU. Namun yang menjadi pembahasan dalam penelitian ini adalah Simpanan Wajib karena nilainya akan berubah setiap tahunnya.

Untuk membantu mengatasi permasalahan tersebut perlu dibentuk suatu pengelompokan Anggota PUSKUD dalam pembayaran Simpanan Wajib. Teknik yang dapat digunakan dalam mengatasi permasalahan ini adalah *Data Mining*.

Data Mining merupakan proses dalam mendapatkan informasi berguna yang berasal dari basis data yang besar (Tan, 2006), hal ini bertujuan untuk membantu dalam pengambilan keputusan (Jindal *dkk*, 2014). Ada empat tugas pokok dalam *Data Mining* diantaranya Pengelompokkan, Klasifikasi, Regresi dan fungsi asosiasi (Sahu *dkk*. 2012). Metode ini nantinya akan membantu proses pengolahan data.

Salah satu metode yang digunakan dalam pengolahan data adalah pengelompokan/*clustering*. Metode ini memungkinkan untuk pengelompokkan *item* menjadi beberapa grup kecil yang mempunyai persamaan esensial (Nugraha *dkk*, 2014). Penelitian sebelumnya mengenai pengelompokkan terhadap anggota Koperasi pernah dilakukan yakni Penerapan *Data Mining* Klustering dengan Menggunakan Algoritma *K-Means* pada Data Nasabah Koperasi Simpan Pinjam dan Kredit pada Graha Mandiri Tegal (Santosa, 2015).

Penelitian *Clustering* ini dilakukan kepada Nasabah pada Koperasi Simpan Pinjam dan setelah ditelusuri belum ada yang melakukan penelitian dengan objek anggota PUSKUD khususnya pada bidang permodalan Simpanan Pokok dan Simpanan Wajib, padahal sedikit banyaknya hal ini dapat berpengaruh pada kelangsungan proses bisnis yang ada pada PUSKUD Riau. Oleh karena itu penelitian dilanjutkan dengan studi kasus diatas menggunakan algoritma *Fuzzy C Means* (FCM).

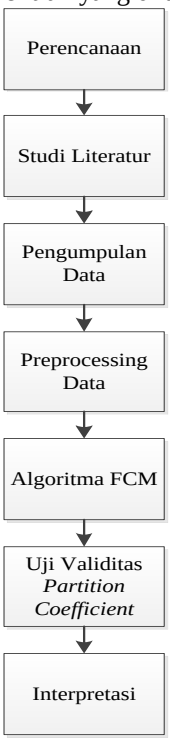
Algoritma ini merupakan salah satu metode *clustering* yang menjadi bagian dari *Hard K-Means*. FCM menggunakan model pengelompokkan *fuzzy* sehingga data tersebut dapat menjadi anggota dari semua klaster yang terbentuk dengan derajat atau tingkat keanggotaan yang berbeda yaitu antara 0 hingga tingkat keberadaan data dalam satu *cluster* ditentukan oleh derajat keanggotaanya (Klawonn, 2007 dalam Sumanto *dan* Wahono, 2011).

Adapun Kelebihan dari algoritma FCM adalah kemampuan hebat untuk mendeteksi *cluster* tingkat tinggi, kemudian dapat menunjukkan hubungan antar pola *cluster* yang berbeda (Sharma *dan* Borana, 2014). Kemudian tujuan dari penelitian ini adalah mendeskripsikan pola-pola hubungan antar atribut pada *cluster* yang menjadi penyebab adanya permasalahan pada Simpanan Wajib anggota.

2. METODE PENELITIAN

Metodologi penelitian ini ditunjukan pada Gambar 1., adapun penjelasannya sebagai berikut:

Pada tahap perencanaan dilakukan pemilihan topik penelitian yang akan dilakukan yaitu Klasterisasi anggota koperasi berdasarkan atribut simpanan wajib, kemudian tahap selanjutnya melakukan studi literatur, yaitu mencari rujukan referensi yang berhubungan dan dapat mendukung topik penelitian yang dilakukan. Langkah selanjutnya mengumpulkan data yang dibutuhkan yaitu data simpanan wajib anggota tahun 2015-2016 yang data tersebut kemudian dilakukan *Preprocessing* untuk memastikan data layak untuk digunakan, selanjutnya proses data mining menggunakan algoritma *Fuzzy C Means*, kemudian dilakukan uji validitas menggunakan *Partition Coefficient*, dan proses yang terakhir dilakukan interpretasi atau evaluasi berdasarkan hasil penelitian yang dilakukan.



Gambar 1. Metodologi Penelitian

2.1. Pengumpulan Data

Pada penelitian ini, data diperoleh dari PUSKUD Provinsi Riau tahun 2015-2016, setelah dilakukan tahapan *preprocessing* maka total data yang digunakan sebanyak 307 *record*.

2.2. *Preprocessing Data*

Tahap *preprocessing* diantaranya *data cleaning*, yaitu melakukan pemberisihan terhadap data yang memiliki kecacatan/ tidak layak digunakan diantaranya *noise*, *outlier* dan *missing value*. Terdapat 325 data awal yang digunakan, kemudian setelah dilakukan pembersihan (*data cleaning/ preprocessing)* didapatkan 18 data *missing value* sehingga data yang dapat diolah adalah sebanyak 307 record data. Kemudian melakukan proses transformasi dari data yang berbentuk teks menjadi bilangan agar dapat diolah menggunakan algoritma seperti yang ditunjukan pada Tabel 1.

Tabel 1. Transformasi Data

No	Atribut	Simbol	Keterangan
1.	Tahun Menjadi Anggota	1	Tahun 80-an
2.		2	Tahun 90-an
3.		3	Tahun 2000-an
4.	Kabupaten/ Kota	1	Bengkalis
5.		2	Indragiri Hilir
6.		3	Indragiri Hulu
7.		4	Kampar
8.		5	Kepulauan Riau
9.		6	Kota Madya Dumai
10.		7	Kota Pekanbaru
11.		8	Kuantan Singingi
12.		9	Lain-lain
13.		10	Pelalawan
14.		11	Rokan Hilir
15.		12	Rokan Hulu
16.		13	Siak

2.3. *Processing Data*

Pada tahap ini data diolah menggunakan algoritma Fuzzy C Means, langkah pertama yang dapat dilakukan adalah menentukan parameter yang akan digunakan. C menunjukan jumlah *cluster* yang akan dibentuk sesuai dengan kebutuhan pada penelitian ini jumlah *cluster* yang dibentuk sebanyak 5, W adalah bobot perpangkatan yang digunakan bobot yang digunakan 2, P^o merupakan fungsi objektif awal yaitu 0, Max Iter adalah maksimal iterasi yang akan dicapai dalam penelitian ini 10 kali maksimal iterasi dan (ε) adalah error terkecil yang diharapkan yaitu sebesar 0,1.

Langkah kedua yang dilakukan adalah membangkitkan bilangan acak/ *random* yang diinisialisasikan dengan U pada matriks partisi awal Tabel 2.

Tabel 2. Matriks U Partisi Awal

U1	U2	U3	U4	U5
0,1	0,2	0,3	0,2	0,2
0,1	0,1	0,4	0,2	0,2
0,2	0,3	0,1	0,3	0,1
0,3	0,2	0,2	0,1	0,2
0,2	0,2	0,2	0,1	0,3
0,2	0,2	0,2	0,3	0,1
0,2	0,1	0,3	0,2	0,2
0,3	0,2	0,3	0,1	0,1
...	...	...	...	...
0,5	0,1	0,2	0,1	0,1

Kemudian mencari nilai pusat *cluster/centroid* dengan menggunakan persamaan 3 yang dilakukan pada setiap nilai cluster 1-5.

Langkah selanjutnya setelah didapatkan hasil keseluruhan *centroid* yang diinisialisasikan dengan fitur v, w, x, y dan z seperti yang ditunjukan pada Tabel 3. Kemudian mencari nilai Fungsi Objektif yang menentukan kriteria berhenti.

Tabel 3. Keseluruhan *Centroid*

CLUSTER	FITUR V	FITUR W	FITUR X	FITUR Y	FITUR Z
1	0,2390	0,5225	0,6249	0,1638	0,6028
2	0,1869	0,5065	0,6560	0,2097	0,5849
3	0,2447	0,5711	0,6205	0,1915	0,5741
4	0,1990	0,5577	0,6477	0,2034	0,5842
5	0,2943	0,5291	0,6014	0,2014	0,5502

Pada iterasi 1 didapatkan nilai P1-P0 |42,574-0|=42,574 > ε (lebih besar dari nilai *error* terkecil) artinya proses perhitungan FCM diulangi dari langkah no 4 (hitung pusat *Cluster*).

Pada penelitian ini proses iterasi dilanjutkan sampai dengan iterasi ke-10 dengan nilai fungsi objektif yang didapatkan sebesar 0,136 artinya nilai fungsi objektif telah berada dibatas ambang *error* terkecil yang diharapkan (0,1), selain itu iterasi yang dilakukan juga sudah mencapai maksimum iterasi yakni 10 kali, dengan demikian proses iterasi pun dihentikan.

2.4. Algoritma *Fuzzy C Means*

Data Mining memiliki hubungan dengan bidang lain seperti sistem basis data, jaringan syaraf tiruan, *pattern recognition*, *data warehouse*, statistik, mesin pembelajaran dan pencarian kembali informasi (Abdullah *dkk*, 2017). Sahu *dkk.* (2012) membagi *data mining* kedalam empat tugas pokok salah satunya adalah *clustering*, yang merupakan teknik pembelajaran pengelompokan data yang serupa, algoritma *clustering* mengelompokkan sejumlah titik yang sama dan memiliki sifat yang sama (Rajagopal, 2011) algoritma yang terkenal pada teknik *clustering* adalah Fuzzy C Means.

. Menurut Klawonn (2007) dalam Sumanto dan Wahono (2011) *Fuzzy C-Means* (FCM) dikenal juga dengan *Fuzzy* ISODATA, algoritma ini merupakan salah satu metode *clustering* yang merupakan bagian dari metode *Hard K-Means*. FCM menggunakan model pengelompokkan *fuzzy* sehingga data dapat menjadi anggota dari semua kelas atau *cluster* yang terbentuk dengan derajat atau tingkat keanggotaan yang berbeda antara 0 hingga tingkat keberadaan data dalam satu *cluster* ditentukan oleh derajat keanggotaanya.

Kelebihan dari algoritma FCM adalah kemampuan hebat untuk mendeteksi *cluster* tingkat tinggi, kemudian dapat menunjukkan hubungan antar pola *cluster* yang berbeda (Sharma *dan* Borana, 2014).

Tahapan dari algoritma FCM adalah sebagai berikut (Klawonn, 2007 dalam Sumanto *dan* Wahono, 2011):

1. Input data yang akan di *cluster* X, berupa matriks berukuran n x m (n, jumlah sampel data; m, atribut setiap data). Xij, data sampel ke-i (i=1,2...n), atribut ke-j (j=1,2...m).
2. Tentukan:
   a. Jumlah Cluster : c;
   b. Pangkat : w;
   c. Maksimum Iterasi : MaxIter;
   d. *Error Rate* : ε_I;
   e. Fungsi Objektif Awal : $P^o = 0$;
   f. Iterasi Awal : t = 1;
3. Bangkitkan bilangan *random* µik, i=1,2...n; k=1,2...c; sebagai elemen-elemen matriks partisi awal U
   $$Q_i = \sum_k^c = 1\mu ik \quad (1)$$
   Dengan j=1,2...n
   Hitung:
   $$\mu_{ik=\frac{\mu_{ik}}{Q_i}} \quad (2)$$
4. Hitung pusat *cluster* ke-k: Vkj, dengan k=1,2...c; dan j=1,2...m (Yan, 1994 dalam Sumanto *dan* Wahono, 2011).
   $$V_{kj} = \frac{\sum_i^n=_1((\mu_{ik})^w * X_{ij})}{\sum_i^n=_1(\mu_{ik})^w} \quad (3)$$
5. Hitung fungsi objektif pada iterasi ke-t, Pt (Yan, 1994 dalam Sumanto *dan* Wahono, 2011).
   $$P_{t=\sum_i^n=_1\sum_k^c=_1([\sum_j^m=_1(X_{ij}-V_{kj})^2](\mu_{ik})^w)} \quad (4)$$
6. Hitung perubahan matriks partisi (Yan, 1994 dalam Sumanto *dan* Wahono, 2011)
   $$\mu_{ik} = \frac{[\sum_j^m=_1(X_{ij}-V_{kj})^2]^{\frac{-1}{w-1}}}{\sum_k^c=_1(X_{ij}-V_{kj})^2]^{\frac{-1}{w-1}}} \quad (5)$$
7. Cek kondisi berhenti:
   a. Jika: (|Pt-Pt-1|<ε_I) atau (t >MaxIter) maka berhenti
   b. Jika tidak : t = t + 1, ulangi langkah ke-4

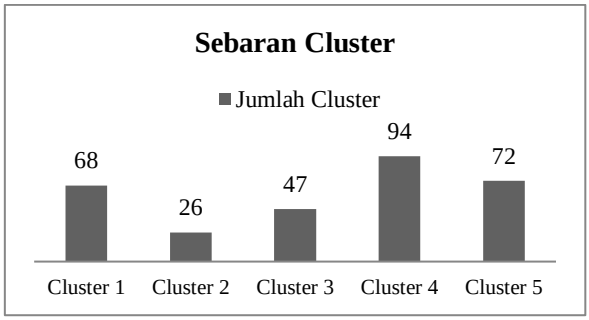
2.5. *Partition Coefficient*

Metode validitas ini untuk mengukur jumlah *overlapping* antar kelompok, rumus dari validitas *partition coefficient* ini adalah (Wang, 2007 dalam Haqiqi *dan* Kurniawan, 2015):

$$PC(c) = \frac{1}{N}\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^2 \quad (6)$$

N merupakan banyaknya jumlah objek dalam penelitian, c banyaknya kelompok, u_{ik}^2 banyaknya nilai keanggotaan objek ke-k dan *centroid* ke-i, index ini mempunyai rentang 1/c sampai 1. Jumlah kelompok yang optimal ditunjukkan dengan nilai PC yang besar (Haqiqi *dan* Kurniawan, 2015).

3. HASIL DAN PEMBAHASAN



Gambar 2. Sebaran Seluruh Cluster

Berdasarkan hasil perhitungan menggunakan algoritma *Fuzzy C Means*, didapatkan sebaran *Cluster* anggota PUSKUD Provinsi Riau dengan rincian, *Cluster* 1 diisi sebanyak 64 anggota, *Cluster* 2 sebanyak 26 anggota, *Cluster* 3 sebanyak 47 anggota, *Cluster* 4 sebanyak 94 anggota dan *Cluster* 5 sebanyak 72 anggota.

Karakteristik *Cluster* 1 pada Gambar 5.2, dapat dilihat hubunganya dengan tahun menjadi anggota diketahui bahwasannya anggota yang mendaftar tahun 80-an sebanyak 61 atau 31% dari total jumlah keseluruhan anggota yang terdaftar tahun 80-an, tahun 90-an sebanyak 7 atau 9,1% dari total jumlah keseluruhan anggota yang terdaftar tahun 90-an dan tahun 2000-an sebanyak 0, pada *Cluster* 1 ini didominasi oleh anggota yang terdaftar pada tahun 80-an.

Berdasarkan penjabaran analisis masing-masing *Cluster*, didapatkan kumpulan analisis *Cluster* 1-5 secara keseluruhan sebagai berikut:

1. Karakteristik *Cluster* 1 memiliki rata-rata pada tahun 80-an saat menjadi anggota PUSKUD Provinsi Riau, dengan Kabupaten/ Kota yang mendominasi yakni Kabupaten Indragiri Hulu dan Indragiri Hilir. Kemudian memiliki Tunggakan > Rp. 1.000.000 yang mendominasi.
2. Karakteristik *Cluster* 2 memiliki rata-rata tahun 80-an dan 90-an pada porsi yang seimbang, untuk wilayah yang mendominasi adalah kabupaten Kampar sedangkan pada Tunggakan > Rp. 100.000 ≤ Rp. 1.000.000 yang mendominasi.
3. Karakteristik *Cluster* 3 memiliki rata-rata pada tahun 80-an menjadi anggota PUSKUD Provinsi Riau, dengan Kabupaten/ Kota yang mendominasi adalah Siak kemudian Tunggakan yang mendominasi pada ≤ Rp. 0
4. Karakteristik *Cluster* 4 memiliki rata-rata tahun 80-an saat menjadi anggota PUSKUD Provinsi Riau, wilayah Kabupaten/ Kota yang mendominasi adalah Rokan Hulu dan memiliki Tunggakan > Rp. 1.000.000
5. Karakteristik *Cluster* 5 memiliki rata-rata tahun 90-an saat menjadi anggota PUSKUD Provinsi Riau, untuk wilayah domisili yang mendominasi adalah Kampar dengan Tunggakan > Rp. 1.000.000 yang mendominasi.

Kemudian diurutkan tingkat kepentingan hubungan permasalahan Simpanan Wajib anggota berdasarkan jumlah anggota dan besarnya tunggakan, karena semakin besar Tunggakan di sanalah letak permasalahan Simpanan Wajib anggota, maka dapat diperoleh hasil sebagai berikut:

Pada posisi pertama diduduki oleh *Cluster* 4 dengan anggota sebanyak 94, dengan perincian jumlah anggota 92 atau 46,7% dari total jumlah keseluruhan anggota yang terdaftar tahun 80-an, dengan wilayah yang mendominasi adalah Kabupaten/ Kota Rokan Hulu sebanyak 23 anggota atau 56,1% dari total jumlah keseluruhan anggota yang berada di Kabupaten Rokan Hulu dengan Tunggakan > Rp. 1.000.000.

Pada posisi kedua diduduki oleh *Cluster* 5 dengan anggota sebanyak 72, dengan perincian sebanyak 39 anggota atau 50,6% dari total jumlah keseluruhan anggota yang terdaftar tahun 90-an, dengan wilayah yang mendominasi adalah Kabupaten/ Kota Kampar sebanyak 20 anggota atau 44,4% dari total jumlah keseluruhan anggota yang berada di Kabupaten Kampar dengan Tunggakan > Rp. 1.000.000.

Pada posisi ketiga diduduki oleh *Cluster* 1 dengan anggota sebanyak 68, dengan perincian sebanyak 61 anggota atau 31% dari total jumlah keseluruhan anggota yang terdaftar tahun 80-an dengan wilayah yang mendominasi adalah Kabupaten/ Kota Indragiri Hulu dan Indragiri Hilir sebanyak 16 atau 66,7% dari total jumlah keseluruhan anggota yang berada di Kabupaten Indragiri Hulu, dan sebanyak 16 atau 76,2% dari total jumlah keseluruhan anggota yang berada di Kabupaten Indragiri Hilir dengan Tunggakan > Rp. 1.000.000.

Pada uji validitas *Partition Coefficient* (PC) Persamaan 7. Diketahui N =307 (jumlah data), sedangkan u_{ik}^2 adalah nilai perpangkatan dari nilai matriks partisi (nilai U baru). Adapun perhitungannya sebagai berikut:

U1	= $(0{,}9274)^2$	= 0,8601
U2	= $(0{,}0162)^2$	= 0,000263
U3	= $(0{,}0124)^2$	= 0,000153
U4	= $(0{,}0334)^2$	= 0,001112
U5	= $(0{,}0107)^2$	= 0,000113

Seluruh hasil perpangkatan adalah sebesar 99,38571, maka jika diolah menggunakan Validitas PC:

$$PC(c) = \frac{1}{307} * 99{,}38571 = 0{,}323732$$

4. KESIMPULAN

Berdasarkan hasil penelitian yang dilakukan yaitu klasterisasi anggota Koperasi menggunakan algoritma FCM dapat mengidentifikasi faktor-faktor yang berpengaruh terhadap permasalahan simpanan wajib anggota dengan parameter C = 5, w = 2 MaxIter = 10 dan ε = 0,1, namun setelah dilakukan uji validitas

menggunakan *Partition Coeffecient* didapatkan hasil 0,3 yang artinya kualitas klaster masih jauh dari kata optimal untuk itu disarankan untuk melakukan percobaan pada parameter jumlah Cluster (C) dalam beberapa versi misalnya jumlah cluster 3,4 dan 5 kemudian bobot (w) yang divariasikan menjadi 2,3 dan 4, kemudian untuk MaxIter juga sebaiknya dinaikan sampai batas maksimal 100 iterasi misalnya dan ε juga diusahakan seminimum mungkin sebesar 1e-6, setelah itu baru dapat dilakukan uji validitas .

5. DAFTAR PUSTAKA

ABDULLAH, USMAN., & EFENDI, M. Sistem Klasifikasi Kualitas Kopra Berdasarkan Warna dan Tekstur Menggunakan Metode Nearest Mean Classifier (NMC) *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*. Vol. 4, No. 4, 2017

FERYANTO, A. 2011. *Koperasi dan Perannya dalam Perekonomian*. Klaten : Saka Mitra Kompetensi, hlm 13-15

HAQIQI, B. N & KURNIAWAN, R. Analisis Perb*dan*ingan Metode Fuzzy C-Means Dan Subtractive Fuzzy C-Means. *Media Statistika*, [S.l.], v. 8, n. 2, p. 59-67, dec. 2015. ISSN 2477-0647

JINDAL, K., SHARMA, M., & SHARMA, DR. B. K. Data Mining to support Decision Process in Decision Support System. *International Journal of Emerging Technology dan Advanced Engineering*. Volume 4, Special Issue 1, 2014

KUSHRINI & LUTHFI, E. T. 2009. Algoritma Data Mining.Yogyakarta : *Dan*i

NUGRAHA, MITA, J. A,. & YUPIE, K,. *Data Mining dengan Metode Clustering untuk Pengolahan Informasi Persediaan Obat pada Puskesmas Pdananaran Semarang*. Semarang : Universitas Dian Nuswantoro 2014

RAJAGOPAL, S. Customer Data Clustering Using Data Mining Technique. *International Journal of Database Management Systems (IJDMS)* Vol.3, No.4, November 2011

RUSDIANTO.2010. *Akuntansi Koperasi*. Jakarta : Erlangga

SAHU,H., SHARMA, S. & GONDHALAKAR, S. A Brief Overview on Data Mining Survey. *International Journal of Computer Technology dan Electronics Engineering* (IJCTEE) Volume 1, Issue 3. 2012

SANTOSA, M. W. *Penerapan Data Mining Klustering dengan Menggunakan Algoritma K-Means pada Data Nasabah Koperasi Simpan Pinjam dan Kredit pada Graha Mdaniri Tegal*. Dokumen Karya Ilmiah. Tugas Akhir. TIF.2015

SHARMA, M., & BORANA, K. Clustering in Data Mining : A Brief Review. *International Journal Of Core Engineering & Management (IJCEM)* Volume 1, Issue 5, August 2014

SUMANTO & WAHONO, R. S. Penerapan Fuzzy C-Means dalam Pemilihan Peminatan Tugas Akhir Mahasiswa. *Seminar Nasional Inovasi Teknologi* Vol 1 No 1. 2011

SUSKANIAH, R. LCA,. JONATHAN, R., & LAU, E. A.Pengaruh Simpanan Pokok dan Simpanan Wajib terhadap Sisa Hasil Usaha pada Koperasi Mitra Sejahtera Samarinda. *Jurnal Ekonomia*. Vol 4, No. 3 2015

TAN, P. N.,STEINBACH, M., & KUMAR, V. 2006. *Introduction to Data Mining*. Boston: Pearson Education

UBAIDILLAH. Peranan Volume Usaha dalam Mencapai Keberhasilan Koperasi.*Jurnal Khasanah Ilmu.* Vol 7 No.1 2016

JURASIK (Jurnal Riset Sistem Informasi & Teknik Informatika) ISSN 2527-5771
Volume 1, Nomor 1, Juli 2016

KOMPARASI KINERJA ALGORITMA *FUZZY C-MEANS* DAN *K-MEANS* DALAM PENGELOMPOKAN DATA SISWA BERDASARKAN PRESTASI NILAIAKADEMIK SISWA

(Studi Kasus : SMP Negeri 2 Pematangsiantar)

Nelson Butarbutar[1], Agus Perdana Windarto[2], Dedi Hartama[3], Solikhun[4]

Mahasiswa STIKOM Tunas Bangsa Pematangsiantar[1]
Dosen STIKOM Tunas Bangsa Pematangsiantar[2,3,4]
E-mail : nbutarbutar@gmail.com[1], zhantura.gusti@gmail.com[2],
solikhun@amiktunasbangsa.ac.id[3], dedyhartama@amiktunasbangsa.ac.id[4]

Abstrak

Berbagai upaya telah dilakukan oleh pihak sekolah untuk meningkatkan prestasi akademik siswa sebagai upaya untuk mencapai standar pendidikan nasional. Salah satunya adalah dengan melakukan bimbingan belajar pada masing-masing siswa, namun hasilnya belum begitu memuaskan. Hal ini disebabkan karena pihak sekolah khususnya bagian pendidikan tidak memahami sepenuhnya kemampuan masing-masing siswa dalam menguasai suatu mata pelajaran khususnya mata pelajaran inti UN. Untuk mengatasi hal ini dengan memanfaatkan teknik clustering akan dilakukan pengelompokan data siswa berdasarkan prestasi nilai akademik yang sumber datanya diperoleh langsung dari bagian pendidikan. Dengan menggunakan teknik clustering, bagian pendidikan akan lebih mudah mendata siswanya berdasarkan kemampuannya masing-masing sesuai dengan prestasi nilai akademiknya. Teknik clustering yang digunakan dalam penelitian ini adalah algoritma Fuzzy C-Means dan K-Means. Berdasarkan hasil penelitian yang dilakukan dapat disimpulkan bahwa teknik clustering yang paling tepat digunakan dalam melakukan pengelompokan data siswa adalah algoritma K-Means dengan jumlah interasi sebanyak 11 untuk mendapatkan cluster data siswa sedangkan algoritma Fuzzy C-Means membutuhkan proses interasi yang panjang sebanyak 35 interasi dan proses perhitungan yang rumit serta hasil cluster data siswa kurang akurat dibandingkan menggunakan algoritma K-Means.

Kata kunci : *Standar pendidikan nasional; UN, clustering; fuzzy C-Means; K- Means*

Pendahuluan

Kurikulum adalah seperangkat rencana dan pengaturan mengenai tujuan, isi dan bahan pelajaran serta cara yang digunakan sebagai pedoman penyelenggaraan kegiatan pembelajaran untuk mencapai tujuan pendidikan tertentu. Tujuan pendidikan itu adalah mencapai standar pendidikan nasional yang meliputi standar isi, proses, kompetensi lulusan, tenaga kependidikan, sarana dan prasarana, pengelolaan, pembiayaan dan penilaian pendidikan. Dalam proses pendidikan di sekolah pihak akademik harus memahami kemampuan masing-masing siswa dalam menguasai suatu ilmu yang diajarkan disekolah karena hal ini sangat menentukan prestasi masing-masing siswa. Dalam sebuah institusi pendidikan kwalitas akademik masing-masing sekolah diukur berdasarkan prestasi nilai akademik masing-masing siswa baik itu nilai akademik setiap semester maupun nilai Ujian Nasional.

Berbagai penelitian telah dilakukan dengan menggunakan teknik *clustering* dalam mengelompokkan data siswa berdasarkan data nilai.Miftahus S., *at al*,.(2012) menggunakan teknik klasifikasi dalam mengelompokkan data nilai siswa untuk memprediksi tingkat kelulusan siswa dalam menghadapi UN.Suprihatin, (2011)

menggunakan teknik *clustering* algoritma *K-Means* dalam menentukan nilai ujian.Anita B. H., at al., (2013) dalam penelitiannya menggunakan algoritma *Fuzzy C-Means* dan TOPSIS dalam mengelompokkan data siswa berdasarkan nilai prestasi akademik mata pelajaran peminatan, minat terhadap jurusan , nilai IQ serta kuota kelas yang tersedia.

Pada saat ini masing-masing sekolah khususnya pada tingkat Sekolah Menengah Pertama berusaha mempersiapkan masing-masing siswanya dalam menghadapi Ujian Nasional yang akan dilaksankan pada saat akhir studi. Berbagai upaya telah dilakukan seperti mengadakan bimbingan belajar, namun hasil yang diharapkan belum begitu memuaskan.Hal ini tentu disebabkan karena pihak sekolah khususnya bagian pendidikan tidak memahami kemampuan masing-masing siswa dalam menguasai suatu mata pelajaran khususnya mata pelajaran inti UN.Untuk itu melalui penelitian ini diusulkan sebuah teknik *clustering* untuk mengelompokkan data siswa berdasarkan kemampuannya dalam menguasai materi mata pelajaran berdasarkan nilai akademiknya.Dengan menggunakan teknik *clustering* bagian pendidikan akan lebih mudah mendata siswanya sesuai dengan kemampuannya masing-masing, sehingga pada saat kenaikan kelas bagian pendidikan dapat menempatkan siswa dalam satu kelas sesuai dengan kemampuan masing-masing siswa agar pihak guru ataupun wali kelas dapat menerapkan metode pengajaran yang tepat sehingga setiap siswa dapat meningkatkan prestasi akademiknya. Dalam penelitian ini teknik *clustering* yang digunakan adalah algoritma *Fuzzy C-Means* dan algoritma *K-Means*. Setelah dilakukan pengelompokan menggunakan kedua algoritma tersebut, selanjutnya akan dilakukan analisis komparasi kinerja kedua algoritma tersebut dalam melakukan pengelompokan data siswa.

Penelitian ini dilakukan untuk memudahkan pihak sekolah khususnya bagian pendidikan dalam mendata siswa sesuai dengan kemampuannya masing-masing melalui prestasi nilai akademik siswa dengan mengusulkan sebuah teknik *clustering* dalam mengelompokkan data siswa.

1. Mengimplementasikan teknik *clustering* metode *Fuzzy C-Means* dan *K-Means* dalam melakukan *clustering* data siswa.

Kajian Teori

Data Mining

Manusia dalam suatu organisasi, sadar atau tidak sadar telah memproduksi berbagai data yang jumlahnya sangat besar.Pada dasarnya data adalah entitas yang tidak memiliki arti, meskipun kemungkinan memiliki nilai di dalamnya. *Data mining* merupakan disiplin ilmu yang mempelajari metode untuk mengestrak pengetahuan atau menemukan pola dari suatu data, sehingga *data mining* sering juga disebut *Knowledge Discovery in Database* (KDD).

Konsep utama dari sebuah transformasi dalam *data mining* adalah suatu kumpulan data yang bersumber dari *database* yang berukuran besar yang diekstrak dan dirangkum untuk menemukan suatu pola berupa informasi yang bermanfaat dan mengandung sebuah pengetahuan yang dapat digunakan untuk membantu pengambil keputusan dalam mengambil kebijakan yang tepat untuk kepentingan sebuah organisasi.

Data mining dapat juga didefenisikan sebagai gabungan dari suatu pengalaman, nilai, informasi kontekstual dan juga pandangan pakar yang memberikan suatu *framework* untuk mengevaluasi dan menciptakan pengalaman baru dan informasi (Thomas H. Davenport, Laurence Prusak).

Clustering

Clustering adalah pengelompokan data, hasil observasi dan kasus ke dalam *class* yang mirip. Suatu *cluster* adalah koleksi data yang mirip antara satu dengan yang lain, dan memiliki perbedaan bila

dibandingkan dengan data dari *cluster* lain. Perbedaan utama algoritma *clustering* dengan klasifikasi adalah *clustering* tidak memiliki target/*class*/label, jadi *unsupervised learning*. *Clustering* sering digunakan sebagai tahap awal dalam proses *data mining*, dengan hasil *cluster* yang terbentuk akan menjadi input dari algoritma berikutnya yang digunakan. Ada beberapa pendekatan yang digunakan dalam mengembangkan metode *clustering*.Dua pendekatan utama tersebut adalah *clustering* dengan pendekatan partisi dan *clustering* dengan pendekatan hirarki.Disamping kedua pendekatan tersebut, ada juga *clustering* dengan pendekatan *automatic mapping* (*Self-Organising Map*/SOM).

Algoritma K-Means

Algoritma *K-Means* merupakan salah satu algoritma *clustering* yang digunakan untuk mempartisi data kedalam beberapa *cluster*, dimana data yang memiliki tingkat kemiripan yang tinggi dikelompokkan dalam satu *cluster* sedangkan data yang memiliki karakteristik yang berbeda dikelompokkan ke dalam *cluster* yang berbeda.

Metode Penelitian

Dalam metode *clustering* konsep utama yang ditekankan adalah pencarian pusat *cluster* secara iteratif, dimana pusat *cluster* ditentukan berdasarkan jarak minimum setiap data pada pusat *cluster*.

Langkah- langkah penyelesaian algoritma *K-Means* dapat diuraikan sebagai berikut :

1. Identifikasi data yang akan di*cluster* dan juga tentukan jumlah *cluster* data, X_{ij} (i =1,2,....n;j = 1,2,....m) dimana n adalah jumlah data yang akan di *cluster* dan m adalah jumlah variabel data.
2. Pada awal interasi pusat setiap *cluster* ditentukan secara acak (sembarang), C_{kj} (k = 1,.....k;j = 1,.....m) .
3. Tentukan jarak setiap data terhadap pusat *cluster* dengan menggunakan *formula Euclidean* berikut :

$$d_{ik} = \sqrt{\sum_{j=1}^{m}(X_{ij} - C_{kj})^2} \quad \text{---------} \ (2.1)$$

4. Kelompokkan data berdasarkan jarak minimum data terhadap pusat *cluster*.
5. Cek kondisi apakah ada data yang masih pindah pada kelompok lain, jika ia lakukan interasi berikutnya dengan menghitung pusat *cluster* berdasarkan nilai rata-rata dari data yang menjadi anggota pada *cluster* yang terbentuk hasil interasi sebelumnya dengan menggunakan rumus :

$$C_{kj} = \frac{\sum_{i=1}^{p} X_{ij}}{p}; \quad \text{----------} \ (2.2)$$

Dimana $X_{ij} \in$cluster ke-k dan adalah jumlah anggota *cluster*

6. Jika tidak ada lagi data yang pindah pada *cluster* lain, maka proses interasi berhenti.

Dalam algoritma *K-Means* suatu data di kelompokkan pada suatu *cluster* jika data tersebut memiliki jarak minimum terhadap pusat *cluster* yang dapat dihitung dengan menggunakan rumus (2.1). Untuk memudahkan pemahaman kita tentang impelementasi penggunaan algoritma *K-Means*, contoh hasil perhitungan algoritma *K-Means* dapat kita lihat pada tabel 1 berikut :

Tabel 1. Contoh Hasil Perhitungan Menggunakan Algoritma *K-Means*

Derajat Keanggotaan			Jarak	Cluster		
d_i	d_i	d_i	Minimum	C	C	C
2.3764	3.6245	1.2547	1.2547	0	0	1
5.2635	1.2356	2.3746	1.2356	0	1	0
3.5476	2.3669	3.4648	2.3669	0	1	0
4.0887	2.0187	3.2863	2.0187	0	1	0
1.2384	3.4658	4.2974	1.2384	1	0	0
4.2929	2.3648	1.2394	1.2394	0	0	1
2.9383	4.2722	3.0883	2.9383	1	0	0
1.0927	2.3937	1.2533	1.0927	1	0	0

Algoritma Fuzzy C-Means (FCM)

Fuzzy C-Means adalah salah satu metode *fuzzy clustering* yang pertama sekali dikembangkan oleh Dunn (1973) dan kemudian diperbaiki oleh Bezdek (1981) sebagai metode yang sering digunakan dalam pengenalan pola (*pattern recognition*).Dalam FCM pengelompokan data dilakukan berdasarkan derajat keanggotaan yang bernilai antara 0 dan 1. Dalam algoritma FCM langkah pertama yang dilakukan adalah menentukan pusat *cluster* yang akan menandai lokasi rata-rata untuk tiap *cluster*. Pada kondisi awal, pusat *cluster* ini masih belum akurat.Tiap tiap data memiliki derajat keanggotaan untuk tiap-tiap *cluster*. Dengan cara memperbaiki pusat *cluster* dan nilai keanggotaan tiap-tiap data secara berulang, maka dapat dilihat bahwa pusat *cluster* akan menuju lokasi yang tepat. Perulangan ini didasarkan pada minimasi fungsi objektif yang menggambarkan jarak dari titik data yang diberikan ke pusat *cluster* yang terbobot oleh derajat keanggotaan titik data tersebut. Output dari FCM bukan merupakan *fuzzy inference system*, namun merupakan deretan pusat *cluster* dan beberapa derajat keanggotaan untuk tiap-tiap titik data.

Langkah-langkah penyelesaian algoritma *Fuzzy C-Means* dapat diuraikan sebagai berikut :

1. Identifikasi data yang akan di *cluster* berupa matriks berukuran n x m (X_{ij} = data sampel ke-i (i = 1,2,...n) dan atribut ke-j (j = 1,2,...m).
2. Menentukan jumlah *cluster* (c), pangkat (w), maksimum interasi (maxiter), *error* terkecil yang diharapkan (Ɛ), fungsi objektif awal (P_0 = 0) dan interasi awal (t = 1).
3. Membangkitkan bilangan *random* μ_{ik} (i = 1,2,..,n;k = 1,2,..,c) sebagai elemen-elemen matriks partisi awal U.
   Menghitung jumlah setiap kolom dengan rumus :
   $$Q_j = \sum_{k=1}^{c} ?_{ik} \quad \text{----------} (2.3)$$
   Mengitung nilai elemen matriks partisi anggota himpunan U dengan rumus :
   $$\mu_{ik} = \frac{?_{ik}}{Q_j} \quad \text{----------} (2.4)$$
4. Menghitung nilai pusat *cluster* ke-k (V_{kj} dengan k = 1,2,..,c;j = 1,2,...,m)
5. Men [illegible] (2.5)
   ke-t [illegible] .
   $$P_t = \sum_{i=1}^{n} \sum_{k=1}^{c} ([\sum_{j=1}^{m} (X_{ij} - V_{kj})^2](\mu_{ik})^w) \quad (2.6)$$
6. Menghitung perubahan matriks partisi U:
   $$\mu_{ik} = \frac{[\sum_{j=1}^{m}(X_{ij} - V_{kj})^2]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}[\sum_{j=1}^{m}(X_{ij} - V_{kj})^2]^{\frac{-1}{w-1}}} \quad (2.7)$$
7. Cek kondisi berhenti:
   a. Jika ($P_t - P_{t-1}$) < Ɛ atau (t < MaxIter) maka interasi berhenti.
   b. Jika tidak : t = t + 1, ulangi langkah 4 sampai 7.

Dalam algoritma *Fuzzy C-Means* suatu data dikelompokkan dalam suatu *cluster*, jika data memiliki nilai jarak maksimum partisi matriks U (μ_{ik}) terhadap pusat *cluster*. Untuk memudahkan pemahaman implementasi algoritma *Fuzzy C-Means*, berikut contoh tabel 2 hasil perhitungan algoritma *Fuzzy C-Means* :

Tabel 2. Contoh hasil perhitungan menggunakan algoritma *Fuzzy C-Means*

Derajat Keanggotaan			Jarak Maksimum	Klaster		
μ_i	μ_i	μ_i		C	C	C
0.456	0.284	0.260	0.456	1	0	0
0.282	0.123	0.595	0.595	0	0	1
0.564	0.234	0.202	0.564	1	0	0
0.234	0.554	0.212	0.554	0	1	0
0.345	0.432	0.223	0.432	0	1	0
0.371	0.123	0.506	0.506	0	0	1
0.653	0.231	0.116	0.653	1	0	0
0.354	0.213	0.433	0.433	0	0	1

Standar Penilaian Pendidikan

Undang-Undang Nomor 20 Tahun 2003 tentang Sistem Pendidikan Nasional. Pasal 1 angka 1 menyatakan bahwa "pendidikan adalah usaha sadar dan terencana untuk mewujudkan suasana belajar dan proses pembelajaran agar peserta didik secara aktif mengembangkan potensi dirinya untuk memiliki kekuatan spiritual keagamaan, pengendalian diri, kepribadian, kecerdasan, akhlak mulia, serta keterampilan yang diperlukan dirinya, masyarakat, bangsa dan negara". Standar Nasional Pendidikan berfungsi sebagai dasar perencanaan, pelaksanaan, dan pengawasan pendidikan dalam rangka mewujudkan pendidikan nasional yang bermutu. Salah satu bagian dari Standar Pendidikan Nasional adalah Standar Penilaian yang bertujuan untuk menjamin :

1. Perencanaan penilaian peserta didik sesuai dengan kompetensi yang akan dicapai dan berdasarkan prinsip-prinsip penilaian.
2. Pelaksanaan penilaian peserta didik secara profesional, terbuka, edukatif, efektif, efisien dan sesuai dengan konteks sosial budaya.
3. Pelaporan hasil penilaian peserta didik secara objektif, akuntabel dan informatif.

Standar Penilaian Pendidikan adalah kriteria mengenai mekanisme, prosedur, dan instrumen penilaian hasil belajar peserta didik.

Hasil dan Pembahasan

Penelitian ini dilakukan untuk mengetahui tingkat kemampuan masing-masing siswa dalam memahami dan menguasai ilmu yang di pelajari, dengan melakukan *clustering* terhadap data siswa yang sudah mengikuti ujian akhir semester genap berdasarkan nilai mata pelajaran inti UN (Bahasa Inggris, Bahasa Indonesia, Matematika, dan IPA) dan nilai rata-rata rapor. Hasil *clustering* data siswa nantinya akan dijadikan sebagai dasar untuk pembagian kelas oleh bagian pendidikan pada saat kenaikan kelas. Melalui hasil penelitian ini, diharapkan pihak pendidikan dapat menentukan kebijakan yang tepat dalam menerapkan metode pengajaran yang akan dilakukan oleh masing-masing guru sesuai dengan kemampuan masing-masing siswa. Selain itu juga, melalui hasil penelitian ini bagian pendidikan dapat memantau perkembangan kemampuan masing-masing siswa dalam menguasai mata pelajaran khususnya mata pelajaran inti UN. Sehingga nantinya setiap siswa lebih siap dalam menghadapi Ujian Nasional, yang akan diadakan pada akhir masa studi serta tingkat kelulusan siswa dapat lebih di maksimalkan. Dalam penelitin ini *clustering* data siswa dilakukan dengan menggunakan algoritma *Fuzzy C-Means* dan algoritma *K-Means*.Selain untuk menentukan *clustering* data siswa, penelitian ini juga dilakukan untuk melakukan analisis komparasi kinerja metode *Fuzzy C-Means* dan metode *K-Means* dalam melakukan pengelompokan data siswa.

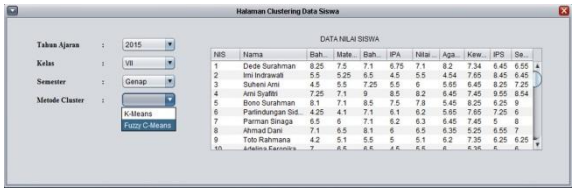
Data penelitian diperoleh langsung dari bagian pendidikan dengan jumlah data sebanyak 150 *record* data siswa. Data yang diperoleh, disimpan pada *database Mysql* agar dapat diproses menggunakan aplikasi *prototypeclustering*. Berikut adalah data nilai siswa yang berhasil diperoleh :

Tabel 3. Data nilai siswa

Nomor	Nama Siswa	Data Nilai Akademik				
		Bahasa Inggris (X_1)	Matematika (X_2)	Bahasa Indonesia (X_3)	IPA (X_4)	Rata-Rata Nilai Rapor (X_5)
1	Dede Surahman	8.25	7.50	7.10	6.75	7.10
2	Irni Indrawati	5.50	5.25	6.50	4.50	5.50
3	Suheni Arni	4.50	5.50	7.25	5.50	6.00
4	Arni Syafitri	7.25	7.10	9.00	8.50	8.20
5	Bono Surahman	8.10	7.10	8.50	7.50	7.80
6	Parlindungan Sidabutar	4.25	4.10	7.10	6.10	6.20
7	Parman Sinaga	6.50	6.00	7.10	6.20	6.30
8	Ahmad Dani	7.10	6.50	8.10	6.00	6.50
9	Toto Rahmana	4.20	5.10	5.50	5.00	5.10
10	Adelina Feronika Situorus	7.00	6.50	6.50	4.50	5.50
11	Putu Donang	4.50	5.20	6.00	5.00	5.50
12	Ali Syahbana	9.00	8.10	8.00	7.50	7.80
13	Romi Dana	8.20	7.50	7.80	6.40	7.60
14	Sulistiwowati	4.30	5.30	6.50	6.50	6.30
15	Butet Manurung	6.20	5.70	7.30	6.30	6.50
16	Rano Karno Sinaga	5.50	6.20	7.30	6.60	7.00
17	Susilowardono	7.20	6.70	7.30	6.20	6.80
18	Roro Mendut	4.20	5.50	7.20	6.50	7.20
19	Endang Surati	5.30	6.40	7.60	6.90	7.00
20	Jhonson Sidabutar	4.25	5.80	6.20	6.20	6.50
21	Parlin Siaituri	7.80	6.40	8.30	7.50	7.70
22	Toto Andara Saragih	8.20	9.10	7.40	6.50	7.90
23	Pangeran g Hutabarat	5.20	4.60	5.20	6.30	5.30
24	Surti Nandani	4.20	4.70	5.40	6.30	6.00
25	Ontong Andi	6.30	6.30	7.30	6.20	6.60
....	....	....	....	....	....	....
....	....	....	....	....	....	....
149	Andi Sinaga	6.20	7.10	6.70	6.50	6.50
150	Permana Sirait	7.20	6.50	7.10	6.85	6.90

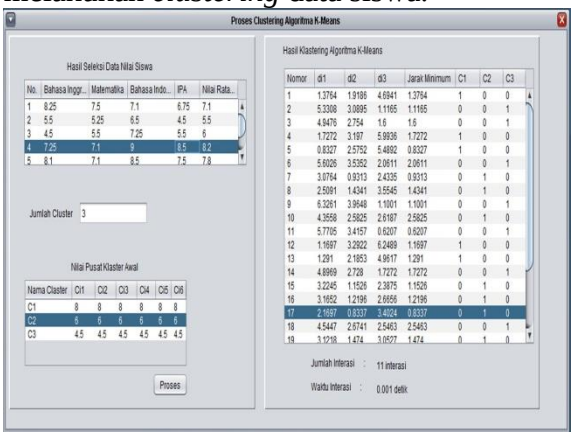
Tahap Pengolahan Data

Untuk mendapatkan *output* dari hasil penelitian yang dilakukan, pengolahan data yang dilakukan adalah menggunakan dua metode teknik clustering. Metode tersebut adalah algoritma *Fuzzy C-Means* dan *K-Means*. Untuk mendapatkan hasil perhitungan *clustering* digunakan aplikasi *prototype clustering* yang dibangun dengan menggunakan bahasa pemrograman java. Berikut adalah tampilan aplikasi clustering data siswa :



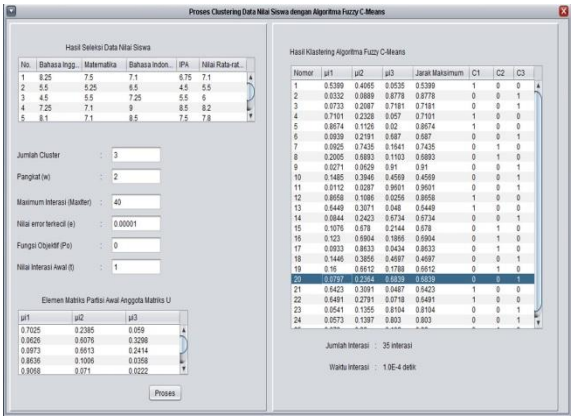
Gambar 1. Halaman c*luster* data siswa

Halaman *Cluster* Data Siswa merupakan halaman yang digunakan untuk memilih metode yang digunakan untuk melakukan *clustering* data siswa.



Gambar 2. Halaman proses algoritma *K-Means*

Halaman Proses Algoritma *K-Means* merupakan halaman yang digunakan untuk menampilkan hasil proses perhitungan algoritma *K-Means*. Dari hasil perhitungan dapat kita lihat bahwa proses interasi berhenti pada interasi ke-11 dengan waktu interasi 0.001 detik.



Gambar 3. Halaman proses algoritma *Fuzzy C-Means*

Halaman Proses Algoritma *Fuzzy C-Means* merupakan halaman yang digunakan untuk menampilkan hasil perhitungan algoritma *Fuzzy C-Means*. Dari hasil perhitungan dapat kita lihat bahwa proses interasi berhenti pada interasi ke-35 dengan waktu interasi 1.0E-4 detik.

Algoritma K-Means

Proses pengelompokan data menggunakan algoritma *K-Means* dilakukan pertama sekali dengan

melakukan proses identifikasi terhadap data yang akan dicluster. Dalam hal ini data yang akan dicluster adalah data siswa yang ada pada tabel 3 di atas. Setelah data yang akan dicluster diketahui maka langkah selanjutnya adalah menentukan pusat cluster awal yang ditentukan secara acak. Langkah selanjutnya adalah melakukan perhitungan algoritma *K-Means* dengan menggunakan aplikasi *prototypeclustering*. Hasil *clustering* ditentukan berdasarkan jarak minimum setiap data pada pusat *cluster* atau proses interasi *clustering* berhenti jika tidak ada lagi data yang berpindah pada *cluster* lain. Berdasarkan hasil percobaan yang dilakukan, proses interasi algoritma *K-Means* berhenti pada interasi ke-11. Hasil *clustering* menggunakan algoritma *K-Means* dapat kita lihat pada tabel 4 berikut :

Tabel 4. Hasil *clustering* data siswa menggunakan algoritma *K-Means*

Nomor	Derajat Keanggotaan			Jarak Minimum	Cluster		
	d_{i1}	d_{i2}	d_{i3}		C_1	C_2	C_3
1	1.3764	1.9186	4.6941	1.3764	1	0	0
2	5.3308	3.0895	1.1165	1.1165	0	0	1
3	4.9476	2.7540	1.6000	1.6000	0	0	1
4	1.7272	3.1970	5.9936	1.7272	1	0	0
5	0.8327	2.5752	5.4892	0.8327	1	0	0
6	5.6026	3.5352	2.0611	2.0611	0	0	1
7	3.0764	0.9313	2.4335	0.9313	0	1	0
8	2.5091	1.4341	3.5545	1.4341	0	1	0
9	6.3261	3.9648	1.1001	1.1001	0	0	1
10	4.3558	2.5825	2.6187	2.5825	0	1	0
11	5.7705	3.4157	0.6207	0.6207	0	0	1
12	1.1697	3.2922	6.2489	1.1697	1	0	0
13	1.2910	2.1853	4.9617	1.2910	1	0	0
14	4.8969	2.7280	1.7272	1.7272	0	0	1
15	3.2245	1.1526	2.3875	1.1526	0	1	0
16	3.1652	1.2196	2.6656	1.2196	0	1	0
17	2.1697	0.8337	3.4024	0.8337	0	1	0
18	4.5447	2.6741	2.5463	2.5463	0	0	1
19	3.1218	1.4740	3.0527	1.4740	0	1	0
20	4.8443	2.6397	1.6567	1.6567	0	0	1
.....	.....	.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....	.....	.....
149	2.7579	0.7272	2.8906	0.7272	0	1	0
150	1.9263	0.6998	3.5197	0.6998	0	1	0

Algoritma Fuzzy C-Means

Proses pengelompokan data menggunakan algoritma *Fuzzy C-Means* dilakukan dengan terlebih dahulu melakukan proses identifikasi terhadap data yang akan di*cluster*. Dalam hal ini data yang akan di*cluster* adalah data siswa seperti yang terdapat pada tabel 3. Setelah data diidentifikasi, langkah selanjutnya adalah melakukan inisialisasi jumlah *cluster*, tingkat kekaburan data (W), maximum interasi, akurasi (e), nilai fungsi objektif awal, dan interasi awal.

Dalam proses *clustering* menggunakan algoritma *Fuzzy C-Means* langkah pertama yang dilakukan adalah menentukan nilai berikut :

1. Jumlah Cluster (c) = 3
2. Pangkat (W) = 2
3. Maksimum Interasi = 40
4. *Error* Terkecil (e) = 10^{-5}

5. Fungsi Objektif Awal(P_0) = 0
6. Interas Awal = 1

Dalam algoritma *Fuzzy C-Means* proses interasi berhenti jika nilai hasil perhitungan fungsi objektif bernilai positif dan ($|P_t - P_{t-1}|$) < e (*error* terkecil). Setelah dilakukan proses interasi menggunakan *Fuzzy C-Means*, proses interasi berhenti pada interasi ke-35 dengan hasil *clustering* data siswa ditentukan berdasarkan jarak maksimum data pada derajat keanggota *cluster* data. Hasil *clustering* data siswa menggunakan *Fuzzy C-Means* dapat kita lihat pada tabel 5 berikut :

Tabel 5. Hasil c*lustering* data siswa menggunakan algoritma *Fuzzy C-Means*

Nomor	Derajat Keanggotaan			Jarak Maksimum	Cluster		
	μ_{i1}	μ_{i2}	μ_{i3}		C_1	C_2	C_3
1	0.5399	0.4065	0.0535	0.5399	1	0	0
2	0.0332	0.0889	0.8778	0.8778	0	0	1
3	0.0733	0.2087	0.7181	0.7181	0	0	1
4	0.7101	0.2328	0.0570	0.7101	1	0	0
5	0.8674	0.1126	0.0200	0.8674	1	0	0
6	0.0939	0.2191	0.6870	0.6870	0	0	1
7	0.0925	0.7435	0.1641	0.7435	0	1	0
8	0.2005	0.6893	0.1103	0.6893	0	1	0
9	0.0271	0.0629	0.9100	0.9100	0	0	1
10	0.1485	0.3946	0.4569	0.4569	0	0	1
11	0.0112	0.0287	0.9601	0.9601	0	0	1
12	0.8658	0.1086	0.0256	0.8658	1	0	0
13	0.6449	0.3071	0.0480	0.6449	1	0	0
14	0.0844	0.2423	0.6734	0.6734	0	0	1
15	0.1076	0.6780	0.2144	0.6780	0	1	0
16	0.1230	0.6904	0.1866	0.6904	0	1	0
17	0.0933	0.8633	0.0434	0.8633	0	1	0
18	0.1446	0.3856	0.4697	0.4697	0	0	1
19	0.1600	0.6612	0.1788	0.6612	0	1	0
20	0.0797	0.2364	0.6839	0.6839	0	0	1
.....	.....	.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....	.....	.....
149	0.0833	0.8324	0.0843	0.8324	0	1	0
150	0.0550	0.9257	0.0193	0.9257	0	1	0

Hasil dan Pembahasan

Berdasarkan perhitungan yang dilakukan dengan menggunakan *prototype* aplikasi *clustering* menggunakan algoritma *K-Meanscluster* data siswa yang terbentuk sebanyak 3 *cluster* dengan anggotanya terdiri dari :

a) *Cluster* pertama (C_1) dengan anggota terdiri dari 1,4,5,12,13,21,22,30,31,36, 37,38,43,44,48,54,55,56,61,62,63,64,65 ,71,72,78,80,81,85,86,87,88,90,97,102, 108,109,116,121,122,123,126,130,131, 135,136,137,142,143,147.
b) *Cluster* kedua (C_2) dengan anggota terdiri dari 7,8,10,15,16,17,19,25,29, 32,33,34,41,45,46,47,52,53,57,58,66,70 ,74,75,76,79,89,93,94,95,96,100,101,10 3,105,106,107,113,114,115,117,120,12 5,129,132,133,134,140,141,144,146,14 9,150.
c) *Cluster* ketiga (C_3) dengan anggota terdiri dari 2,3,6,9,11,14,18,20,23,24, 26,27,28,35,39,40,42,49,50,51,59,60,67 ,68,69,73,77,82,83,84,91,92,98,99,104, 110,111,112,118,119,124,127,128,138, 139,145,148.

Berdasarkan perhitungan dengan menggunakan teknik *Fuzzy C-Means* maka data siswa dikelompokkan kedalam 3 *cluster* dengan anggotanya terdiri dari :

a) *Cluster* pertama (C_1) yang anggotanya terdiri dari 1,4,5,12,13,21,22,30,31,36 ,37,38,43,44,48,55,56,61,63,64,65,71,7 2,78,81,85,86,87,88,90,97,102,108,109, 116,122,126,130,131,135,136,142,143, 147.
b) *Cluster* kedua (C_2) dengan anggotanya terdiri dari 7,8,15,16,17,19,25,29,32, ,41,45,46,47,52,53,54,57,58,62,66,70,7 4,76,79,80,93,94,95,96,100,101,103,10 5,106,107,113,114,115,117,120,121,12 3,125,129,132,133,134,137,140,141,14 6,149,150.
c) *Cluster* ketiga (C_3) dengan anggotanya terdiri dari 2,3,6,9,10,11,14,18,20,23, 24,26,27,28,33,35,39,40,42,49,50,51,59 ,60,67,68,69,73,75,77,82,83,84,89,91,9 2,98,99,104,110,111,112,118,119,124,1 27,128,138,139,144,145,148.

Analisis Komparasi Kinerja Algoritma Fuzzy C-Means dan K-Means

Untuk menganalisis komparasi kinerja kedua algoritma tersebut dalam melakukan pengelompokan data siswa berdasarkan prestasi akademik, data nilai siswa akan dikelompokkan kedalam 3 *cluster* dengan ketentuan *cluster* pertama (C1) rata-rata nilai siswa adalahlebih besar atau sama dengan 7.24, dan *cluster* kedua dengan rata-rata nilai siswa lebih besar atau sama dengan 6 dan lebih kecil dari 7.24, sedangkan untuk *cluster* ketiga rata-rata nilai siswalebih kecil dari 6. Dari hasil interasi kedua algoritma tersebut dapat kita lihat bahwa hasil interasi algoritma *K-Means* lebih akurat dalam melakukan pengelompokan data siswa, seperti terlihat pada data nomor 10 dan nomor 33 dimana nilai rata-rata siswa adalah 6 dan 6.14 pada algoritma *K-Means* data siswa dikelompokkan pada *cluster* kedua sesuai dengan ketentuan yang diinginkan sedangkan hasil *cluster* algoritma *Fuzzy C-Means* data siswa dikelompokkan pada *cluster* ketiga yang tentunya tidak sesuai dengan syarat yang diinginkan sebelumnya. Jadi dapat disimpulkan bahwa algoritma *K-Means* lebih tepat digunakan untuk mengelompokkan data siswa. Perbedaan hasil *clustering* algoritma *K-Means* dan *Fuzzy C-Means* dapat kita lihat pada tabel 6 berikut :

Tabel 6. Hasil *clustering* kedua algoritma

Nomor	Nama	Rata-rata Nilai	Cluster Hasil Algoritma K-Means	Cluster Hasil Algoritma Fuzzy C-Means
1	Dede Surahman	7.34	C1	C1
2	Imi Indrawati	5.45	C3	C3
...	...	...	...	...
...	...	...	...	...
9	Toto Rahmana	4.98	C3	C3
10	Adelina Veronika Situmorang	6	C2	C3
...	...	...	...	...
...	...	...	...	...
32	Iwan Sahilah	6.24	C2	C2
33	Ilmu Prahmana	6.14	C2	C3
34	Toto Andi Surahman	6.78	C2	C2
...	...	...	...	...
...	...	...	...	...
53	Dara Siring	6.22	C2	C2
54	Fernando Hutabarat	7.26	C1	C2
55	Ragel Sutono	7.56	C1	C1
...	...	...	...	...
...	...	...	...	...
61	Wadin Pardamean Manik	7.36	C1	C1
62	Beta Siagian	7.34	C1	C2
63	Kaku Siagian	8.52	C1	C1
...	...	...	...	...
...	...	...	...	...
74	Abdi Silalahi	7.2	C2	C2
75	Abdi Pramono	6.1	C2	C3
76	Panca Prahmana	7.16	C2	C2
...	...	...	...	...
...	...	...	...	...
79	Toto Sidabutar	7.08	C2	C2
80	Torang Sinaga	7.28	C1	C2
81	Bedy Manik	8.3	C1	C1
...	...	...	...	...
...	...	...	...	...
88	Ferdinand Sidabutar	8.4	C1	C1
89	Marsanda Manik	6.06	C2	C3
90	Endang Manik	7.4	C1	C1
...	...	...	...	...
...	...	...	...	...
120	Didik Borna	6.66	C2	C2
121	Andika Ambar	7.26	C1	C2
122	Palentin Manurung	8.04	C1	C1
123	Polanda Sinaga	7.24	C1	C2
...	...	...	...	...
...	...	...	...	...
136	Titin Butabutar	7.66	C1	C1
137	Nasibil Saragih	7.28	C1	C2
138	Santi Sinaga	5.6	C3	C3
...	...	...	...	...
143	Cakrawati Sulis	8.54	C1	C1
144	Sisi Sulistini	6.08	C2	C3
145	Rara Indri	5.16	C3	C3
...	...	...	...	...
149	Andi Sinaga	6.6	C2	C2
150	Permana Sirait	6.91	C2	C2

Kesimpulan

Teknik *clustering* dapat digunakan untuk meningkatkan prestasi akademik siswa dengan cara melakukan *clustering* data siswa berdasarkan kemampuannya melalui prestasi nilai akademik, sehingga pihak akademik dapat memantau perkembangan prestasi siswa dan dapat menentukan kebijakan yang tepat dalam menentukan metode pengajaran yang tepat sesuai dengan kemampuan siswa. Berdasarkan hasil penelitian yang dilakukan teknik *clustering* yang paling tepat digunakan adalah metode *K-Means*.

Daftar Pustaka

Maria E. S., et al, 2015, *Sistem Penentuan Jurusan Sekolah Menengah Atas Negeri 1 Karangmojo*, Seminar Nasional Teknologi Informasi dan Multimedia 2015, STMIK AMIKOM Yogyakarta, ISSN : 2302 – 3805.

Aniq Noviciatie Ulfah dan Shofwatul Uyun, 2015, *Analisis Kinerja Algoritma Fuzzy C-Means dan K-Means pada Data Kemiskinan*, Jatisi Vol. 1 No. 2, ISSN : 2407 - 4322

Suprihatin, 2011, *Clustering K-Means Untuk Penentuan Nilai Ujian*, JUSI Vol. I No. 1, ISSN : 2087 – 8737.

Mardiani, 2014, *Perbandingan Algoritma K-Means dan EM Untuk Clusterisasi Nilai Mahasiswa Berdasarkan Asal Sekolah*, Citec Journal Vol. I No. 4, ISSN : 2354 - 5771

Erfan Agil Putranto, et al, 2012, *Sistem Pendukung Keputusan Penjurusan Siswa Kelas X SMA Negeri 2 dengan Metode Fuzzy C-Means dengan Penggunaan Daya Dukung Minat*, JURNAL ITSMART Vol. 1 No. 2, ISSN : 2301 – 7201

Cakra Ramadhana, et al, 2013, *Data Mining dengan Algoritma Fuzzy C-Means Clustering dalam Kasus Penjualan di PT. Sepatu Bata*, Seminar Nasional Teknologi Informasi dan Komunikasi Terapan, ISBN : 979-26-0266-6.

Tb. Ai Munandar, et al, 2013, *Clustering Data Nilai Mahasiswa Untuk Pengelompokan Konsentrasi Jurusan Menggunakan Fuzzy Cluster Means*, Seminar Nasional Aplikasi Teknologi Informasi (SNATI) Yogyakarta, ISSN : 1907 – 5022.

Idni Irsalina, et al, 2014, *Clustering Gender Berdasarkan Nilai Maksimum Minimum Amplitudo Suara Berbasis Fuzzy C-Means (FCM)*, Prosiding SNATIF Ke-1, ISBN : 978-602-1180-04-4.

Idni Irsalina, et al, 2014, *Clustering Gender Berdasarkan Nilai Maksimum Minimum Amplitudo Suara Berbasis Fuzzy C-Means (FCM)*, Prosiding SNATIF Ke-1, ISBN : 978-602-1180-04-4

Anita Budi Hastuti, et al, 2013, *Implementasi Metode Fuzzy C-Means dan TOPSIS Dalam Membangun Sistem Pendukung Keputusan Penentuan Jurusan SMA (Studi Kasus : Penentuan Jurusan di SMA Negeri 1 Wonosari)*, JURNAL DASI, Vol. 14 No. 2, ISSN : 1411 – 3201.

Miftahus Sholihin dan Aizatus Sholikhiyah, 2012, *Prediksi Tingkat Kelulusan Siswa Dalam UAN di SMP Negeri 2 Deket Dengan Menggunakan Metode Teorema Bayes*, Jurnal Teknika, Vol. 4 No. 2, ISSN : 2085 – 0859.

Buletin Ilmiah Mat. Stat. Dan Terapannya (Bimaster)
Volume 02, No.1(2013), hal. 21-26

CLUSTERING LULUSAN MAHASISWA MATEMATIKA FMIPA UNTAN PONTIANAK MENGGUNAKAN ALGORITMA *FUZZY C-MEANS*

Cary Lineker Simbolon, Nilamsari Kusumastuti, Beni Irawan

INTISARI

Clustering adalah proses pengelompokan data ke dalam cluster berdasarkan parameter tertentu sehingga obyek-obyek dalam sebuah cluster memiliki tingkat kemiripan yang tinggi satu sama lain dan sangat tidak mirip dengan obyek yang lain pada cluster yang berbeda. Fuzzy C-Means termasuk dalam salah satu teknik clustering. Seperti teknik clustering lainnya. Fuzzy C-Means juga mengelompokkan sejumlah obyek. Pada jurnal ini teknik Fuzzy C-Means digunakan untuk mengelompokkan lulusan jurusan Matematika FMIPA Universitas Tangjungpura (UNTAN). Lulusan dibagi kedalam empat cluster berdasarkan IPK dan lama studi. Hasil dari penelitian ini adalah diperoleh empat pusat cluster atau center. Untuk cluster 1 terdiri dari lulusan dengan kisaran IPK 3,15 dan lama studi 5,09 tahun , cluster 2 terdiri dari lulusan dengan kisaran IPK 2,88 dan lama studi 7,32 tahun, cluster 3 terdiri dari lulusan dengan kisaran IPK 3,48 dan lama studi 4,37 tahun serta cluster 4 terdiri dari lulusan dengan kisaran IPK 2,89 dan lama studi 5,91 tahun. Lulusan yang paling banyak anggotanya ada pada cluster 4. Hasil yang diperoleh menunjukkan bahwa masih banyak lulusan jurusan Matematika di Fakultas MIPA UNTAN yang menempuh lama studi lebih dari 5 tahun.

Kata Kunci : *Fuzzy, Fuzzy C-Means, clustering*

PENDAHULUAN

Proses *clustering* adalah proses pengelompokan data ke dalam *cluster* berdasarkan parameter tertentu sehingga obyek-obyek dalam sebuah *cluster* memiliki tingkat kemiripan yang tinggi satu sama lain dan sangat tidak mirip dengan obyek yang lain pada *cluster* yang berbeda [1]. Pembentukan *cluster* data merupakan salah satu teknik yang digunakan dalam mengetahui pola kecenderungan suatu data. Analisis *cluster* merupakan proses membagi data dalam suatu himpunan ke dalam beberapa kelompok yang kesamaan karakteristik datanya dalam suatu kelompok lebih besar daripada kesamaan karakteristik data tersebut dengan data dalam kelompok lain [2].

Fuzzy C-Means merupakan suatu teknik *clustering* data yang mana keberadaan tiap-tiap titik data dalam suatu *cluster* ditentukan oleh derajat keanggotaannya. Teknik ini pertama kali diperkenalkan oleh Jim Bezdek pada tahun 1981 [3]. Kelebihan dari *Fuzzy C-Means* adalah dapat melakukan *clustering* lebih dari satu variabel secara sekaligus. Penelitian dengan metode *Fuzzy C-Means* telah dilakukan Hadi untuk pembagian kelas kuliah mahasiswa. Hasilnya dibandingkan dengan hasil menggunakan algoritma *subtraktif*. Kesimpulan yang diperoleh bahwa algoritma *Fuzzy C-Means* lebih cocok untuk alokasi mahasiswa menjadi beberapa kelas karena masukan berupa jumlah *cluster* yang diinginkan [4]. Metode *Fuzzy C-Means* digunakan untuk menentukan nilai akhir kuliah [5]. Dengan menggunakan metode ini, nilai akhir ditentukan secara alami karena berdasarkan pada kecenderungan masing-masing data pada kelompoknya. *Fuzzy C-Means* digunakan untuk *clustering* data *performance* mengajar dosen [6]. Dari 4 pusat *cluster* yang di set di awal iterasi, ada 2 *cluster* yang mempunyai nilai sama, yaitu *cluster* 3 dan *cluster* 4. Dari hasil *clustering*, dihasilkan 3 *cluster performance* mengajar dosen. Aplikasi algoritma *Fuzzy C-Means* diterapkan untuk pengelompokan lulusan Universitas Islam Negeri (UIN) Malang semester ganjil tahun akademik *2005/2006* [7]. *Fuzzy C-Means* digunakan untuk meneliti tentang bahan pangan berdasarkan kandungan nutrisinya [8].

Dari hasil penelitian tersebut, diperoleh 4 *cluster* yaitu golongan hijau, golongan kuning, golongan jingga, dan golongan merah

Penelitian ini menggunakan metode *clustering* dengan *Fuzzy C-Means.* Permasalahan yang dibahas dalam penelitian ini adalah bagaimana cara kerja algoritma *Fuzzy C-Means* untuk menyelesaikan masalah *clustering* lulusan mahasiswa Matematika FMIPA UNTAN Pontianak. Tujuan dari penelitian adalah melakukan pengelompokan lulusan mahasiswa Matematika FMIPA UNTAN Pontianak yang dibatasi dari tanggal 24 Februari 2007 sampai 31 Maret 2012 dan mencari titik pusat *cluster* dengan *Fuzzy C-Means clustering* berdasarkan nilai IPK dan lama studi.

FUZZY C-MEANS CLUSTERING

Fuzzy C-Means clustering adalah suatu teknik *clustering* yang banyak digunakan dalam aplikasi *clustering*. *Fuzzy C-Means* menerapkan pengelompokan *fuzzy*, dimana setiap data dapat menjadi anggota dari beberapa *cluster* dengan derajat keanggotaan yang berbeda-beda pada setiap *cluster*. *Fuzzy C-Means* merupakan algoritma iteratif, yang menerapkan iterasi pada proses *clustering* data. Tujuan dari *Fuzzy C-Means* adalah untuk mendapatkan pusat *cluster* yang nantinya akan digunakan untuk mengetahui data yang masuk ke dalam sebuah *cluster*.

Algoritma *Fuzzy C-Means* adalah sebagai berikut [3]:

1. Menentukan data yang akan dicluster *X*, berupa matriks berukuran *n* x *m* (*n*=jumlah sampel data, *m*=atribut setiap data). X_{ij}=data sampel ke-*i* (*i*=1,2,...,*n*), atribut ke-*j* (*j*=1,2,...,*m*).
2. Tentukan jumlah *cluster* (c), pangkat (w), maksimum iterasi (MaxIter), *error* terkecil yang diharapkan (ζ), fungsi obyektif awal (P_0=0), iterasi awal (t=1).
3. Bangkitkan bilangan *random* μ_{ik}, *i*=1,2,...,*n*; *k*=1,2,...,*c*; sebagai elemen-elemen matriks partisi awal U. Matriks partisi (U) pada pengelompokan *fuzzy* memenuhi kondisi sebagai berikut [9]:

$$\mu_{ik}\ \varepsilon\,[0,1]; \qquad 1 \le i \le n; \qquad 1 \le k \le c$$

   μ_{ik} adalah derajat keanggotaan yang merujuk pada seberapa besar kemungkinan suatu data bisa menjadi anggota ke dalam suatu *cluster*.

   Hitung jumlah setiap kolom (atribut):

$$Q_i = \sum_{k=1}^{c} \mu_{ik}$$

$$Q_i = \mu_{i1} + \mu_{i2} + \ldots + \mu_{ic}$$

   dengan *i*=1,2,...,*n*

4. Hitung pusat *cluster* ke-k: V_{kj}, dengan *k*=1,2,...,*c*; dan *j*=1,2,...,*m*

$$V_{kj} = \frac{\sum_{i=1}^{n}\left(\left(\mu_{ik}\right)^{w} X_{ij}\right)}{\sum_{i=1}^{n}\left(\mu_{ik}\right)^{w}} \tag{1}$$

5. Hitung fungsi obyektif pada iterasi ke-t, P_t:
   Fungsi obyektif digunakan sebagai syarat perulangan untuk mendapatkan pusat *cluster* yan tepat. Sehingga diperoleh kecenderungan data untuk masuk ke *cluster* mana pada step akhir. Untuk iterasi awal nilai t=1.

$$P_t = \sum_{i=1}^{n} \sum_{k=1}^{c} \left(\left[\sum_{j=1}^{m} \left(X_{ij} - V_{kj} \right)^2 \right] \left(\mu_{ik} \right)^w \right) \tag{2}$$

6. Hitung perubahan matriks partisi:

$$\mu_{ik} = \frac{\left[\sum_{j=1}^{m} \left(X_{ij} - V_{kj} \right)^2 \right]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c} \left[\sum_{j=1}^{m} \left(X_{ij} - V_{kj} \right)^2 \right]^{\frac{-1}{w-1}}} \tag{3}$$

7. Cek kondisi berhenti:
   a. $|P_t - P_{t-1}| < \zeta$) atau (t>MaxIter) maka berhenti;
   b. Jika tidak, iterasi dinaikkan t=t+1, ulangi langkah ke-4.

HASIL DAN PEMBAHASAN

Data yang digunakan dalam penelitian ini adalah data lulusan mahasiswa Matematika dari tanggal 24 Februari 2007 sampai 31 Maret 2012. Data berupa nilai IPK dan lama studi sebanyak 93 lulusan. Dalam proses *clustering*, IPK dijadikan sebagai X_{i1} dan lama studi dijadikan X_{i2}. Pengolahan data pada penelitian ini menggunakan bantuan program yaitu Matlab. Dari *clustering* yang dilakukan diperoleh hasil yaitu nilai fungsi obyektif selama iterasi, pusat *cluster* atau *center* serta derajat keanggotaan lulusan untuk setiap *cluster* pada iterasi terakhir. Dalam penelitian ini, proses iterasinya berhenti pada iterasi ke-27 karena nilai $|P_t - P_{t-1}| < \xi$. Nilai fungsi obyektif pada iterasi terakhir yang diperoleh adalah 9,355243. Hasil dari *clustering* yang dilakukan dapat dilihat di bawah ini:

1. Nilai fungsi obyektif selama 27 iterasi dapat dilihat pada Tabel 1 di bawah ini:

Tabel 1. Nilai Fungsi Obyektif Selama 27 Iterasi

Iterasi ke-	Fungsi Obyektif
1	28,832597
2	21,933190
3	18,872350
4	14,825657
5	13,204085
6	12,404592
7	12,076182
8	11,634414
9	10,799451
10	9,994893
⋮	⋮
27	9,355243

2. Nilai pusat *cluster* atau *center*

   Pada iterasi terakhir (iterasi ke-27), pusat *cluster* V_{kj} yang dihasilkan dengan k=1,2,3,4 dan j=1,2 adalah:

$$V_{kj} = \begin{pmatrix} 3{,}15 & 5{,}09 \\ 2{,}88 & 7{,}32 \\ 3{,}48 & 4{,}37 \\ 2{,}89 & 5{,}91 \end{pmatrix}$$

Nilai ini merupakan nilai dari koordinat keempat titik pusat *cluster* dan memberikan garis besar tiap *cluster* yaitu:

a. Untuk pusat *cluster* 1 terdiri dari lulusan dengan kisaran lama studi 5,09 tahun dan kisaran IPK 3,15.
b. Untuk pusat *cluster* 2 terdiri dari lulusan dengan kisaran lama studi 7,32 tahun dan kisaran IPK 2,88.
c. Untuk pusat *cluster* 3 terdiri dari lulusan dengan kisaran lama studi 4,37 tahun dan kisaran IPK 3,48.
d. Untuk pusat *cluster* 4 terdiri dari lulusan dengan kisaran lama studi 5,91 tahun dan kisaran IPK 2,89.

3. Derajat keanggotaan lulusan untuk sertiap *cluster* pada iterasi terakhir (iterasi ke-27) dapat dilihat pada Tabel 2 di bawah ini:

Tabel 2. Derajat Keanggotaan Lulusan Untuk Setiap *Cluster* Pada Iterasi Terakhir

Lulusan ke	Derajat Keanggotaan (μ) Lulusan Untuk Setiap *Cluster* Pada Iterasi Terakhir			
	μ_{i1}	μ_{i2}	μ_{i3}	μ_{i4}
1	0,1203	0,0161	0,8209	0,0427
2	0,4042	0,0145	0,5209	0,0604
3	0,2821	0,0205	0,6276	0,0698
4	0,3930	0,0175	0,5211	0,0684
5	0,1137	0,0148	0,8317	0,0398
6	0,0709	0,0073	0,9005	0,0213
7	0,8566	0,0096	0,0742	0,0596
8	0,2463	0,0134	0,6905	0,0498
9	0,6898	0,0198	0,0622	0,2282
10	0,4090	0,0383	0,0808	0,4719
⋮	⋮	⋮	⋮	⋮
93	0,0810	0,7202	0,0538	0,1450

Dari derajat keanggotaan lulusan pada iterasi terakhir dapat diperoleh informasi mengenai kecenderungan lulusan untuk masuk ke *cluster* mana. Derajat keanggotaan terbesar menunjukkan bahwa kecenderungan tertinggi lulusan untuk masuk menjadi anggota *cluster* tersebut. Misalnya untuk lulusan ke-3, dapat menjadi:

(i). Anggota *cluster* pertama dengan derajat keanggotaan 0,2821
(ii). Anggota *cluster* kedua dengan derajat keanggotaan 0,0205

(iii). Anggota *cluster* ketiga dengan derajat keanggotaan 0,6276
(iv). Anggota *cluster* keempat dengan derajat keanggotaan 0,0698

Derajat keanggotaan terbesarnya terletak di *cluster* ketiga, maka lulusan ke-3 akan dimasukkan kedalam *cluster* ketiga. Hasil selengkapnya pengelompokkan ke-93 lulusan kedalam 4 *cluster* dapat dilihat pada Tabel 3 di bawah ini:

Tabel 3. Anggota Pada Keempat *Cluster*

Cluster	**Beranggotakan Lulusan Nomor**
1	7, 9, 11, 12, 18, 19, 21, 23, 24, 25, 28, 30, 37, 45, 48, 57, 58, 59, 60, 62, 66, 68, 69, 72, 73, 84.
2	47, 49, 54, 64, 75, 79, 85, 92, 93.
3	1, 2, 3, 4, 5, 6, 8, 29, 32, 44, 46, 55, 56, 61, 63, 71, 74, 78, 82, 83, 86, 87, 89, 90, 91.
4	10, 13, 14, 15, 16, 17, 20, 22, 26, 27, 31, 33, 34, 35, 36, 38, 39, 40, 41, 42, 43, 50, 51, 52, 53, 65, 67, 70, 76, 77, 80, 81, 88.

Dari Tabel 3 di atas dapat diperoleh:

a. *Cluster* 1 beranggotakan lulusan dengan IPK 2,67 sampai 3,65 dan lama studi 4,75 sampai 5,50 tahun sebanyak 26 lulusan.
b. *Cluster* 2 beranggotakan lulusan dengan IPK 2,60 sampai 3,28 dan lama studi 6,75 sampai 8,42 tahun sebanyak 9 lulusan.
c. *Cluster* 3 beranggotakan lulusan dengan IPK 3,07 sampai 3,81 dan lama studi 4,00 sampai 4,83 tahun sebanyak 25 lulusan.
d. *Cluster* 4 beranggotakan lulusan dengan IPK 2,38 sampai 3,51 dan lama studi 5,42 sampai 6,42 tahun sebanyak 33 lulusan.

PENUTUP

Berdasarkan *clustering* yang dilakukan diperoleh 4 *cluster*. Dari hasil keempat *cluster* tersebut diketahui bahwa pada *cluster* 4 memiliki anggota lulusan yang paling banyak dengan 33 lulusan. *Cluster* 4 terdiri dari lulusan dengan kisaran lama studi 5,91 tahun. Hal ini menunjukkan bahwa masih banyak mahasiswa jurusan Matematika di Fakultas MIPA Untan Pontianak yang menempuh lama studi lebih dari 10 semester atau 5 tahun. Hasil ini dapat dijadikan sebagai bahan pertimbangan jurusan dalam meningkatkan IPK mahasiswa untuk menyelesaikan masa studinya dengan cepat.

DAFTAR PUSTAKA

[1]. Asfi M. Pelabelan Otomatis Citra Menggunakan *Fuzzy C-Means* Untuk Sistem Temu Kembali Citra [M.Sc thesis]. Bogor: Institut Pertanian Bogor; 2008.
[2]. Bahar. Penentuan Jurusan Sekolah Menengah Atas Dengan Algoritma *Fuzzy C-Means* [MT thesis]. Semarang: Universitas Dian Nuswantoro; 2011.
[3]. Kusumadewi S, Purnomo H. *Aplikasi Logika Fuzzy Untuk Pendukung Keputusan*. Edisi ke-2. Yogyakarta: Graha Ilmu; 2010.
[4]. Hadi HY. Pembagian Kelas Kuliah Mahasiswa Menggunakan Algoritma Pengklasteran *Fuzzy*. Semarang: Universitas Diponegoro; 2005.

[5]. Khoiruddin AA. 2007, Menentukan Nilai Akhir Kuliah Dengan *Fuzzy* C-*Means*. *Seminar Nasional Sistem dan Informatika* (SNSI). 2007 Nov 16.
[6]. Luthfi ET. *Fuzzy* C-*Means* Untuk *Clustering* Data (Studi Kasus: Data *Performance* Mengajar Dosen). *Seminar Nasional Teknologi* (SNT). 2007 Nov 24.
[7]. Aziz A. Aplikasi Algoritma *Fuzzy C-Means Clustering* Untuk Pengelompokan Lulusan. Malang: Universitas Islam Negeri; 2010.
[8]. Kusumadewi S. Klasifikasi Kandungan Nutrisi Bahan Pangan Menggunakan *Fuzzy C-Means*. *Seminar Nasional Aplikasi Teknologi Informasi* (SNATI). 2007 Jun 16.
[9]. Kusumadewi S, Hartati S. *Integrasi Sistem Fuzzy dan Jaringan Syaraf*. Edisi ke-2. Yogyakarta: Graha Ilmu; 2010.

CARY LINEKER SIMBOLON : Fakultas MIPA Universitas Tanjungpura, Pontianak, lineker_cs@yahoo.com
NILAMSARI KUSUMASTUTI : Fakultas MIPA Universitas Tanjungpura, Pontianak, uminilam@yahoo.com
BENI IRAWAN : Fakultas MIPA Universitas Tanjungpura, Pontianak, benicsc@yahoo.com

2012 4th International Conference on Bioinformatics and Biomedical Technology
IPCBEE vol.29 (2012) © (2012) IACSIT Press, Singapore

Automatic Epithelial Cells Detection of Pap smears images using Fuzzy C-Means Clustering

Izzati Muhimmah [1 +], Rahadian Kurniawan [1] and Indrayanti [2]

[1] Department of Informatics, Universitas Islam Indonesia, Yogyakarta, Indonesia
[2] Faculty of Medicine, Universitas Muhammadiyah Yogyakarta, Indonesia

Abstract. The task of epithelial cells counting of pap smears images is still a burden for the pathologists. The computer-assisted system to address this challenge will benefit them. However, identification of epithelial cells is nontrivial task due to the complexity nature of pap smears data. This article proposes combination of distance-metric and Fuzzy C-Means clustering to identify the nuclei of epitheleal cells for any given pap smears images. The proposed method is evaluated using Liquid Based Preparation/ ThinPrep, which contain 350 epithelial cells which are confirmed by our expert Pathologist. The proposed methodology successfully automatically identify of cells with sensitivity rate of 90.86% and specificity rate of 81.62% in comparison to the expert assessment on NCI Bethesda Systems of pap smears data.

Keywords: pap smears, epithelial cells, nuclei detection, FCM

1. Introduction

The challenge of pap smears slide reading is known to be very time-consuming and tedious. It is due to the characteristics of pap smears images which suffer from low contrast, in homogeneities (cells tend not to absorbed the stained materials equally), and cluttered (uneven spreading of cells on slide) [1]. Often in practical setting, the task of counting the number of epithelial cells under microscope to measure its adequacy was not done because it is very difficult to do manually.

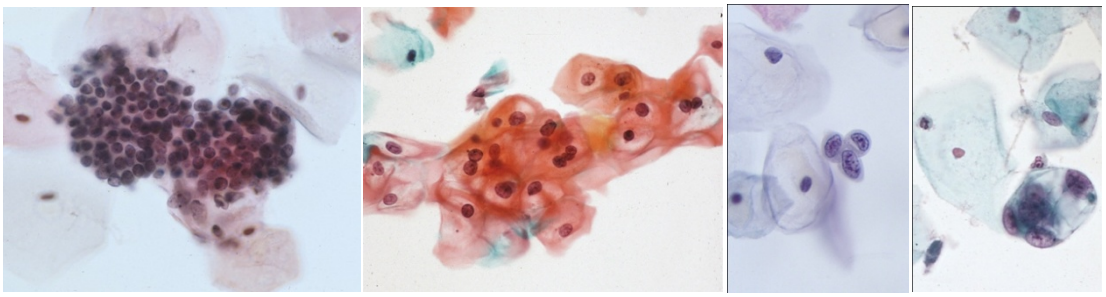


Fig. 1: Examples of pap smears images. Note that the images broadly varied w.r.t. magnifications and color intensities.

Pap smears images are varied with respect to magnification, contrast agent, cell types. These broad variations can be seen in Figure 1. As consequent, automatic detection using fixed parameters cannot be done. In other words, predefined thresholds on color intensities or imposing a single value of cells metric may not work.

There are some previous works to address this challenge by mean of automatic cells detection. Several approaches have been proposed such as: used deformable templates to identify cells with applied a generalized Hough transform to roughly detect round-like shapes [1], used a new criterion function based on

+ Corresponding author. Tel.: + 62274895287; fax: +62274895007.
E-mail address: izzati@uii.ac.id.

statistical structure of the objects in cell image [9], automatically finding the threshold value and detect the edges of cytoplasm and nucleus edges using moving k-means and Modified Seed Based Region Growing [10]. Although these methods present promising results, they were still using grayscale image and not yet tested on RGB images. Thus, this paper investigates automatic method for detection of epithelial cell location in RGB Pap Smear images.

The reminder of this paper is outlined as follows: the proposed methods for epithelial cells detection are described in Section 2. The data and the evaluation methods are explained in Section 3. Section 4 shows our results and discussions on our findings including pointing out the future works

2. Epithelial cells detection

2.1. Preprocessing

As mentioned earlier, the Pap smears images may suffer from low contrast. So, the preprocessing step is needed in order to tackle this. In this step we adopted method proposed by Plissiti, et al [2] which can be summarized as follows.

In a first step, we perform the contrast limited adaptive histogram equalization [3] and global thresholding using the method proposed by Otsu [4] to the red, green and blue channels of the initial image into three binary images. Subsequently, a binary mask is obtained as the result of a logical OR operation of those three binary images. The detected areas of the binary mask are extended with a morphological dilation. Then, all the connected components with an area smaller than 500 pixels are removed. This is necessary for the exclusion of image artifacts that may interfere in the next steps.

This pre-processing stage allows us to have regions of interest (ROI) removed from its background. Yet, the ROI may consist of the targeted-nuclei and cytoplasm. Thus, further step is needed in order to detect the nuclei only.

2.2. Candidate of Nuclei Detection

One can identify nuclei from cytoplasm is from its darker color (low intensity) than the surrounding Cytoplasm tends to have lighter color (high intensity). For this reason we search for intensity valleys using the h-minima transform [5] in the red, green and blue channels of the outcome preprocessing image. Next, we subtract a threshold h from every pixel of the outcome image. A morphological reconstruction [6] step is then performed. In this step, we use the resulted image as a marker and the initial image as a mask. Then, we subtract the red, green and blue channels of the outcome preprocessing image with the outcome image from morphological reconstruction step.

However, the resulting images of the region-minima approach may have several markers within the targeted-nuclei. Also, there are some markers on the cytoplasm. So, edge is the best feature to impose selection only within the nuclei boundary.

Furthermore, the number of markers are reduced using a combination of Euclidean distance measure and their intensities similarity. Two markers which have shorter distances will be compared their colour intensities. Markers with low intensity will survive, whereas markers with high intensity will be lost. Illustration of this stage can be seen in Figure 2.

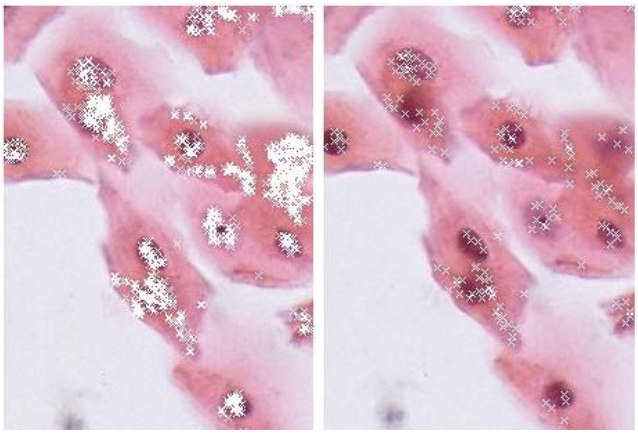


Fig. 2: nuclei detection and refinement.

2.3. Refinement of Candidate Cell Nuclei Centroids

Centroids that have been detected in the previous stage will be reduced by performing morphological dilation at each pixel where centroids are located. Next, we look for pixels with the lowest intensity as new centroids location of the area formed by the algorithm. For all the obtained centroids we apply the following rule:

```
i =1
Repeat
    ∀p = (x,y) ∈ LA(i)
    Select r = {p | min(I(p))}
    i++
Until i=nL
```

LA is an area that has a label equal to *i*, *nL* is the number of labeled area, *I* is a grayscale image of the original image.

2.4. Reduction in Number of Centroids by Distance-Metric

The number of markers are reduced using a combination of Euclidean distance-metric and their intensities similarity. Two markers which have shorter distances will be compared their color intensities. Markers with low intensity will survive, whereas markers with high intensity will be lost.

2.5. Clustering

Despite of the vast markers reduction on the candidate nuclei detection stage, there are remain some markers within nuclei or on cytoplasm. We aimed to get one marker to represent one nuclei. To achieve that, we proposed the Fuzzy C-Means Clustering as follows:

1. Two clusters are defined, one represents the nuclei and another belongs to not nuclei.
2. The classification criteria are the closest distance between markers and similarity on their intensity values.
3. Do min-max normalization on both these criteria.
4. A positive class is defined when the average intensity of the cluster is lower than the other.

This proposed algorithm is effective to identify the nuclei in one-to-one relation. The result of this stage can be seen in Figure 3.

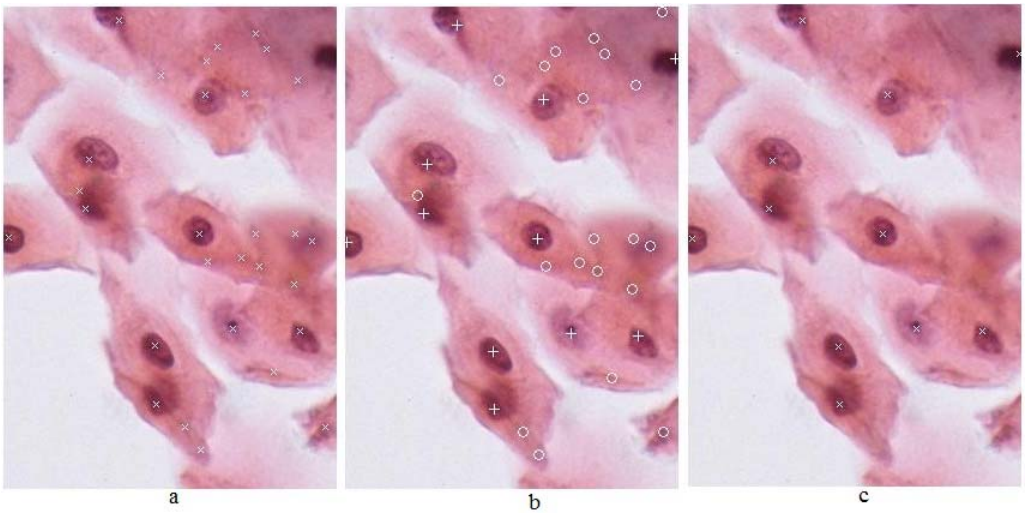


Fig. 3: Clustering stages: (a) initial image as result of distance-metric, (b) result of FCM: (+) denotes positive nuclei and (○) denotes not nuclei, and (c) one-to-one map of nuclei.

3. Evaluation method

3.1. Data

We tested our proposed methods on 11 pap smears images of NCI Bethesda System [7]. These consist of five images of medium magnification, and six of high ones. These images have 350 nuclei and have been confirmed by our pathologist [8]. This number was obtained as consistent nuclei identification by our expert [8]. It should be noted that the two readings were done in random ordered and had eight days interval in between the two readings to minimize the recall bias by our expert pathologist.

3.2. Evaluation methods

Furthermore, as a measure of the computational efficiency of the segmentation method, we present in Table I the processing times of the individual steps of the method developed in MATLAB using a Pentium 2.66 GHz and 4 GB of RAM.

TABLE 1 EXECUTION TIME OF THE PROPOSED METHOD

Step of the proposed method	Time in sec.(mean ± std)
Preprocessing	3.65 ±1.40
Candidates of nuclei detection	20.56 ± 8.58
Refinement of Candidate Cell Nuclei Centroids	3.18 ± 1.35
Reduction in Number of Centroids by Distance-Metric	2.17 ± 3.03

4. Results and Discussion

We tested our proposed methods with respect to its detection rate. The detection rate is compared with the expert truth (Table II, second column) in terms of its specificity and sensitivity. The detection rates of the proposed method can be seen in Table II. The average sensitivity rate was 90.86% and specificity rate was 81.62% which are quite promising.

TABLE 2 RESULT OF THE PROPOSED METHOD

File (*.jpg)	Expert Truth	True positive	True negative	False positive	False negative	Sensitivity	Specificity
1982	100	96	72	28	4	96%	72%
3079	122	108	122	0	14	89%	100%
6676	21	21	21	4	0	100%	84%
7214	28	25	31	2	3	89%	94%
7808	8	8	25	16	0	100%	61%

9945	24	23	14	12	1	96%	54%
6451	11	5	13	7	6	45%	65%
2064	6	6	6	0	0	100%	100%
2905	8	5	12	4	3	63%	75%
3835	15	14	19	4	1	93%	83%
9299	7	7	7	0	0	100%	100%
Sum	350	318	342	77	32	90.86%	81.62%

As can be seen in Table II, file 6451 and 2905 had a very low sensitivity and specivity rate. This was mainly caused by inhomogenity of its intensities (see Figure 4 for illustration). The cytoplasms on this image were darker than its nuclei. We modelled our methodolgy with assumption that nuclei is darker than cytoplams. Although our model correctly marks the nuclei on candidate detection stage, but these markers were lost on clustering stage.

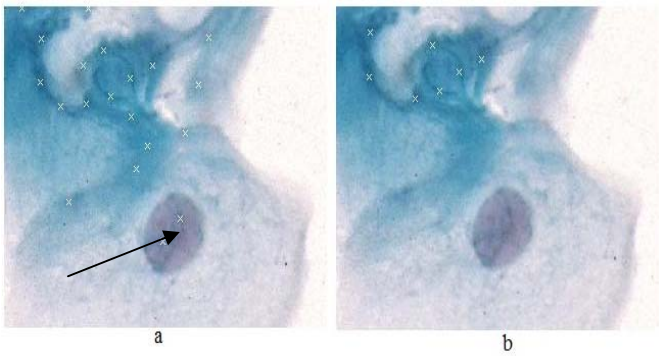


Fig. 4: False negative case. (a) Correct nuclei identification on candidate nuclei detection stage, but (b) lost on clustering stage.

On the other hand, there were cases where markers on the inflammatory cells (see arrow on Figure 5) rather than on the nuclei. This is caused by the intensity of the inflammatory cells were lower than those of the nuclei's.

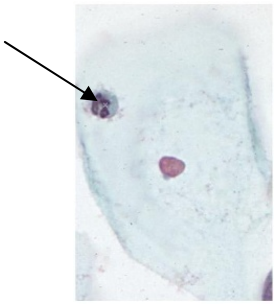


Fig. 5: Example of false positive case. The marker tends to stand on inflammatory cells rather than on the nuclei.

Despite our aim to obtain a one marker-to-one nuclei mapping, there are some nuclei that have several markers on. This is affected by variations on magnification values used in NCI Bethesda Systems [7]. Hence, further investigations need to be done to determine data-driven distance criteria for the clustering stage.

Our proposed methodology works quite well on identifying the epithelial cells in one-to-one mapping. This can serve as basis for further investigation on classifiying whether a marker stands on which type of epithelial cells. This will be the further avenue of our research.

5. References

[1] A. Garrido, N. Pérez de la Blanca, "Applying deformable templates for cell image segmentation", *Pattern Recognition*, vol. 33, 2000, pp. 821-832.

[2] ME. Plissiti, C. Nikou, and A. Charchanti, "Automated detection of Cell Nuclei in Pap Smear Images Using

Morphological Reconstruction and Clustering," *IEEE Trans on Information technology in Biomedicine.*, vol. 15, no 2, March. 2011, pp. 233-241.

[3] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*. San Diego, CA: Academic, 1994, pp. 474–485.

[4] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man Cybernetics.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[5] P. Soille, *Morphological Image Analysis: Principles and Applications*. New York: Springer-Verlag, 1999.

[6] L. Vincent, "Morphological grayscale reconstruction in image analysis: Applications and efficient algorithms," IEEE Trans. Image Process., vol. 2, no. 2, pp. 176–201, Apr. 1993.

[7] NCI Bethesda System (http://nih.techriver.net/)

[8] R. Kurniawan, "Otomatisasi Deteksi Jumlah Nuclei pada Citra Pap Smear Menggunakan Kriteria Jarak & Clustering," M.S. thesis, Dept. Informatics., Universitas Islam Indonesia, Yogyakarta, Indonesia, 2011.

[9] E. Bak, K. Najarian, and J. P. Brockway, "Efficient segmentation framework of cell images in noise environments," in Proc. 26th Int. Conf. IEEE Eng. Med. Biol., Sep., 2004, vol. 1, pp. 1802–1805.

[10] N. A. Mat Isa, "Automated edge detection technique for Pap smear images using moving K-means clustering and modified seed based region growing algorithm," Int. J. Comput. Internet Manag., vol. 13, no. 3, pp. 45–59, 2005.

Fuzzy c-Means Classifier for Incomplete Data Sets with Outliers and Missing Values

Hidetomo Ichihashi, Katsuhiro Honda
Graduate School of Engineering
Osaka Prefecture University
1-1 Gakuen-cho, Sakai, Osaka 599-8531 Japan
ichi@cs.osakafu-u.ac.jp

Abstract

A novel membership function and a fuzzy clustering approach derived from a viewpoint of iteratively reweighted least square (IRLS) techniques resolve the problem of singularity in the regular fuzzy c-means (FCM) clustering. An FCM classifier using the membership function and Mahalanobis distances makes class memberships of outliers less clear-cut, which thus resolve the problem of classification based on normal populations or normal mixtures. The ways of handling singular covariance matrices and missing values are also furnished, which improve the generalization capability of the classifier. Computational experiments show high classification performance on several well-known benchmark data sets.

1 Introduction

The regular FCM [1] assigns very fuzzy memberships to data points that are far from all cluster centroids. This property distinguishes the regular FCM from the entropy regularized FCM (FCM-e) as proved by Miyamoto [2]. Many statistical clustering approaches, such as Gaussian mixture models (GMM) [3] and deterministic annealing (DA) by Rose [4], share this property of FCM-e. Though the regular FCM is advantageous in this point, its membership functions tend to become spiky and singular in shape at cluster centroids when data set is partitioned fuzzily. This problem is mitigated by introducing a slightly different membership function to an FCM algorithm derived from the view point of iteratively reweighted least-squares (IRLS) [5] technique [6, 7]. The ways of handling singular covariance matrices and missing values are devised for IRLS-FCM, which improve generalization capability (classification accuracy on unseen objects) of the classifier. Classification performances of the trained classifier on unseen objects are tested by 10-fold cross validation (10-fold CV) with a partition into ten subsets of the same size. Computational experiments show high performance on several well-known benchmark data sets.

2 FCM Clustering with Mahalanobis Distances and IRLS

FCM clustering partitions data set by introducing memberships to fuzzy clusters. Let m dimensional vector $\boldsymbol{v}_i$ denote prototype parameter (i.e., cluster centroid). u_{ik} denotes the membership of k-th object data to i-th cluster.

$$d_{ik}^2 = (\boldsymbol{x}_k - \boldsymbol{v}_i)^\top A_i^{-1} (\boldsymbol{x}_k - \boldsymbol{v}_i) \tag{1}$$

is a measure of distance from $\boldsymbol{x}_k$ to i-th cluster prototype, where A_i is a covariance matrix of data samples of the i-th cluster.

$$A_i = \frac{\sum_{k=1}^n u_{ik} (\boldsymbol{x}_k - \boldsymbol{v}_i)(\boldsymbol{x}_k - \boldsymbol{v}_i)^\top}{\sum_{k=1}^n u_{ik}}. \tag{2}$$

$$\boldsymbol{v}_i = \frac{\sum_{k=1}^n u_{ik} \boldsymbol{x}_k}{\sum_{k=1}^n u_{ik}}. \tag{3}$$

$$\pi_i = \frac{\sum_{k=1}^n u_{ik}}{\sum_{j=1}^c \sum_{k=1}^n u_{jk}} = \frac{1}{n} \sum_{k=1}^n u_{ik}. \tag{4}$$

The M-estimators [5, 8] try to reduce the effect of outliers by replacing the squared residuals with ρ-function, which is chosen to be less increasing than square. Instead of solving directly this problem, we can implement it as the IRLS and minimize.

$$L_{IRLS} = \sum_{i=1}^c \sum_{k=1}^n w(d_{ik}) d_{ik}^2. \tag{5}$$

Proceedings of the 2005 International Conference on Computational Intelligence for Modelling, Control and Automation, and International Confe
Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'05)
0-7695-2504-0/05 $20.00 © 2005 IEEE

where we set as $w(d_{ik}) = u_{ik}$. See [7] for detail derivation of the IRLS-FCM algorithm. While IRLS approaches in general does not guarantee the convergence to a global minimum, experimental results have shown reasonable convergence points.

To facilitate competitive function in the algorithm we need to define the weight function to be normalized as:

$$u_{ik} = \frac{u^*_{ik}}{\sum_{l=1}^{c} u^*_{lk}}. \tag{6}$$

Examples of u^* which made the IRLS stably convergent in our experiment were as follows

$$\begin{aligned} u^{*(1)}_{ik} &= \pi_i/(\eta + d^2_{ik}/0.1)^{1/\lambda}|A_i|^{-1/\gamma}, & (7)\\ u^{*(2)}_{ik} &= \pi_i\lambda/(\lambda + d^2_{ik})^2|A_i|^{-1/\gamma}, & (8)\\ u^{*(3)}_{ik} &= \pi_i/(d^2_{ik})^{1/\lambda}|A_i|^{-1/\gamma}, & (9)\\ u^{*(4)}_{ik} &= \pi_i\exp(-d^2_{ik}/\lambda)|A_i|^{-1/\gamma}. & (10) \end{aligned}$$

The clustering algorithm can be written as follows:

Algorithm: IRLS-FCM

Step 1: Initialize $u_{ik}, i = 1, ..., c, k = 1, ..., n$ randomly.

Step 2: Calculate $\boldsymbol{v}_i, i = 1, ..., c$ by using (3) .

Step 3: Calculate A_i and $\pi_i, i = 1, ..., c$ by using (2) and (4) .

Step 4: Calculate $u_{ik}, i = 1, ..., c, k = 1, ..., n$ by using (6) and (7).

Step 5: If iteration number exceeds the predetermined value then terminate, else go to *Step 2*.

The termination criteria is simplified in the above algorithm to reduce the computation, but it should be recommended to check the convergence of the objective function (5). Note that (8) is the weight function of Geman-McClure M-estimator and λ is a scale parameter that controls the convexity of the robust ρ function. Equation (9) produces an algorithm similar to the Gath-Geva's modified FCM, which thus suffers from the problem of regular FCM known as singularity. Characteristics of the FCM families are compared in [7].

After completing the clustering for each class, the classification is performed by computing class memberships. Let α_q denote the mixing proportion of class q, i.e., the $a\ priori$ probability of class q. If we adopt (7), the class membership of k-th data $\boldsymbol{x}_k$ to class q is computed using

$$\begin{aligned} u^{*(1)}_{qjk} &= \pi_{qj}/(1 + d^2_{qjk}/0.1)^{1/\lambda}|A_{qj}|^{-1/\gamma}, & (11)\\ \tilde{u}_{qk} &= \alpha_q \sum_{j=1}^{c} u^{*(1)}_{qjk} & (12) \end{aligned}$$

and normalizing as in (6). It should be noted that if we use (10) with $\alpha = 2, \gamma = 2$ and homoscedastic A_{qi}, then the IRLS-FCM is the same as probabilistic neural network (PNN) by Streit and Luginbuhl [9].

3 Handling Singular Matrices

The algorithm using Mahalanobis distances should include steps to check that the covariance matrices are nonsingular and hence invertible. If a covariance matrix have an entry on the diagonal which is approaching zero then the algorithm terminates. When estimating covariance structures in high-dimensions, but not over-constraining model flexibility, Tipping and Bishop proposed a way to control the number of parameters in the mixture of probabilistic principal component analysis (MPCA) [10]. This approach was applied to fuzzy c-varieties (FCV) clustering and discussed its close relationship with MPCA in [11].

Eigenvalue decomposition of a covariance matrix A_i is written as:

$$A_i = P_i \Delta_i P_i^\top, \tag{13}$$

where P_i is an $r \times r$ matrix of eigenvectors of A_i. Δ_i =diag$(\delta_{i1}, ..., \delta_{ir})$ is an $r \times r$ diagonal matrix of eigenvalues. r equals the dimensionality of input samples.

Let A'_i denotes an approximation of A_i in (7)-(10) as:

$$A'_i = P^p_i(\Delta^p_i - \sigma_i I_r)P^{p\top}_i + P_i(\sigma_i I_r)P_i^\top, \tag{14}$$

where P^p_i is an $r \times p$ matrix of eigenvectors corresponding to the p largest eigenvalues, where $p < r - 1$. P^p_i is an $r \times p$ matrix and Δ^p_i is a $p \times p$ diagonal matrix. p is chosen so that all A'_is are nonsingular and the classifier maximizes its generalization capability.

$$\sigma_i = (\mathrm{trace}(A_i) - \Sigma^p_{l=1}\delta_{il})/(r - p). \tag{15}$$

Inverse of A'_i is

$$A'^{-1}_i = P^p_i((\Delta^p_i)^{-1} - \sigma_i^{-1} I_r)P^{p\top}_i + \sigma_i^{-1} I_r. \tag{16}$$

Proceedings of the 2005 International Conference on Computational Intelligence for Modelling, Control and Automation, and International Confe
Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'05)
0-7695-2504-0/05 $20.00 © 2005 IEEE
IEEE COMPUTER SOCIETY

4 Handling Missing Values

Missing values are common in many real world data sets. Usually entire incomplete data samples with missing values are eliminated in preprocessing for clustering and classifying applications. In [12, 13], not the entire data samples but only the missing values are ignored by multiplying "0" weights over the corresponding reconstruction errors. In this section, we propose an approach to clustering and classification without eliminating or ignoring missing values but with estimating the values. The estimation is done by using least square method of fuzzy Mahalanobis distances. Equation (5) is the weighted sum of the squared fuzzy Mahalanobis distances between data points and cluster centroids. The missing values are some elements of data vector $\boldsymbol{x}_k$, which are estimated by the least square technique, that is, the missing elements are the solution to the system of linear equations derived from differentiating (1) with respect to the missing elements of $\boldsymbol{x}_k$.

Let $x_{kli}, l = 1, ..., r$, be the elements of centered $\boldsymbol{x}_k$ (i.e., $x_{kli} = x_{kl} - v_{il}$) and the j-th element x_{kji} be a missing element. Then the system of linear equations can be written as:

$$\begin{pmatrix} 2A_i^{-1} & U \\ U & Z \end{pmatrix} \boldsymbol{x}_{ki}^* = \boldsymbol{b}_{ki}, \tag{17}$$

where

$$U = \mathrm{diag}(1 \cdots 1\ 0_j\ 1 \cdots 1), \tag{18}$$

$$\boldsymbol{x}_{ki}^* = (x_{k1i}^* \cdots x_{kri}^*\ \mu_1 \cdots \mu_{j-1}\ 0\ \mu_{j+1} \cdots \mu_r), \tag{19}$$

and

$$\boldsymbol{b}_{ki} = (0_1 \cdots 0_r\ x_{k1i} \cdots x_{k\ j-1\ i}\ 0\ x_{k\ j+1\ i} \cdots x_{kri}). \tag{20}$$

"diag" denotes diagonal matrix and 0_j denotes that the j-th element is zero. Z is an $r \times r$ zero matrix. When more than two elements of $\boldsymbol{x}_k$ are missing, corresponding elements in (18)-(20) are also replaced by 0. All zero rows and all zero columns are eliminated from (17) and then we obtain the least square estimates of all the missing values by adding v_{il} to x_{kli}. We use the estimated values for each cluster in both the clustering and classifying.

5 Numerical Experiments

Figs.1-4 show the clustering results where all A_i's are unit matrices and π_i's are all 1 for simplicity's sake. In each figure, (d) shows the result depicted by the defuzzification using the maximum membership approach. The stars represent cluster centroids and the membership function of the cluster depicted by circles is shown in (a), that of the cluster depicted by triangles is shown in (b), and that of the cluster depicted by squares is shown in (c). Note that as the partition is fuzzy, there are no precise boundaries, so we give curved surface of membership function instead. Fig.1 shows the fuzzily clustered result by the regular FCM where λ denotes the weighting exponent on u_{ik}. Densely populated areas are captured by cluster centroids. By FCM-e, the memberships of points that are far from the center of mass of the training set become close to 1 or 0, so the center of the membership function moves toward outside the figure, though the cluster centroids in Fig.1 and Fig.4 are almost in the same locations. In other words, the regular FCM is more robust than FCM-e. As shown in Fig.1, membership functions become spiky at the centroids by the regular FCM and are singular in shape when the data set is partitioned fuzzily. This is alleviated by introducing $u^{*(1)}$ as shown in Fig.2. Fig.4 shows the results by the regular FCM with a large fuzzifier ($\lambda = 7$), namely a high temperature in the terminology of DA. Despite the fuzzy clustering result with the high temperature, the cluster centroids do not come close together and the phase transition in DA, which changes the number of effective clusters, does not occur.

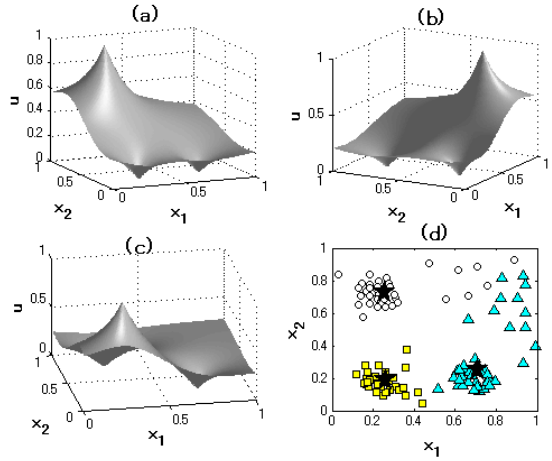


Figure 1. Clustering results by regular FCM with $\lambda = 3$.

The regular FCM gives ambiguous memberships for data points that are far from all the centroids as shown in Fig.1. This property distinguishes the regular FCM from FCM-e ($u^{*(4)}$) as proved in [2]. Fig.3 shows the fuzzily clustered result by FCM-e. The memberships of points that are far from all cluster centroids become close to 1 or 0 as shown in Fig.3 (b) (see the membership values of the points in close proximity to $x_1 = 1$ and $x_2 = 1$).

Figs.5-6 show the clustering results for an artificial 2D data formed by two ellipse clusters. Mhalanobis distances defined by (1)-(4) and (6) are used. The first row shows

Proceedings of the 2005 International Conference on Computational Intelligence for Modelling, Control and Automation, and International Confe
Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'05)
0-7695-2504-0/05 $20.00 © 2005 IEEE

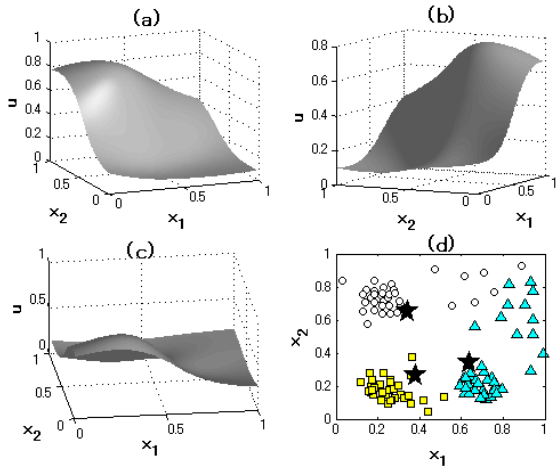


Figure 2. Clustering results by IRLS-FCM with $u^{*(1)}$ and $\lambda = 0.5$.

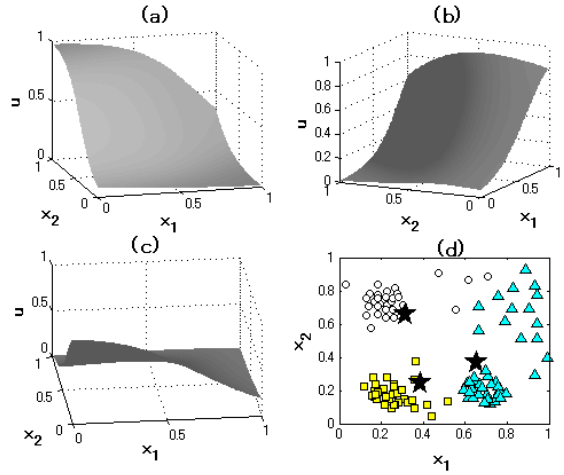


Figure 3. Clustering results by FCM-e with $\lambda = 0.12$.

3D graphics of posterior probability or cluster membership. The second row shows contour curves of nonnormalized membership function. The third row shows the clustering results. Fig.5 shows the results by GMM, which is equivalent to IRLS-FCM with $u_{ik}^{*(4)}$, $\lambda = 2$ and $\gamma = 2$, and Fig.6 shows the results by IRLS-FCM with $u^{*(1)}$, $\lambda = 1$ and $\gamma = 1$. By GMM, the points near the origin (0, 0) are clustered crisply and their memberships (i.e., posterior probabilities) are close to 0 or 1. But they are clustered fuzzily by IRLS-FCM with $u^{*(1)}$, that is, those memberships are nearly 0.5.

Fig.7 shows the classification results of the iris 2D data with 3 classes by IRLS-FCM with $u^{*(1)}$, $\lambda = 1, \gamma = 20$, p=2 and $c = 1$, where only the two variables, namely petal length (x_3) and petal width (x_4), are used and 5 new points are added at left upper and right bottom corners. Classes are shown by closed (filled) and open (white) marks of circles, triangles and squares. The open marks represent given classes and closed marks represent estimated classes. 5-Fold cross validation (5-fold CV) was deployed for evaluation and 9 points including all artificially added points were misclassified (error rate was 5.81%). The error rate was comparable to the classification based on the normal populations or the quadratic discriminant ($u_{ik}^{*(4)}$, $\lambda = 2, \gamma = 2$ and $c = 1$). The maximum class memberships of the artificially added 5 points were between 0.40 and 0.70, whereas the posterior probabilities by the normal distributions were between 0.96 and 1.0. By the proposed membership function, the classifier gives an ambiguous decision for outliers. The decision is deemed to be more prudent and rational than those by the classifier based on posterior probabilities and normal mixture (i.e., PNN), because the additional five test points are far from all class centroids.

To compare our results on the iris data to other algorithms, we cite the results of C4.5 and its Bagged and Boosted versions from [14] and Clustering Genetic Programming from [15] in Table 1. The algorithm is evaluated using a 10-fold CV, which is applied to each data set 10 times using a default partition into ten subsets of the same size. The performance is the average, standard deviation (s.d.), best and worst of misclassification error (%) over 10 folds. $u^{*(4)}$ in Table 1 represents the membership function in (10), but the resultant algorithm is equal to PNN except that PNN uses homoscedastic mixture (i.e., $A_{qj} = A$). When $c = 1$, it is the same as the classifier based on normal distributions with unequal covariance matrices, which thus results in a quadratic classifier. $u^{*(1)}$ represents the proposed IRLS-FCM whose membership function resolves the singularity in the regular FCM. The iteration number is fixed to 50 when $c > 1$ that was adequate for the objective function value to converge in our experiments. We selected the best parameter values and cluster number from numerous experiments using 10-fold CV with a default partition. For deterministic algorithms, the standard deviation (s.d.), best and worst performances are not reported. 9) in Table 1 show the result by the proposed algorithm with unit covariance matrices $A_{qj} = I$, hence $|A_{qj}|$=1 and the mixing proportion $\pi_{qj} = 1$. If we look at the results of our algorithm on the iris data set in Table 1 we see that the best performance is achieved by our algorithm with $u^{*(1)}$.

10) in Table 1 shows the result of discriminating only the iris versicolor from the other two iris subspecies, i.e., two-class problem. As shown in 11), the classification performance is clearly improved by increasing the number of

Proceedings of the 2005 International Conference on Computational Intelligence for Modelling, Control and Automation, and International Confe
Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'05)
0-7695-2504-0/05 $20.00 © 2005 IEEE
IEEE COMPUTER SOCIETY

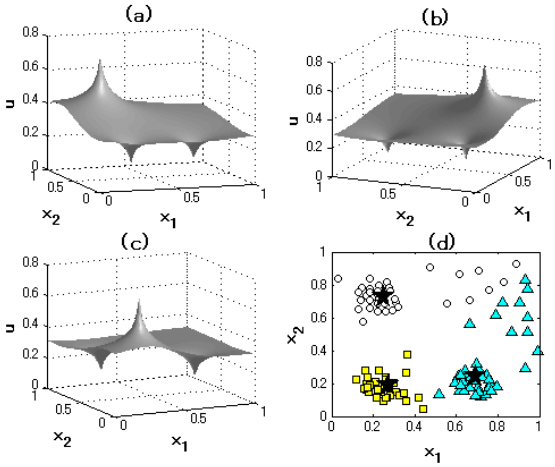


Figure 4. Clustering results by regular FCM with $\lambda = 7$.

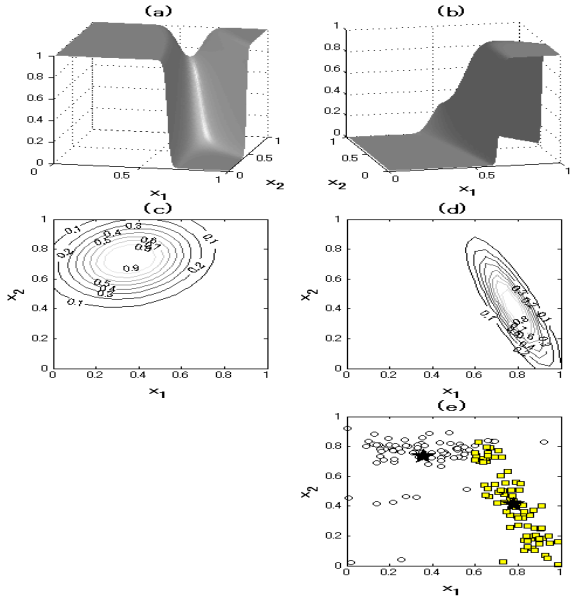


Figure 5. Clustering results by GMM with $\lambda = 2, \gamma = 2$.

clusters to two in each class. This example shows that we need to implement the clustering algorithm, though IRLS-FCM produces high classification performance on many data sets when $c = 1$. Similar results are obtained for the iris verginica as shown in 12) and 13). 14) shows that for distinguishing the iris setosa, we need only one cluster in each of the two classes.

We used 8 data sets: Iris, Wisconsin breast cancer, Ionosphere, Glass, Liver disorder, Diabetes, Sonar and Wine. These data sets are available from the UCI ML repository (http://www.ics.uci.edu/ mlearn/). All attribute values of each data set were normalized to zero mean and unit variance.

Table 2-4 shows the results. Incomplete samples in the breast cancer were eliminated from the training and test sets. In case when A_{qj} was singular or noninvertible, we reduced the number p and for some data sets the smaller p improved generalization capability. The results by 10-fold CV with randomized partitions are shown in Table 4. The simplified version ($A_{qj} = I$) of our algorithm did not perform up to its potential except for the breast cancer data and sonar data, and the fuzzy Mahalanobis distances improved the performance of our IRLS-FCM for the other 6 data sets though we used only one cluster for each class (i.e., $c = 1$). The boldface letters in Tables 4-5 indicate that those classification error rates are smaller than the other approach according to the paired t test with significance level $p = 0.05$ in the comparison between IRLS-FCM (Table 4) and the k-nearest neighbor classifier (k-NN) in Table 5.

Despite the continuous increase in memory capacity and CPU speed, especially in data mining, storage and efficiency issues become even more and more prevalent. For this reason we also measured the compression ratios ($(p+1) \times c \times$ number of classes)/ (number of data samples) of the trained classifiers. The k-NN is a well-established high performance classifier in many domains but the compression ratio is 100% and no good. Table 5 shows the average classification error rate $\pm$ standard deviation (s.d.) of k-NN by our numerical experiment with normalized data (i.e., attribute values are with zero mean and unit variance). The results without normalization reported in [17] (in the parentheses), and that by the supervised learning vector quantization (LVQ1) [16] reported in [17] are also shown. Compression ratios of LVQ1 are displayed with (%) in Table 5. LVQ1 performance reported in [17] is slightly lower than that of k-NN. The classification performance of the proposed IRLS-FCM classifier is better than or equal to k-NN for all 8 data sets. The efficiency of the proposed method in terms of memory consumption is also high.

Fig.8 shows the classification and estimation results on the iris 2D data of 2-class problem with missing values by IRLS-FCM (10-fold CV) with $u^{*(1)}$, $\lambda = 0.5, \gamma = 2.5$ and $c = 1$. Only the two variables, namely x_3 and x_4, are used. The missing values are depicted by solid lines. Squares represent the corresponding estimated values. Because the number of clusters is one ($c = 1$) for each class, the single class consisted of two subspecies (i.e., setosa and verginica) forms a slim and long ellipsoidal cluster. Therefore, the missing values on the right side and on the upper side are

Proceedings of the 2005 International Conference on Computational Intelligence for Modelling, Control and Automation, and International Confe
Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'05)
0-7695-2504-0/05 $20.00 © 2005 IEEE

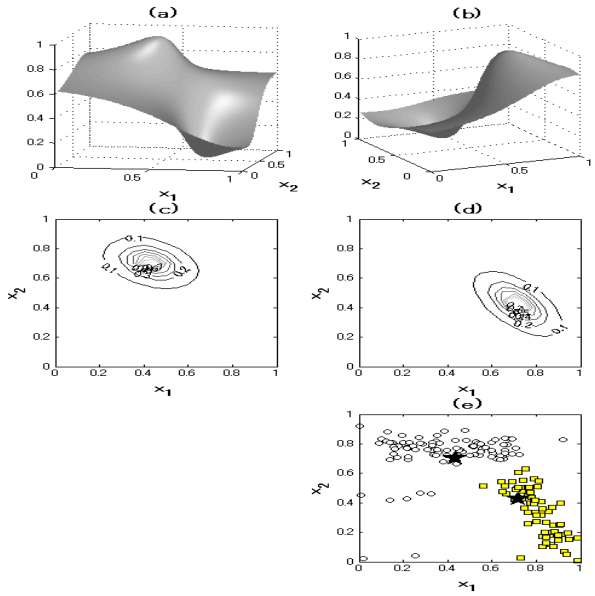


Figure 6. Clustering results by IRLS-FCM with $u^{*(1)}$, $\lambda = 1$ and $\gamma = 1$.

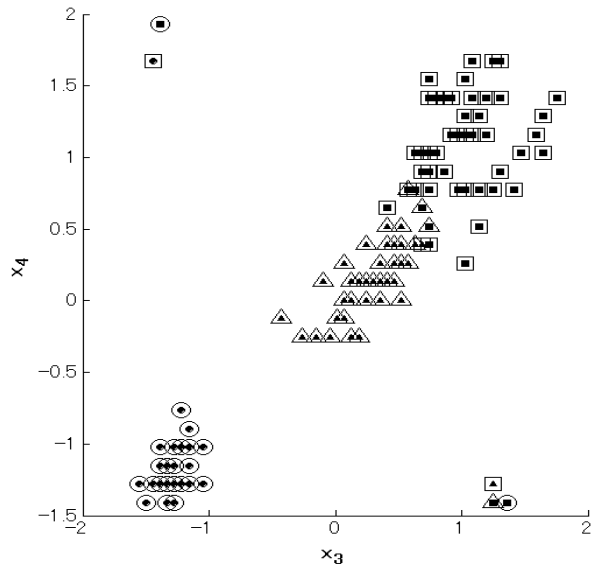


Figure 7. Classification results (5-fold CV) of Iris 2D data by IRLS-FCM with $u^{*(1)}$, $\lambda = 1, \gamma = 20$ and $c = 2$.

estimated at upper right corner of the figure. This is alleviated by increasing the cluster number to two as shown in Fig.9.

The breast cancer data consisted of 699 data samples includes 16 missing feature values. We tested the classifier performance using the missing value handling option on this data set with $p = 9, \lambda = 0.62, \gamma = 10$ and $c = 1$. By 10-fold CV, 25 samples were misclassified including only 1 sample with a missing attribute value. The number of miss-classified data is decreased by one compared with the result on the set without missing attribute values. The performance was not deteriorated by including incomplete data with missing values.

6 Conclusion

We have proposed an FCM classifier based on IRLS. High classification performance on several well-known benchmark data sets was achieved. The posterior probabilities from normal populations for classification decision are replaced by the class memberships which are more ambiguous for outliers. Thorough comparisons between the proposed classifier and the classifier based on normal population or normal mixtures in terms of receiver operating characteristics are now undergoing.

References

[1] J. C. Bezdek: *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, 1981.

[2] Z.-Q. Liu, S. Miyamoto (Eds.), *Softcomputing and Human-Centered Machines*, Springer-Verlag, 2000.

[3] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.

[4] K. Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems," *Proc. of the IEEE*, vol.86, no.11, pp.2210-2239, 1998.

[5] P. W. Holland, and R. E. Welsch, "Robust regression using iteratively reweighted least-squares, "*Communications in Statistics*, vol. A6, no. 9, pp. 813-827, 1977.

[6] H. Ichihashi, K. Honda: , "FCM clustering from the view point of iteratively reweighted least squares," *Proc. of 2005 IEEE International Conference on Fuzzy Systems*, pp.873-878, 2005.

[7] H. Ichihashi, K. Honda, T. Yagi, "FCM classifier based on IRLS and fuzzy Mahalanobis distances," *Proc. of the 6th International Symposium on Advanced Intelligent Systems*, Yeosu, Korea, 2005.

Proceedings of the 2005 International Conference on Computational Intelligence for Modelling, Control and Automation, and International Confe
Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'05)
0-7695-2504-0/05 $20.00 © 2005 IEEE

Table 1. Results on Iris data set by 10-fold CV with a default partition

	algorithms and parameters
1)	Clustering Genetic Programming
2)	C4.5
3)	Bagged C4.5
4)	Boosted C4.5
5)	normal distribution, $u^{*(4)}, \lambda = 2, \gamma = 2, c = 1$
6)	PNN, $u^{*(4)}, \lambda = 2, \gamma = 2, p = 2, c = 2$
7)	$u^{*(1)}, \lambda = 1.2, \gamma = 8, \eta = 1, c = 1$
8)	$u^{*(1)}, \lambda = 1.2, \gamma = 8, \eta = 1, c = 2$
9)	$u^{*(1)}, \lambda = 0.7, A_{qj} = I, \gamma = \pi_{qj} = 1, \eta = 1, c = 6$
10)	versicolor, $u^{*(1)}, \lambda = 2, \gamma = 5.1, \eta = 1, c = 1$
11)	versicolor, $u^{*(1)}, \lambda = 0.6, \gamma = 3, \eta = 1, c = 2$
12)	verginica, $u^{*(1)}, \lambda = 1, \gamma = 2, \eta = 1, c = 1$
13)	verginica, $u^{*(1)}, \lambda = 0.5, \gamma = 2.5, \eta = 1, c = 2$
14)	setosa, $u^{*(1)}, \lambda = 1, \gamma = 2, \eta = 1, c = 1$

	average	s.d.	best	worst
1)	2.1	0.2	2.0	2.7
2)	5.9			
3)	5.0			
4)	5.0			
5)	2.0			
6)	3.73	0.68	2.67	4.67
7)	1.33			
8)	1.33	0.0	1.33	1.33
9)	3.33	0.0	3.33	3.33
10)	6.00			
11)	1.33	0.0	1.33	1.33
12)	3.33			
13)	1.27	0.36	0.67	2.00
14)	0.0			

Table 2. Datasets and parameters of IRLS-FCM ($u^{*(1)}$) used in the experiments

	data sets, number of attributes, samples and classes, and parameters
Ir	Iris, 4, 150, 3, $\lambda = 1.2, \gamma = 8, \eta = 1, p = 4, c = 1$
Br	Wisconsin breast cancer, 9, 683, 2, $\lambda = 0.4, A_{qj} = I, \pi_{qj} = \alpha = 1, \eta = 1, c = 30$
Io	Ionosphere, 33, 351, 2, $\lambda = 0.9, \gamma = 40, \eta = 35, p = 4, c = 1$
Gl	Glass, 9, 214, 6, $\lambda = 1, \gamma = 22, \eta = 8, \alpha = 1, p = 4, c = 1$
Li	Liver, 6, 345, 2, $\lambda = 1, \gamma = 7, \eta = 1, p = 6, c = 1$
Di	Diabetes, 8, 768, 2, $\lambda = 0.5, \gamma = 15, \eta = 1, p = 5, c = 1$
So	Sonar, 60, 208, 2, $\lambda = 0.2, A_{qj} = I, \pi_{qj} = \alpha = 1, \eta = 1, c = 20$
Wi	Wine, 13, 178, 3, $\lambda = 1, \gamma = 30, \eta = 1, \alpha = 1, p = 10, c = 1$

Table 3. Results by 10-fold CV with a default partition

	average	s.d.	best	worst
Ir	1.33			
Br	2.80	0	2.80	2.80
Io	4.29			
Gl	29.52			
Li	29.71			
Di	23.82			
So	11.10	0.70	10.00	12.50
Wi	0			

[8] P. J. Huber. *Robust Statistics*. New York:Wiley, first edition, 1981.

[9] R. L. Streit and T. E. Luginbuhl, "Maximum likelihood training of probabilistic neural networks," *IEEE Transactions on Neural Networks*, vol.5, no.5, pp.764-783, 1994.

[10] M.E. Tipping, C.M. Bishop, "Mixtures of probabilistic principal component analysers, "*Neural Computation*, Vol.11, pp.443-482, 1999.

[11] K. Honda, H. Ichihashi, "Regularized Linear Fuzzy Clustering and Probabilistic PCA Mixture Models," *IEEE Trans. Fuzzy Syst.*, vol.13, no.4, pp.508-516, 2005.

[12] K. Honda, H. Ichihashi, "Linear fuzzy clustering techniques with missing values and their application to local principal component analysis," *IEEE Trans. Fuzzy Syst.*, vol.12, no.2, pp.183-193, 2004.

[13] K. Honda, H. Ichihashi, "Component-wise robust linear fuzzy clustering for collaborative filtering," *Int. J. Approximate Reasoning.*, vol.37, no.2, pp.127-144, 2004.

[14] Y. Freund and R. Schapire, "Experiments with a new boosting algorithm, "*In Proc. 13th International Conference on Machine Learning*, pp.148-146, Morgan Kaufmann, 1996.

[15] J. Eggermont, J.N. Kok and W.A. Kosters, "Genetic programming for data classification: Partitioning the

0-7695-2504-0/05 $20.00 © 2005 IEEE

Table 4. Results by 10-fold CV with random partitions

	average	s.d.	best	worst	compression
Ir	**2.27**	0.33	2.00	2.67	10.67 %
Br	**2.84**	0.11	2.65	3.09	8.78 %
Io	**4.80**	0.21	4.57	5.14	2.85 %
Gl	**30.71**	1.77	28.10	34.76	14.02 %
Li	**31.21**	0.99	29.12	32.35	4.06 %
Di	**24.28**	0.35	23.55	24.87	1.56 %
So	**11.15**	1.32	9.00	13.50	19.23 %
Wi	**0.41**	0.46	0.0	1.18	18.54 %

Table 5. Classification error rates $\pm$ s.d. of k-NN. Results by 10-fold CV with random partitions. We used attribute values with zero mean and unit variance and the results without normalization reported in [17] are also shown in the parentheses. Compression ratios of LVQ1 are displayed with (%).

	k-NN	LVQ1	
Ir	3.73 ± 0.44 (3.3 ± 0.6)	3.7 ± 0.4	5.9 %
Br	3.32 ± 0.18 (3.0 ± 0.2)	3.7 ± 0.4	5.9 %
Io	15.46 ± 0.63 (13.9 ± 0.7)	13.6 ± 0.8	6.8 %
Gl	**29.57** ± 1.23 (27.7 ± 1.2)	31.7 ± 2.0	45 %
Li	34.71 ± 1.34 (32.7 ± 1.6)	33.7 ± 1.9	8.4 %
Di	**24.22** ± 0.37 (25.3 ± 0.7)	26.5 ± 0.9	3.4 %
So	15.10 ± 0.77 (18.2 ± 1.4)	21.7 ± 2.4	19 %
Wi	2.29 ± 0.49 (26.1 ± 1.9)	27.3 ± 1.5	32 %

search space, , "*In Proc. 19th Annual ACM Symposium on Applied Computing*, Nicosia, Cyprus, March 14-17, pp. 1001-1005, 2004.

[16] T. Kohonen, *Self-Organization and Associative Memory*, 3rd Edn. Springer, Berline 1989.

[17] C.J. Veenman, M.J.T. Reinders, "The nearest subclass classifier: a compromise between the nearest mean and nearest neighbor classifier," *IEEE Transactions on PAMI*, vol.27, no.9, pp.1417-1429,2005.

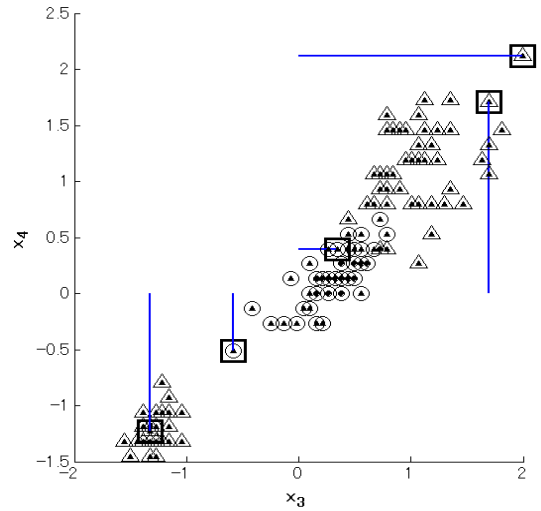


Figure 8. Classification and estimation results on iris 2-class 2D data with missing values by IRLS-FCM (10-fold CV) with $u^{*(1)}$, $\lambda = 0.5, \gamma = 2.5$ and $c = 1$.

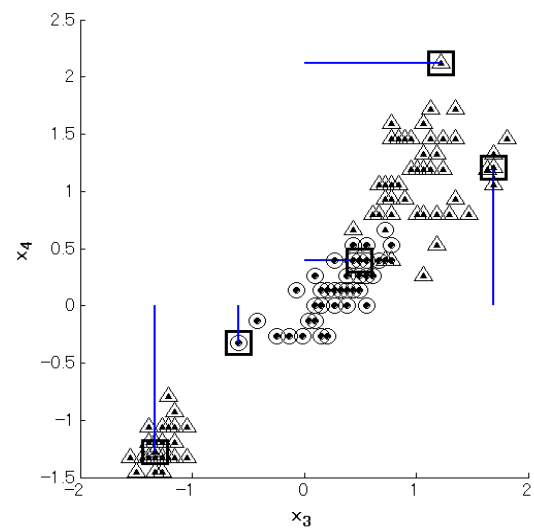


Figure 9. Classification and estimation results on iris 2-class 2D data with missing values by IRLS-FCM (10-fold CV) with $u^{*(1)}$, $\lambda = 0.5, \gamma = 2.5$ and $c = 2$

Proceedings of the 2005 International Conference on Computational Intelligence for Modelling, Control and Automation, and International Confe
Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'05)
0-7695-2504-0/05 $20.00 © 2005 IEEE

Tulips University of Tsukuba Repository

Fuzzy c-means as a regularization and maximum entropy approach

著者別名	宮本 定明
内容記述	the seventh International Fuzzy Systems Association World Congress, June 25-29, 1997, University of Economics, Prague, Czech Republic
journal or publication title	IFSA'97 Prague : proceedings [of the] seventh International Fuzzy Systems Association World Congress
page range	86-92
year	1997-06
URL	http://hdl.handle.net/2241/118283

FUZZY C - MEANS AS A REGULARIZATION AND MAXIMUM ENTROPY APPROACH

Sadaaki Miyamoto[a], Masao Mukaidono[b]

[a] *University of Tsukuba, Institute of Information Sciences and Electronics, Tsukuba-shi, Ibaraki 305, Japan*
[b] *Department of Computer Science, Meiji University, Kawasaki-shi 214, Japan*

Abstract. A maximum entropy method has been proposed by one of the authors as a variation of the fuzzy c - means. In this paper this method and the standard fuzzy c-means are regarded as *regularizations* of the crisp k - means. As a result, further variations of the maximum entropy method are developed that are parallel to the corresponding methods of the ordinary fuzzy c - means. Moreover, the fuzzy k - nearest neighbor classification rule which has been derived from the standard fuzzy c - means is transformed into the corresponding method within the maximum entropy framework.

Keywords. Fuzzy c - means, maximum entropy, regularization, fuzzy classifier, k - nearest neighbor rule.

1 Introduction

Fuzzy c - means clustering (abbreviated as FCM) [5, 6, 2], which is the fuzzy version of the k - means [17], has been studied by many researchers. Recently, one of the authors has proposed a maximum entropy method [15, 16] that can be used instead of the standard FCM.

In this paper we show that the standard fuzzy c - means and the maximum entropy method are different types of *regularizations* for the crisp k - means.

Regularization is an old technique to solve ill-posed problems of functional equations [23] and has been applied to many real problems. For example, the ridge estimator in regression analysis [10] is closely related to regularization. Regularization in general implies modification of a given problem that is singular in some sense into a regular problem. The singular problem is difficult to solve but the latter problem is easier to handle. The latter problem is called the regularization of the original problem when the solution of the regularized problem approximates the original solution.

We recognize fuzzy c - means to be a regularization for the crisp k - means. It is strange to say that the well-known method of k - means is singular, since the solution of the crisp k - means is by no means ill-posed.

Although the crisp k - means is a well-posed problem, the crisp solution is characterized by the extremal points when the optimal clustering is generalized into the fuzzy membership case [2]. Thus, we regard extremal points as singular solutions, whereas intermediate points are considered to be regular solutions.

This idea is fruitful in the sense that we have possibility of other regularizations than the standard one by Dunn [5, 6] and Bezdek [2].

The concept of the regularization in FCM has briefly been suggested in Miyamoto [20]. In this paper we show how this concept works for the maximum entropy approach. As a result we obtain a number of new algorithms that are parallel to the standard FCM.

Moreover, a method of supervised fuzzy classification that has been derived from the k - nearest neighbor rule [4] can be transformed into the corresponding method of classification by the maximum entropy approach.

Thus, what we present is not a single method but a framework of maximum entropy in which we can develop a family of new methods.

2 Fuzzy c - means as a regularization

2.1 Review of fuzzy c - means

Let us briefly review fuzzy c - means for discussing regularizations. In the following we are referring to description by Bezdek [2].

Let us first assume that the individuals to be clustered be $X = \{x_1, ..., x_n\}$ and suppose that $x_1, ..., x_n$ are points of p dimensional Euclidean space, although we consider other spaces later. Remark also that the number of clusters is given by c. Hereafter we refer to the methods as c - means, not as k - means.

The introduction of the standard fuzzy c - means starts from a version of crisp c - means algorithm formulated as an optimization problem. Namely, the function to be minimized is

$$J_1(U, v) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik} d(x_k, v_i), \tag{1}$$

where $U = (u_{ik})$ and u_{ik} is a binary variable showing whether x_k belongs to the cluster i ($u_{ik} = 1$) or not

$(u_{ik} = 0)$; $v = (v_1, ..., v_c)$ and each v_i is the center for the cluster i. $d(x_k, v_i)$ is a dissimilarity measure between x_k and v_i, and for the most part $d(x_k, v_i) = \|x_k - v_i\|_2^2$ where $\|\cdot\|_2$ is the Euclidean norm. Since each individual belongs to one and only one cluster, the admissible set for u_{ik} is

$$M_c = \{(u_{ik}) | \sum_{i=1}^{c} u_{ik} = 1, u_{ik} \in \{0,1\}, k = 1, ..., n\}.$$

In most cases of different c - means, a two stage alternative optimization procedure is used for minimization of the objective function.

Procedure CM.

CM1. Set initial values for $\bar{U}$ and $\bar{v}$.

CM2. Minimize $J(U, \bar{v})$ with respect to $U \in M$ and let the optimal solution be $\bar{U}$.

CM3. Minimize $J(\bar{U}, v)$ with respect to v and let the optimal solution be $\bar{v}$.

CM4. Check the stopping criterion. If the criterion is not satisfied, go to **CM2**.

If we put $J = J_1$ and the admissible set $M = M_c$, the above procedure is reduced to a standard algorithm of crisp c -means which is sometimes referred to as Forgy's method [1, 7].

The first step toward the fuzzy c - means is to generalize the binary valued U to fuzzy U. Thus, u_{ik} may take any value in the unit interval, and therefore the admissible set becomes

$$M_f = \{(u_{ik}) | \sum_{i=1}^{c} u_{ik} = 1, u_{ik} \in [0,1], k = 1, ..., n\}.$$

This generalization is insufficient. Suppose that we use the procedure **CM** with $J = J_1$ and $M = M_f$. We obtain crisp solutions, since the objective function is linear for u_{ik} and hence the step **CM2** is a linear programming. It is well-known that the optimal solution for a linear programming is on an extremal point: in this case u_{ik} is zero or unity.

Consequently, Dunn [5, 6] and Bezdek [2] studied a modified criterion

$$J_m(U, v) = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m d(x_k, v_i)$$

using a parameter $m > 1$. By using this parameter, we have fuzzy memberships; it is easily seen that $0 < u_{ik} < 1$, unless the point x_k is on a center v_i.

It is well-known that the solution for $\bar{U}$ with a given $\bar{v}$ in step **CM2** is

$$\bar{u}_{ik} = [\sum_{j=1}^{c} (\frac{d(x_k, \bar{v}_i)}{d(x_k, \bar{v}_j)})^{\frac{1}{m-1}}]^{-1}, \tag{2}$$

while the solution for $\bar{v}$ with a given $\bar{v}$ in step **CM3** is

$$\bar{v}_i = \frac{\sum_{k=1}^{n} (\bar{u}_{ik})^m x_k}{\sum_{k=1}^{n} (\bar{u}_{ik})^m} \tag{3}$$

in the case of the Euclidean space $d(x_k, v_i) = \|x_k - v_i\|_2^2$. Moreover we sometimes write d_{ik} instead of $d(x_k, v_i)$. In particular, the last symbol may imply variations of the dissimilarity $d(x_k, v_i)$.

As mentioned earlier, we regard binary solutions as extremal and hence singular, when the admissble set M_f of fuzzy memberships is assumed. Thus, the regularizing parameter is m, and the objective function is transformed by using this parameter which results in regularized solutions that are fuzzy. Notice that when we refer to regularizations, the regularized functional is similar to the original one and the regularized solution is *near* the original solution. The last statement is difficult to prove exactly, but we have many empirical evidence that fuzzy solutions approximate corresponding crisp solutions.

What is the advantage of taking this standpoint of the regularization? The answer is that this argument implies that other types of regularizations are possible. Indeed, a recent method by one of the authors is considered to be a typical regularization of the crisp c - means.

In order to see this, notice that a typical regularization for the optimization of J is done by adding regularizing functional K with a positive parameter α:

$$J_\alpha = J + \alpha K$$

and the minimization of J_α is considered instead of minimizing J. The regularizing parameter α should be adjusted so that good quality solutions are obtained. The maximum entropy method has the last form, whereas the standard fuzzy c - means does not. In this sense the standard fuzzy c - means is a *nonlinear* regularization.

2.2 Maximum entropy approach

The maximum entropy method [15, 16] has been introduced as the following optimization problem:

$$\begin{aligned} &\textit{maximize} \quad -\sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik} \log u_{ik} \\ &\textit{subject to} \quad \sum_{i=1}^{c} u_{ik} = 1, \quad J_1(U, v) = K \end{aligned}$$

where $J_1(U, v)$ is given by (1) and K be a parameter. Let the Lagrange multipliers be ν_k and λ, and the Lagrange function be $\mathcal{L}$. Then we have

$$\begin{aligned} \mathcal{L} \quad = \quad & -\sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik} \log u_{ik} \\ & +\nu_k (\sum_{i=1}^{c} u_{ik} - 1) + \lambda (J_1 - K). \end{aligned}$$

The stationary point of $\mathcal{L}$ is

$$u_{ik} = \frac{e^{-\lambda d_{ik}}}{\sum_{j=1}^{c} e^{-\lambda d_{jk}}}. \tag{4}$$

where $d_{ik} = d(x_k, v_i)$ for simplicity. The multiplier ν_k has already been determined while λ is still in (4). This parameter should be determined so that the constraint $J_1(U, v) = K$ is satisfied. The number K is, however, difficult to determine beforehand.

The last statement implies that the multiplier λ should not be treated as a uniquely determined number, but an adjustable parameter. Thus, actually we consider a regularized problem which is equivalent to the above:

$$\begin{aligned} &\textit{minimize} \quad J_1(U, v) + \lambda^{-1} \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik} \log u_{ik} \\ &\textit{subject to} \quad \sum_{i=1}^{c} u_{ik} = 1, \end{aligned}$$

where λ is a positive regularizing parameter. Notice that the number K can be neglected. It is easily seen that the solution of the last problem is given by the same formula of (4). Remark also that in the Euclidean space $d(x_k, v_i) = \|x_k - v_i\|_2^2$, the center $\bar{v} = (\bar{v}_1, ..., \bar{v}_c)$ in **CM** is given by

$$\bar{v}_i = \frac{\sum_{k=1}^{n} \bar{u}_{ik} x_k}{\sum_{k=1}^{n} \bar{u}_{ik}}. \tag{5}$$

When $\bar{u}_{ik}$ is replaced by $(\bar{u}_{ik})^m$, the above formula becomes that of calculating centers in the standard fuzzy c - means.

2.3 Logarithmic transformation of dissimilarity

A simple calculation shows a relationship between (2) and (4). Let $\bar{d}_{ik} = d(x_k, \bar{v}_i)$ and $\alpha = \frac{1}{m-1} > 0$ in (2). Since $a^z = e^{z \log a}$, we have

$$\begin{aligned} \bar{u}_{ik} &= [\sum_{j=1}^{c} (\frac{\bar{d}_{ik}}{\bar{d}_{jk}})^{\alpha}]^{-1} && (6) \\ &= [\sum_{j=1}^{c} e^{\alpha(\log \bar{d}_{ik} - \log \bar{d}_{jk})}]^{-1} && (7) \\ &= \frac{e^{-\alpha \log \bar{d}_{ik}}}{\sum_{j=1}^{c} e^{-\alpha \log \bar{d}_{jk}}}. && (8) \end{aligned}$$

The last expression is similar to (4). Indeed, the equation (4) coincides with (8) by the substitutions $\lambda = \alpha$ and $d_{jk} = \log \bar{d}_{jk}$.

Since the calculations of the centers are different between the both methods, we cannot say that the two methods are equivalent. Nevertheless, the above logarithmic transformation shows that the underlying idea is essentially the same.

3 Methods derived from maximum entropy approach

3.1 Methods of clustering

The maximum entropy approach is not a single method, but is providing a framework in which different algorithms are incorporated. Namely, the consideration in the previous section immediately leads us to a list of new methods of clustering. Some of them are as follows.

(I) As shown by Bezdek and other researchers, there are a number of variations such as the fuzzy c - varieties [2] and fuzzy c - regression [8, 21]. These models can be reconsidered by the maximum entropy approach.

(II) Recently, L_1 and L_p spaces are considered by a number of researchers [3, 9, 18, 19]. Their methods can also be studied within the present framework.

Let us examine these statements in more detail, by showing two examples: one in (I) and the other in (II). In the alternative optimization **CM**, u_{ik} is calculated with a given $\bar{v}$ in **CM2**, and v is calculated with a given $\bar{U}$ in **CM3**. This structure of the algorithm facilitates development of new methods in the maximum entropy framework.

Example 1 (FCV).

Let us take an example in the fuzzy c - varieties, in which the objective function is

$$J(U, w, s) = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m d_{ik}, \tag{9}$$

where

$$d_{ik} = \|x_k - w_i\|^2 - \langle x_k - w_i, s_i \rangle^2. \tag{10}$$

Namely, the cluster i is represented by a line $w_i + \beta s_i$ where β is a scalar variable, w_i means the center of the cluster, and s_i is the unit vector showing the direction of the line. The line shows the principal axis of the cluster. Thus, the dissimilarity is measured by the distance between an individual and the line.

Consequently, the alternative algorithm requires three steps:

Procedure FCV.

FCV1. Set initial values for $\bar{U}$, $\bar{w}$, $\bar{s}$.

FCV2. Minimize $J(\bar{U}, w, \bar{s})$ with respect to w. Let the optimal solution be $\bar{w}$.

FCV3. Minimize $J(\bar{U}, \bar{w}, s)$ with respect to s. Let the optimal solution be $\bar{s}$.

FCV4. Minimize $J(U, \bar{w}, \bar{s})$ with respect to U. Let the optimal solution be $\bar{U}$.

FCV5. Check stopping criterion. If the criterion is not satisfied, go to FCV2.

It is known that in FCV2,

$$\bar{w}_i = \frac{\sum_{k=1}^{n} (\bar{u}_{ik})^m x_k}{\sum_{k=1}^{n} (\bar{u}_{ik})^m}.$$

In FCV3, let $\bar{s}_i$ be the normalized eigenvector for the maximum eigenvalue of the following matrix

$$A^i = \sum_{\ell=1}^{c} (\bar{u}_{i\ell})^m (x_\ell - \bar{w}_i)(x_\ell - \bar{w}_i)^T.$$

In FCV4,

$$\bar{u}_{ik} = [\sum_{j=1}^{c} (\frac{\bar{d}_{ik}}{\bar{d}_{jk}})^{\frac{1}{m-1}}]^{-1}.$$

as usual. (For simplicity, we write $\bar{d}_{ik} = \|x_k - \bar{w}_i\|^2 - \langle x_k - \bar{w}_i, \bar{s}_i \rangle^2$.)

Now, it is easy to obtain the maximum entropy method in which the objective function L is given by

$$L = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik} d_{ik} + \lambda^{-1} \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik} \log u_{ik}.$$

The solution in FCV2 is

$$\bar{w}_i = \frac{\sum_{k=1}^{n} \bar{u}_{ik} x_k}{\sum_{k=1}^{n} \bar{u}_{ik}}.$$

In FCV3, let $\bar{s}_i$ be the normalized eigenvector for the maximum eigenvalue of the matrix

$$B^i = \sum_{\ell=1}^{c} \bar{u}_{i\ell} (x_\ell - \bar{w}_i)(x_\ell - \bar{w}_i)^T.$$

Finally, in FCV4,

$$\bar{u}_{ik} = \frac{e^{-\lambda \bar{d}_{ik}}}{\sum_{j=1}^{c} e^{-\lambda \bar{d}_{jk}}}.$$

which is the same formula as (4).

Example 2 (L_1 FCM).

As another example, let us consider the L_1 space based fuzzy c - means [3, 9, 18]. Namely we assume that

$$d(x_k, v_i) = \|x_k - v_i\|_1 = \sum_{\ell=1}^{p} |x_k^\ell - v_i^\ell|, \tag{11}$$

where x_k^ℓ is the ℓth component of the vector x_k. Thus, the dissimilarity is given by the L_1 norm.

Miyamoto and Agusta [18] show a fast algorithm in the step **CM3**. (Remark that u_{ik} is calculated by the same formula (2).) As in the Euclidean case, the solution $\bar{v}_i$ is calculated componentwise. Hence we consider ℓth component alone. Assume that when $\{x_1^\ell, ..., x_n^\ell\}$ is ordered, subscripts are changed using a permutation function $q(k)$, $k = 1, ..., n$, that is, $x_{q(1)}^\ell \leq x_{q(2)}^\ell \leq ... \leq x_{q(n)}^\ell$. Using $\{x_{q(k)}^\ell\}$, the following algorithm calculates $\bar{v}_i^\ell$, in which we put $\gamma_{ik} = (u_{ik})^m$.

Algorithm CC.

begin
  $S := -\sum_{k=1}^{n} \gamma_{ik}$;
  $r := 0$;
  while ($S < 0$)do begin
    $r := r + 1$;
    $S := S + 2\gamma_{iq(r)}$
  end;
  output $v_i^\ell = x_{q(r)}^\ell$ as the ℓth coordinate of
  the cluster center $\bar{v}_i$
end.

Now, we can develop a maximum entropy algorithm in which the objective function is

$$J_1(U, v) + \lambda^{-1} \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik} \log u_{ik},$$

where $J_1(U, v)$ is defined by (1) and the dissimilarity given by (11). By a discussion parallel to the previous work [18], the center is calculated by the algorithm CC *with the replacement* $\gamma_{ik} = u_{ik}$. Notice that the membership u_{ik} is calculated by (4). It is obvious that the latter method has the advantage of less calculation, since calculation of the power m in $(u_{ik})^m$ is necessary in the former method.

These considerations show that a variety of other methods can easily be transformed into the corresponding algorithms in the framework of the maximum entropy.

3.2 Fuzzy classifiers

Clustering is frequently called unsupervised classification, whereas supervised classification means determination of classification rules given a number of classes of individuals. Frequently the latter is simply referred to as classification problem. There are various methods for this class of problems, which include classical descriminant analysis [11] as well as the nonparametric technique of the k - nearest neighbor method (kNN).

Fuzzy supervised classification techniques have been studied as extensions of crisp nonparametric methods. In particular, a fuzzy kNN and a method of fuzzy nearest prototype has been studied by Keller *et al.* [13].

Their fuzzy kNN classifier, which assigns the membership of x in the class i, is given by

$$u_j(x) = \sum_{i=1}^{k} u_{ji} [\sum_{\ell=1}^{k} (\frac{d(x,y_i)}{d(x,y_\ell)})^{\frac{1}{m-1}}]^{-1} \tag{12}$$

where $y_1, ..., y_k$ are k nearest individuals from x, and $d(x, y_i) = \|x - y_i\|_2^2$.

Moreover, the fuzzy nearest prototype classifier therein is

$$u_j(x) = [\sum_{\ell=1}^{c} (\frac{d(x,z_j)}{d(x,z_\ell)})^{\frac{1}{m-1}}]^{-1} \tag{13}$$

where z_i $(i = 1, ..., c)$ are prototype vectors for the class i.

It is immediate to see that these classifiers are closely related to the rule (2) of determining memberships in the standard fuzzy c - means.

Now, we obtain, by analogy, fuzzy classifiers from the maximum entropy principle:

(I) Fuzzy kNN classifier:

$$u_j(x) = \frac{\sum_{i=1}^{k} u_{ji} e^{-\lambda d(x,y_i)}}{\sum_{\ell=1}^{k} e^{-\lambda d(x,y_\ell)}}. \tag{14}$$

($y_1, ..., y_k$ are k nearest individuals from x.)

(II) Fuzzy prototype classifier:

$$u_j(x) = \frac{e^{-\lambda d(x,z_j)}}{\sum_{\ell=1}^{c} e^{-\lambda d(x,z_\ell)}}. \tag{15}$$

(z_i, $i = 1, ..., c$, are prototype vectors for the class i.)

An advantage of the present classifiers is that the combination of exponential functions provides a clear view for behaviors of the function. For example, the prototype classifier satisfies

$$u_j(x)^{-1} = 1 + \sum_{\ell \neq j} e^{-\lambda(d(x,z_\ell) - d(x,z_j))}. \tag{16}$$

When λ is large enough, the fuzzy classifier is shown to approach the crisp nearest prototype classifier by which x is classified into the class of the nearest prototype. To see this, let z_j be the nearest prototype to x, then the terms $d(x, z_\ell) - d(x, z_j)$ are all positive. When $\lambda \to 0$, the second term in (16) becomes sufficiently small and hence $u_j(x) \to 1$. From $\sum_\ell u_\ell(x) = 1$, we have $u_\ell(x) \to 0$ $(\ell \neq j)$. Thus, the classifier tends to crisp one. It should be remarked that this property is obtained from the assumption that the parameter $\lambda > 0$ is *adjustable*.

Another observation is that the above classifiers are not limited in the Euclidean space, since $d(x, v)$ may be defined in terms of other spaces such as the L_1 space. Moreover, we can derive fuzzy classifiers based on fuzzy c - varieties by using

$$u_j(x) = \frac{e^{-\lambda d(x,V_j)}}{\sum_{\ell=1}^{c} e^{-\lambda d(x,V_\ell)}}, \tag{17}$$

where V_j, $j = 1, ..., c$, is not a single vector, but a set of parameters characterizing a variety. In relation to the FCV clustering shown as Example 1, $d(x, V_j)$ is the distance from x to the line representing the class.

4 Methods related to regularization

Regularization is somewhat similar to the penalty method for optimization by considering the regularizing functional to be a penalty. Another relation between regularization and bicriteria optimization is observed. Namely, the regularized objective function $J_\alpha = J + \alpha K$ keeps balance between the original objective J and the regularizing functional αK.

The latter idea of the bicriteria optimization has been used in fuzzy clustering [24]. Combining their method and the consideration herein, we have a new algorithm using the maximum entropy. The detailed description is omitted here.

Another example of bicriteria fuzzy clustering is the possibilistic approach by Krishnapuram and Keller [14]. They remove the assumption of the fuzzy partition $\sum_{i=1}^{c} u_{ik} = 1$ and consider an objective function

$$J = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m d_{ik} + \sum_{i=1}^{c} \eta_i \sum_{k=1}^{n} (1 - u_{ik})^m.$$

Is should be noticed that the second term is introduced in order to *regularize* the objective function: without this term the solution becomes trivial. At the same time the first and second terms should be balanced in the optimization, and hence the method is regarded as a bicriteria clustering algorithm.

An interesting exercise is to design a regularized objective function using the entropy that leads us to a very simple membership allocation scheme. See Appendix.

5 Conclusions

We have presented a number of methods that are parallel to existing methods within the framework of maximum entropy. Thus, the major purpose of the present paper is to show potential of this framework or approach. Therefore examination of performances of the individual methods presented here by numerical simulations and real data will be done as further works. In this sense, the present paper is still a preliminary

report, and more studies should be done using this approach of the maximum entropy.

The crucial point in the present approach is that the parameter λ is rather free to choose, whereas $m = 2$ has frequently been used in the standard FCM, since this value simplifies the calculation. From the viewpoint of regularization, a regularizing parameter should not uniquely be determined. For example, properties of a regularized method can be made clear by moving the regularizing parameter, as shown in Section 3.2.

For some readers it seems strange to use the maximum entropy for clustering, since the entropy criterion has sometimes been referred to in validation of clusters [2]. Actually we abandan the use of entropy in validating clusters by employing the entropy criterion in cluster generation. Thus, the relationship between the criterion for clustering and that for validation is delicate in general.

As another direction for future studies, mixture with other methods and the approach here is promising. There have been studies of clustering using Kohonen's associative memory [12]. Pal, Bezdek, and Tsao [22] develop a crisp clustering algorithm GLVQ based on the learning vector quantization. Fuzzy version of this algorithm using regularization should be attempted.

Since the membership allocation rule obtained by the maximum entropy has a simple form of exponential functions, there are rooms for observing many theoretical properties. Thus, the present approach provides a new viewpoint for fuzzy clustering which has many possibilities for future studies.

Acknowledgment

This research has partly been supported by TARA (Tsukuba Advanced Research Alliance), the University of Tsukuba.

Appendix

Let us consider an objective function that leads us to a simple membership allocation scheme of an exponential type. Put

$$L = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik} d_{ik} + \sum_{i=1}^{c} \zeta_i^{-1} \sum_{k=1}^{n} u_{ik} (\log u_{ik} - 1),$$

in which $\sum_{i=1}^{c} u_{ik} = 1$ is not assumed. From

$$\frac{\partial L}{\partial u_{ik}} = d_{ik} + \zeta_i^{-1} \log u_{ik} = 0$$

we have

$$u_{ik} = e^{-\zeta_i d_{ik}}.$$

References

[1] M. R. Anderberg, *Cluster Analysis for Applications*, Academic Press, New York, 1973.

[2] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum, New York, 1981.

[3] L. Bobrowski and J. C. Bezdek, c-means clustering with the ℓ_1 and ℓ_∞ norms, *IEEE Trans. on Systems, Man, and Cybern.*, Vol. 21, pp. 545–554, 1991.

[4] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.

[5] J. C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, *J. of Cybernetics*, Vol.3, pp. 32–57, 1974.

[6] J. C. Dunn, Well-separated clusters and optimal fuzzy partitions, *J. of Cybernetics*, Vol.4, pp. 95–104, 1974.

[7] E. W. Forgy, Cluster analysis of multivariate data: efficiency vs. interpretability of classifications, *Biometrics*, Vol.21, pp. 768–769 (Abstract) 1965.

[8] R. J. Hathaway and J. C. Bezdek, Switching regression models and fuzzy clustering, *IEEE Trans. on Fuzzy Syst.*, Vol.1, pp. 195–204, 1993.

[9] K. Jajuga, L_1-norm based fuzzy clustering, *Fuzzy Sets and Systems*, Vol. 39, pp. 43–50, 1991.

[10] M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics, Vol. 2, 4th Ed.*, Griffin, London, 1979.

[11] M. G. Kendall, A. Stuart, and J. K. Ord, *The Advanced Theory of Statistics, Vol. 3, 4th Ed.*, Griffin, London, 1983.

[12] T. Kohonen, *Self-Organization and Associative Memory*, Springer-Verlag, Heiderberg, 1989.

[13] J. M. Keller, M. R. Gray, and J. A. Givens, Jr., A fuzzy k - nearest neighbor algorithm, *IEEE Trans., on Syst., Man, and Cybern.*, Vol. 15, pp. 580–585, 1985.

[14] R. Krishnapuram and J. M. Keller, A possibilistic approach to clustering, *IEEE Trans. on Fuzzy Syst.*, Vol.1, pp. 98–110, 1993.

[15] R.-P. Li and M. Mukaidono, Gaussian clustering and its application to rock classification, *Proc. of Eleventh Fuzzy System Symposium*, Japan Society of Fuzzy Theory and Systems, pp .697–698, 1995.

[16] R.-P. Li and M. Mukaidono, A maximum entropy approach to fuzzy clustering, *Proc. of the 4th IEEE Intern. Conf. on Fuzzy Systems (FUZZ-IEEE/IFES'95)*, Yokohama, Japan, March 20-24, 1995, pp. 2227–2232.

[17] J. B. McQueen, Some methods of classification and analysis of multivariate observations, *Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.*, pp. 281–297, 1967.

[18] S. Miyamoto and Yudi Agusta, An efficient algorithm for ℓ_1 fuzzy c-means and its termination, *Control and Cybernetics* Vol. 24, pp. 421–436, 1995.

[19] S. Miyamoto and Yudi Agusta, Efficient algorithms for ℓ_p fuzzy c-means clustering and their termination properties. *Fifth Conference of the International Federation of Classification Societies (IFCS 96)*, Kobe, Japan, March 27-30, Vol.2, pp. 255–258, 1996.

[20] S. Miyamoto, Fuzzy c - means and its variations, *Journal of the Japan Society of Fuzzy Theory and Systems*, Vol.8, pp. 423–430, 1996 (in Japanese).

[21] Y. Nakamori, Fuzzy clustering and regression analysis, *Journal of the Japan Society of Fuzzy Theory and Systems*, Vol.8, pp. 431–439, 1996 (in Japanese).

[22] N. K. Pal, J. C. Bezdek, and E. C.-K. Tsao, Generalized clustering networks and Kohonen's self-organizing scheme, *IEEE Trans. on Neural Networks*, Vol.4, pp. 549–557, 1993.

[23] A. N. Tihonov, Solutions of incorrectly formulated problems and the regularization method, *Dokl. Akad. Nauk. SSSR*, Vol.151, pp. 1035–1038, 1963.

[24] H.-F. Wang and G.-Y. Wu, Bi-criteria fuzzy clustering systems, *IV IFSA World Congress*, São Paulo, Brazil, 1995, Vol.1, pp. 633–636, 1995.

METODE CLUSTERING DENGAN ALGORITMA FUZZY C-MEANS UNTUK REKOMENDASI PEMILIHAN BIDANG KEAHLIAN PADA PROGRAM STUDI TEKNIK INFORMATIKA

Muhammad Faisal Mirza A11.2009.04930
Program Studi Teknik Informatika –S1
Fakultas Ilmu Komputer
Universitas Dian Nuswantoro, Jl. Nakula 1 No. 5-11. Semarang
Faisal4asi@gmail.com

ABSTRAK

Fakultas ilmu komputer merupakan salah satu fakultas favorite di Universitas Dian Nuswantoro terutama pada jurusan Teknik Informatika. Jurusan Teknik Informatika sendiri rencananya akan dibagi menjadi 3 bidang keahlian yaitu Keahlian *Informatic*, *Software develop and enginer*, dan *Networking*. Dalam pembagian bidang keahlian tersebut banyak mahasiswa akan mengalami kesulitan untuk menentukan keahlian mana yang akan diambilnya. Maka dari itu butuh rekomendasi untuk mereka.*Fuzzy C-means* adalah suatu teknik pengklasteran data yang mana keberadaan pada setiap titik data dalam suatu klaster ditentukan oleh derajat keanggotaan. Beberapa penelitian telah menghasilkan kesimpulan bahwa algoritma Fuzzy C-Means dapat dipergunakan untuk mengelompokkan data berdasarkan atribut-atribut tertentu. Pada penelitian ini akan digunakan algoritma Fuzzy C-Means untuk mengelompokkan mahasiswa berdasarkan transkip nilai mata kuliah prasayarat untuk rekomendasi penjurusan. tujuan dari penelitian ini adalah menerapkan metode *clustering* dengan Algoritma *Fuzzy C-Means* dalam kasus pengelompokkan mahasiswa berdasarkan transkip nilai mata kuliah prasyarat untuk rekomendasi pemilihan bidang keahlian.

Kata kunci : clustering, rekomendasi penjurusan mahasiswa

BAB I
PENDAHULUAN

1.1. Latar Belakang

Fakultas ilmu komputer merupakan salah satu fakultas favorite di Universitas Dian Nuswantoro terutama pada jurusan Teknik Informatika. Jurusan Teknik Informatika sendiri rencananya akan dibagi menjadi 3 bidang keahlian yaitu Keahlian *Informatic*, *Software develop and enginer*, dan *Networking*. Dalam pembagian bidang keahlian tersebut banyak mahasiswa akan mengalami kesulitan untuk menentukan keahlian mana yang akan diambilnya. Maka dari itu butuh rekomendasi untuk mereka, Untuk merekomendasikan mahasiswa yang begitu banyak tentu tidak mudah karena kita harus mengelola data yang begitu besar yang memiliki jumlah field dan jumlah record yang begitu banyak. Teknologi *data mining* merupakan salah satu alat bantu untuk penggalian data pada basis data berukuran besar dan dengan spesifikasi tingkat kerumitan yang telah banyak digunakan pada banyak domain aplikasi seperti perbankan maupun bidang telekomunikasi [1]. oleh karena itu teknologi *data mining* dapat dimanfaatkan untuk melakukan pengelompokkan mahasiswa berdasarkan nilai mata kuliah yang terkait dengan bidang keahlian yang ada sehingga

dapat memberikan gambaran kepada mahasiswa untuk memilih program keahlian yang sesuai dengan kemampuan mereka.

Clustering merupakan salah satu metode data mining. *Clustering* berguna untuk mengelompokkan data (objek) yang didasarkan pada informasi yang ditemukan dalam data yang menggambarkan objek tersebut dan hubungan diantaranya [2].pada penelitian yang dilakukan oleh jimmy, Sherwin [4] tentang segmentasi citra spot dengan menggunakan pendekatan *Fuzzy C-means* menyimpulkan bahwa metode logika samar *(Fuzzy C-Means Clustering)* memiliki tingkat kestabilan ouput / hasil yang lebih baik daripada pendekatan metode konvensional *(K-Means Clustering)*. sedangkan pada penelitian yang dilakukan oleh Budi, Rizal [7] tentang pembagian kelas kuliah mahasiswa menggunakan algoritma pengklasteran Fuzzy C-Means menyimpulkan bahwa Fuzzy C-means cocok untuk alokasi mahasiswa menjadi beberapa kelas. Oleh karena itu dalam penelitian ini akan menggunakan algoritma *Fuzzy C-means*. *Fuzzy C-means Clustering* adalah salah satu algoritma *clustering* yang merupakan bagian dari metode *Hard K-Means*. *Fuzzy C-means* menggunakan model pengelompokan *Fuzzy* sehingga data dapat menjadi anggota dari semua kelas atau cluster terbentuk dengan derajat atau tingkat keanggotaan yang berbeda antara 0 hingga 1 [5].

1.2. Rumusan Masalah

Berdasarkan **uraian yang kami kemukakan di atas maka dalam penelitian ini dapat diambil rumusan masalah sebagai berikut "**bagaimana menerapkan metode *Clustering* dengan alogritam *Fuzzy C-means* dalam kasus pengelompokkan mahasiswa berdasarkan transkip nilai mata kuliah prasyarat sebagai rekomendasi untuk mengambil bidang keahlian yang sesuai dengan kemampuan mahasiswa**.**

1.3. Batasan Masalah

Batasan masalah untuk penelitian ini meliputi :

a. Penelitian ini diberlakukan untuk program studi Teknik Informatika S1 Fakultas Ilmu Komputer Universitas Dian Nuswantoro.
b. Penelitian ini menggunakan data mahasiswa angkatan 2009.
c. Penelitian ini menggunakan transkip nilai mata kuliah prasyarat.
d. Penelitian ini diaplikasikan menggunakan matlab 7.10.

1.4. Tujuan Penelitian

Berdasarkan rumusan masalah diatas maka tujuan penelitian ini adalah menerapkan metode *clustering* dengan Algoritma *Fuzzy C-Means* dalam kasus pengelompokkan mahasiswa berdasarkan

transkip nilai mata kuliah prasyarat untuk rekomendasi pemilihan bidang keahlian.

1.5. Manfaat Penelitian

Manfaat dari penelitian ini adalah :

a. Bagi Penulis, penelitian ini berguna untuk menambah wawasan mengenai *data mining* metode *Clustering* dengan algoritma *Fuzzy C-means* beserta penerapannya.
b. Bagi Program Studi Teknik Informatika, penelitian ini merupakan salah satu upaya untuk membantu mahasiswanya dalam memilih bidang keahlian.
c. Hasil penelitian ini juga dapat dimanfaatkan sebagai referensi untuk penelitian selanjutnya.

BAB II
LANDASAN TEORI DAN TINJAUAN PUSTAKA

2.1 Tinjauan Pustaka

Penelitian mengenai pengelompokkan data untuk mendapatkan informasi tentang rekomendasi penjurusan bukanlah baru pertama kali ini dilakukan , sudah ada penelitian terdahulu tentang penerapan metode *clustering* tersebut. Penelitian terdahulu yang relevan dengan penelitian ini adalah sebagai berikut.

wijaya [3] dengan judul "Analisis Algoritma *K-means* untuk Sistem Pendukung Keputusan Penjurusan Siswa di MAN Binong Subang". Sistem tersebut memberikan gambaran untuk penjurusan berdasarkan nilai siswa semester lalu.

jimmy, Sherwin [4] dengan judul "Segmentasi Citra Spot dengan Menggunakan Pendekatan *Fuzzy C-means*". Sistem tersebut memberikan perbandingan segmentasi citra spot antara Fuzzy C-means dengan K-means.

Budi, rizal [7] dengan judul "Pembagian Kelas Kuliah Mahasiswa Menggunakan Algoritma Pengklasteran *Fuzzy"* sistem tersebut bertujuan untuk membuat pembagian kelas mahasiswa berdasarkan nilai prestasi pada mata kuliah yang menjadi prasyarat untuk menempuh mata kuliah yang baru.

2.2 Landasan teori

2.2.1 Data mining

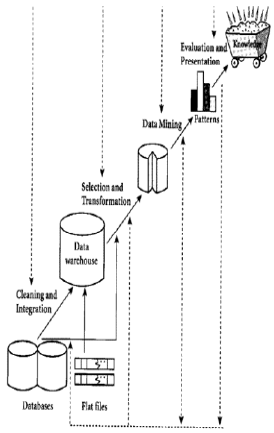
Secara sederhana *data mining* adalah penambangan atau penemuan informasi baru dengan mencari pola atau aturan tertentu dari sejumlah data yang sangat besar[9]. *Data mining*, sering juga disebut sebagai *knowledge discovery in database* (KDD). KDD adalah kegiatan yang meliputi pengumpulan, pemakaian data, historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar [10]. *Data mining* adalah kegiatan menemukan pola yang menarik dari data dalam jumlah besar, data dapat disimpan dalam *database, data warehouse*, atau penyimpanan informasi lainnya. *Data mining* berkaitan dengan bidang ilmu – ilmu lain, seperti *database system*, *data*

warehousing, statistik, *machine learning*, *information retrieval*, dan komputasi tingkat tinggi. Selain itu, *data mining* didukung oleh ilmu lain seperti *neural netwo*rk, pengenalan pola, *spatial data analysis*, *image database*, *signal processing* [10]. *Data mining* didefinisikan sebagai proses menemukan pola-pola dalam data. Proses ini otomatis atau seringnya semiotomatis. Pola yang ditemukan harus penuh arti dan pola tersebut memberikan keuntungan, biasanya keuntungan secara ekonomi. Data yang dibutuhkan dalam jumlah besar[12]. Karakteristik *data mining* sebagai berikut :

a. *Data mining* berhubungan dengan penemuan sesuatu yang tersembunyi dan pola data tertentu yang tidak diketahui sebelumnya.
b. *Data mining* biasa menggunakan data yang sangat besar. Biasanya data yang besar digunakan untuk membuat hasil lebih dipercaya.
c. *Data mining* berguna untuk membuat keputusan yang kritis, terutama dalam strategi [9].

Berdasarkan beberapa pengertian yang telah disebutkan diatas, maka dapat ditarik kesimpulan bahwa *data mining* adalah suatu teknik untuk menggali informasi yang tersembunyi pada gunungan data. Kata mining sendiri berarti usaha untuk mendapatkan sedikit barang berharga dari sejumlah besar material dasar.

Karena data mining adalah suatu rangkaian proses, data mining dapat dibagi menjadi beberapa tahap. Tahap-tahap tersebut bersifat *interaktif* di mana pemakai terlibat langsung atau dengan perantaraan *knowledge base*. Tahap-tahap ini diilustrasikan di Gambar 2.1:



Gambar 2.1 tahap-tahap data mining

Tahap-tahap data mining ada 6 yaitu :

a. Pembersihan data (data cleaning)

Pembersihan data merupakan proses menghilangkan noise dan data yang tidak konsisten atau data tidak relevan. Pada umumnya data yang diperoleh, baik dari database suatu perusahaan maupun hasil eksperimen, memiliki isian-isian yang tidak sempurna seperti data yang hilang, data yang tidak valid atau juga hanya sekedar salah

ketik. Selain itu, ada juga atribut-atribut data yang tidak relevan dengan hipotesa data mining yang dimiliki. Data-data yang tidak relevan itu juga lebih baik dibuang. Pembersihan data juga akan mempengaruhi performasi dari teknik data mining karena data yang ditangani akan berkurang jumlah dan kompleksitasnya.

b. Integrasi data (data integration)

Integrasi data merupakan penggabungan data dari berbagai database ke dalam satu database baru. Tidak jarang data yang diperlukan untuk data mining tidak hanya berasal dari satu database tetapi juga berasal dari beberapa database atau file teks. Integrasi data dilakukan pada atribut-aribut yang mengidentifikasikan entitas-entitas yang unik seperti atribut nama, jenis produk, nomor pelanggan dan lainnya. Integrasi data perlu dilakukan secara cermat karena kesalahan pada integrasi data bisa menghasilkan hasil yang menyimpang dan bahkan menyesatkan pengambilan aksi nantinya. Sebagai contoh bila integrasi data berdasarkan jenis produk ternyata menggabungkan produk dari kategori yang berbeda maka akan didapatkan korelasi antar produk yang sebenarnya tidak ada.

c. Seleksi Data (Data Selection)

Data yang ada pada database sering kali tidak semuanya dipakai, oleh karena itu hanya data yang sesuai untuk dianalisis yang akan diambil dari database. Sebagai contoh, sebuah kasus yang meneliti faktor kecenderungan orang membeli dalam kasus market basket analysis, tidak perlu mengambil nama pelanggan, cukup dengan id pelanggan saja.

d. Transformasi data (Data Transformation)

Data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam data mining. Beberapa metode data mining membutuhkan format data yang khusus sebelum bisa diaplikasikan. Sebagai contoh beberapa metode standar seperti analisis asosiasi dan clustering hanya bisa menerima input data kategorikal. Karenanya data berupa angka numerik yang berlanjut perlu dibagi-bagi menjadi beberapa interval.

Proses ini sering disebut transformasi data.

e. Proses mining

Merupakan suatu proses utama saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data.

f. Evaluasi pola (pattern evaluation)

Untuk mengidentifikasi pola-pola menarik kedalam knowledge based yang ditemukan. Dalam tahap ini hasil dari teknik data mining berupa pola-pola yang khas maupun model prediksi dievaluasi untuk menilai apakah hipotesa yang ada memang tercapai. Bila ternyata hasil yang diperoleh tidak sesuai hipotesa ada beberapa alternatif yang dapat diambil seperti menjadikannya umpan balik untuk memperbaiki proses data mining, mencoba metode data mining lain yang lebih sesuai, atau menerima hasil ini sebagai suatu hasil yang di luar dugaan yang mungkin bermanfaat.

g. Presentasi pengetahuan (knowledge presentation)

Merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna. Tahap terakhir dari proses data mining adalah bagaimana memformulasikan keputusan atau aksi dari hasil analisis yang didapat. Ada kalanya hal ini harus melibatkan orang-orang yang tidak memahami data mining. Karenanya presentasi hasil data mining dalam bentuk pengetahuan yang bisa dipahami semua orang adalah satu tahapan yang diperlukan dalam proses data mining. Dalam presentasi ini, visualisasi juga bisa membantu mengkomunikasikan hasil data mining [11].

2.2.2 Clustering

Clustering adalah pekerjaan mengelompokkan data (objek) yang didasarkan hanya pada informasi yang ditemukan dalam data yang menggambarkan objek tersebut dan hubungan diantaranya [11]. Tujuannya adalah agar objek-objek yang bergabung dalam sebuah kelompok merupakan objek-objek yang mirip (atau berhubungan) satu sama lain dan berbeda (atau tdak berhubungan)

dengan objek dalam kelompok yang lain. Lebih besar kemiripanya (homogenitas) dalam kelompok dan lebih besar perbedaanya dianara kelompok yang lain [13]. Menurut keanggotaan dalam kelompok, pengelompokkan dapat dibagi menjadi dua, yaitu eksklusif dan tumpang tindih. Dalam kategori eksklusif sebuah data bisa dipastikan hanya menjadi anggota satu kelompok dan tidak menjadi anggota kelompok lain. Metode pengelompokkan yang masuk dalam kategori ini adalah K-means dan DBSCAN, sedangkan yang masuk kategori tumpang tindih adalah metode yang membolehkan sebuah data menjadi anggota di lebih dari satu kelompok yaitu Fuzzy C-means, pengelompokkan hierarki [13].

2.2.3 Fuzzy C-means

Fuzzy C-means adalah suatu teknik pengklasteran data yang mana keberadaan pada setiap titik data dalam suatu klaster ditentukan oleh derajat keanggotaan. Konsep dasar Fuzzy C-means, yang pertama adalah menentukan pusat klaster, yang berfungsi untuk menandai lokasi rata-rata untuk tiap-tiap klaster. Pada kondisi awal, pusat klaster ini masih belum akurat. Tiap-tiap titik data memiliki derajat keanggotaan untuk tiap-tiap klaster. Dengan cara memperbaiki pusat klaster dan derajat keanggotaan tiap-tiap titik data secara berulang maka dapat dilihat bahwa pusat klaster akan bergerak menuju lokasi yang tepat. Perulangan ini didasarkan pada minimisasi fungsi objektif yang menggambarkan jarak dari titik data yang diberikan ke pusat klaster yang terbobot oleh derajat keanggotaan titik data tersebut. Keluaran dari Fuzzy C-means bukan merupakan sistem inferensi kabur, namun merupakan deretan pusat klaster dan beberapa derajat keanggotaan untuk tiap tiap titik data. Algoritma Fuzzy C-means sebagai berikut.

a) Masukan data yang akan diklaster U , berupa matriks berukuran $n \times m$ (n = jumlah sampel data, m = atribut setiap data). ik U = data sampel ke-i (i =1,2,...,n), atribut ke- k(k =1,2,...,m).

b) Menetapkan nilai pangkat w >1 (misal: w =2), Eps (galat terkecil) (misal: 10-5), MaxIter (misal:100), jumlah klaster c > 1, dan t = 0 ; Menetapkan fungsi objektif awal: P (c) t secara acak;

c) Menetapkan matriks partisi (c) f μ awal sembarang, sebagai berikut.

$$\mu_f(C) = \begin{bmatrix} \mu_{11}[u_1] & \mu_{21}[u_1] & . & . & . & \mu_{C1}[u_1] \\ \mu_{12}[u_2] & \mu_{22}[u_2] & . & . & . & \mu_{C2}[u_2] \\ . & . & & . & & . \\ . & . & & . & & . \\ . & . & & . & & . \\ \mu_{1N}[u_N] & \mu_{2N}[u_N] & . & . & . & \mu_{CN}[u_N] \end{bmatrix} \quad (1)$$

d) Menaikkan nomor iterasi: t = t +1.

e) Menghitung pusat vektor tiap-tiap klaster untuk matrik partisi tersebut sebagai berikut.

$$\boldsymbol{v_{ij}} = \frac{\sum_{K=1}^{N}(\mu_{ik})^{w}\boldsymbol{x_{kj}}}{\sum_{K=1}^{N}(\mu_{ik})^{w}}$$

(2)

Memodifikasi tiap-tiap nilai keanggotaan sebagai berikut.

- Jika k fi y ≠ v ,

$$\boldsymbol{u_{ik}(y_k)} = \left[\sum_{g=1}^{c}\left(\frac{|\boldsymbol{u_k} - \boldsymbol{v_{ij}}|^2}{|\boldsymbol{u_k} - \boldsymbol{v_{gi}}|^2}\right)^{1/(w-1)}\right]^{-1}$$

(3)

- Jika $y_k = v_{fi}$,

  $\mu_{ik}(y_k) = 1$, jika $i = g$;

  $\mu_{ik}(y_k) = 0$, jika $i \neq g$;

(4)

f) Menghitung fungsi objektif:

$$\boldsymbol{P_t(C)} = \sum_{k=0}^{N}\sum_{i=1}^{c}(\boldsymbol{\mu_{ik}})^{w}\,|\boldsymbol{u_{ik}} - \boldsymbol{v_{ij}}|^2$$

(5)

g) Memodifikasi matriks partisi sebagai berikut:

No	Judul	Pengarang	jenis
1	Data Mining Konsep dan Aplikasi Menggunakan Matlab	Eko Prasetyo	Buku
2	Analisis Algortima K-Means Untuk Sistem Pendukung Keputusan Penjurusan Siswa Di Man Binong Subang	Arim wijaya	jurnal
3	Pembagian Kelas Kuliah Mahasiswa Menggunakan Algoritma Pengklasteran Fuzzy	Helmy Yulianto Hadi, R. Rizal Isnanto, Budi Setiyono	Jurnal

$$\mu_f(C) = \begin{bmatrix} \mu_{11}[u_1] & \mu_{21}[u_1] & . & . & . & \mu_{C1}[u_1] \\ \mu_{12}[u_2] & \mu_{22}[u_2] & . & . & . & \mu_{C2}[u_2] \\ . & . & & . & & . \\ . & . & & . & & . \\ . & . & & . & & . \\ \mu_{1N}[u_N] & \mu_{2N}[u_N] & . & . & . & \mu_{CN}[u_N] \end{bmatrix}$$

(6)

h) Mengecek kondisi untuk berhenti, yaitu:

$(|P_t(C) - P_{t-1}(C)| < Eps\,)atau(t > MaxIter)$

Jika ya berhenti, dan jika tidak ulangi kembali ke langkah-5 [7].

BAB III
METODE PENELITIAN

3.1 Jenis Penelitian

Penelitian yang dilaksanakan adalah jenis penelitian *eksperimen*, yaitu Mencoba untuk melakukan implementasi algoritma *Fuzzy C-Means* dalam pengelompokkan berdasarkan transkrip nilai mata kuliah prasyarat yang berhubungan dengan bidang keahlian.

3.2 Metode Pengumpulan Data

Untuk membuat algoritma *Fuzzy C-means*, beberapa teknik pengumpulan data dan variable digunakan, diantaranya adalah :

a. Studi kepustakaan dan literatur, digunakan untuk mendapatkan data awal tentang dasar *Fuzzy C-means*. Contoh studi kepustakaan dan literature yang saya gunakan adalah :

Tabel 3.1 Literatur

b. Wawancara mendalam, yaitu dilakukan untuk memperoleh data tentang pembagian bidang keahlian di program studi Teknik Informatika Fakultas Ilmu Komputer Universitas Dian Nuswantoro. Wawancara ini dilakukan dengan kaprogdi Teknik Informatika. Data wawancara tersebut dapat dilihat dibawah ini :

Tabel 3.2 Wawancara

No	Pertanyaan	Jawaban
1	Jurusan apa saja yang akan ada di Jurusan teknik informatika Fakultas ilmu komputer di udinus..?	-Informatik -software develop and enginer -Network
2	Akan diterapkan disemester berapa penjurusan tersebut..?	Akan diterapkan disemester 5
3	Bagaimana konsep penjurusan tersebut..?	Saat ini masih peminatan.., belum ada rekomendasi penjurusan.
4	Kenapa belum di terapkan yang rekomendasi penjurusan	Untuk saat ini konsepnya belum siap.

3.3 Sumber Data

a. Jenis data yang digunakan dalam penelitian ini berupa data primer dan data sekunder: Data Primer: berupa data tentang mekanisme pelaksanaan pemilihan bidang keahlian yang diperoleh dengan wawancara kepada kepala program studi Teknik Informatika Fakultas Ilmu Komputer Universitas Dian Nuswantoro.

Tabel 3.3 Mekanisme Pemlihan Bidang

No	Pertanyaan	Jawaban
1	Akan diterapkan disemester berapa penjurusan tersebut..?	Akan diterapkan disemester 5
2	Bagaimana konsep penjurusan tersebut..?	Saat ini masih peminatan.., belum ada rekomendasi penjurusan.

b. Data Sekunder: berupa data mahasiswa dan nilai mata kuliah prasayarat yang diperoleh dari ruang data (database akademik).

3.4 Tempat dan Waktu Penelitian

Dalam melaksanakan penelitian ini, tempat yang dijadikan bahan penelitian yaitu Program Studi Teknik Informatik Fakultas Ilmu Komputer Universitas Dian Nuswantoro tepatnya di ruang data dan di ruang dosen di meja kepala program studi Teknik Informati Fakultas Ilmu Komputer Universitas Dian Nuswantoro. Waktu penelitian dilakukan pada bulan Maret 2013.

3.5 Tahap-Tahap Penelitian

Tahap-tahap penelitian ini meliputi

1. Tahap kompetensi untuk pemilihan peminatan

   Mata kuliah yang wajib di ambil sebelum peminatan adalah sebagai berikut :

   a. Kalkulus
   b. Fisika
   c. Matematika diskrit
   d. Dasar pemrograman
   e. Rekayasa perangkat lunak
   f. Algoritma pemrograman
   g. Sistem digital
   h. Sistem operasi
   i. Organisasi dan arsitektur komputer

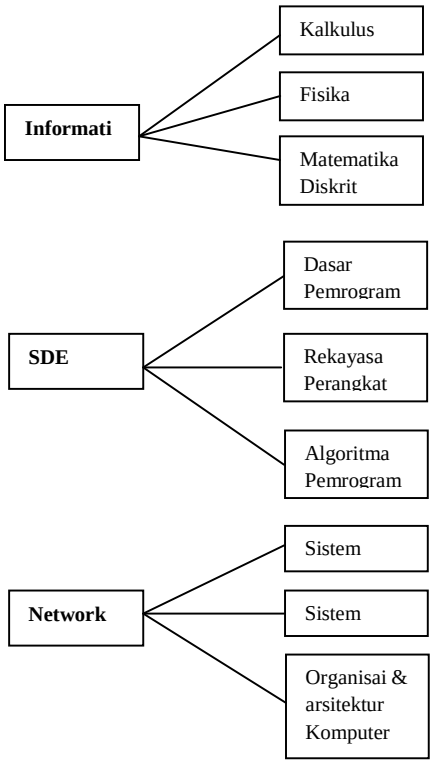
   Dan berikut adalah bidang keahlian yang akan menjadi pilihan mahasiswa:

   a. Informatik
   b. Software Developt and Enginer
   c. Network

2. Tahap Praprocessing Data

   Tahap praprocessing data adalah tahap dimana kita mencari nilai dari masing-masing mata kuliah bidang keahlian. Nilai dari mata kuliah bidang keahlian didapat dari mencari nilai rata-rata dari

setiap mata kuliah prasyarat dari masing-masing bidang keahlian. Untuk mengetahui mata kuliah prasyarat dari masing-masing bidang keahlian kita dapat melihat korelasi antara mata kuliah prasyarat dengan bidang keahlian sebagai berikut :



Gambar 3.1 : korelasi antara bidang keahlian dengan matakuliah prasyarat.

Berikut cara perhitungan nilai rata-rata mata kuliah prasyarat berdasarkan bidang keahlian.

a. Bidang keahlian Informatik

Informatik = $\frac{kalkulus + fisika + matematika\ diskrit}{3}$

b. Bidang Software Developt and Enginer

SDE = $\frac{Dasar\ pemrograman + algoritma\ pemrograman + RPL}{3}$

c. Bidang Network

Network = $\frac{Sistem\ operasi + sistem\ digital + organisasi\ dan\ arsitektur\ ko}{3}$

3. Tahap Pengolahan Data

Tahap pengolahan data adalah tahap dimana data yang telah di dapat nilai rata-ratanya tadi diproses sesuai dengan algoritma Fuzzy C-means untuk di cari nilai derajat keanggotaannya. Nilai derajat keanggotan ini yang akan digunakan sebagai acuan untuk menentukan mahasiswa masuk ke kelompok apa.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

1. Algoritma Fuzzy C-means untuk rekomendasi penjurusan dapat di terapkan dengan memanfaatkan keluaran U_i sebagai penentu setiap mahasiswa masuk ke kelompok mana dan memanfaatkan hasil keluaran V_i untuk menentukan setiap kelompok diidentifikasikan sebagai jurusan apa.
2. Dari hasil pengelompokkan kita juga dapat mengetahui rata-rata nilai

mahasiswa di setiap jurusan dengan memanfaatkan hasil keluaran V_i.

3. Algoritma Fuzzy C-means menggunakan iterasi yang berulang-ulang untuk menentukan pengelompokkan. Iterasi tersebut diulang hingga mendekati fungsi objektif yang paling optimal.

5.2 Saran

1. Penilitian ini hanya sebagai penerapan dari algoritma Fuzzy C-means. Agar lebih bermanfaat disarankan di bangun sistem informasinya.
2. Agar mendapatkan hasil clustering yang lebih optimal disarankan melakukan penelitian lebih lanjut untuk memodifikasi algoritma Fuzzy C-means atau menggabungkan algoritma Fuzzy C-means dengan algoritma lain.

Implementasi Fuzzy C-Means untuk Pengelompokan Citra Batik Berdasarkan Motif dengan Fitur Tekstur

Alvian A. Pratama, Nanik Suciati, Diana Purwitasari
Jurusan Teknik Informatika, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember
Jl. Arief Rahman Hakim, Surabaya 60111
E-mail: nanik@if.its.ac.id

Abstrak*— Batik Indonesia oleh UNESCO telah ditetapkan sebagai Warisan Kemanusiaan untuk Budaya Lisan dan Nonbendawi sejak 2 Oktober, 2009. Akan tetapi, banyak masyarakat Indonesia sendiri yang masih belum familier dengan jenis-jenis batik. Oleh karena itu, dalam tugas akhir ini diimplementasikan sistem yang bisa digunakan untuk membantu mengelompokkan jenis-jenis batik. Proses pengelompokkan citra membutuhkan fitur citra untuk dibandingkan. Di dalam tugas akhir ini, fitur citra yang digunakan untuk mengenali tekstur batik didapatkan melalui beberapa metode. Metode-metode tersebut adalah *Discrete Wavelet Transform (DWT), Rotated Wavelet Filter (RWF),* dan *Grey Level Co-occurrence Matrix (GLCM)*. Kemudian, hasil ekstraksi akan dikelompokkan ke dalam motifnya dengan menggunakan metode pengklasteran *Fuzzy C-Means* (FCM). Penggunaan FCM untuk memberikan label pada batik yang memiliki lebih dari satu motif memberikan hasil yang cukup akurat. Sayangnya hasil pengelompokan secara umum ternyata tidak terlalu baik. Hal ini kemungkinan besar dikarenakan banyaknya terjadi penghitungan nilai optimum lokal sehingga iterasi proses pengklasteran terhenti sebelum waktunya. Di sisi lain, penggunaan fitur tambahan RWF dan GLCM untuk melengkapi DWT mampu memberikan hasil pengenalan motif yang lebih baik. Dari percobaan yang dilakukan, nilai *Overall F-Measure* tertinggi didapatkan saat penggunaan *fuzziness* 2 yaitu sebesar 0,364. Hal ini menunjukkan bahwa pemilihan nilai *fuzziness* yang tepat diperlukan untuk mendapatkan hasil pengklasteran yang baik.

Kata Kunci : Discrete Wavelet Transform*, *Fuzzy C-Means Clustering*, *Grey Level Coocurrence Matrix*, Pengenalan motif batik, *Rotated Wavelet Filter

I. PENDAHULUAN

BATIK Indonesia, sebagai warisan budaya telah memiliki sejarah yang cukup panjang. Sehingga saat ini berbagai jenis batik saling mempengaruhi satu sama lain yang mengakibatkan adanya beberapa jenis batik yang memiliki lebih dari satu motif (biasanya di sebut batik *multi-label*). Hal ini terkadang menyulitkan proses pengklasifikasian yang ingin dilakukan.

Sebelumnya telah dilakukan tugas akhir yang membuat sistem untuk pengklasifikasian batik. Dalam tugas akhir tersebut digunakan *Rotated Wavelet Filter* untuk mengenali tekstur batik dan *Neural Network* untuk pengklasifikasiannya [1]. Yang membedakan dengan penelitian tersebut adalah, dalam tugas akhir ini akan digunakan beberapa metode untuk proses pengenalan tekstur. Metode-metode tersebut adalah *Grey Level Co-occurrence Matrix* (GLCM), *Discrete Wavelet Transform* (DWT), dan *Rotated Wavelet Filter* (RWF). Sedangkan proses pengklasteran akan menggunakan *Fuzzy C-Means* (FCM). Dengan menggunakan FCM, hasil pengenalan pola batik bisa memberikan lebih dari satu motif, sehingga bisa diaplikasikan untuk batik *multi-label*.

II. METODE PENELITIAN

Secara garis besar, terdapat tiga tahap utama yang perlu dilakukan dalam sebuah proses pengklasteran. Pertama adalah mengambil objek *input*. Kedua adalah proses ekstraksi fitur sebagai representasi objek. Tahap terakhir adalah proses pengklasteran itu sendiri.

Citra batik yang dijadikan objek dalam penelitian ini didapatkan dari buku Batik: Pengaruh Zaman dan Lingkungan karangan Santosa Doellah [2]. Citra tersebut berupa citra berwarna sehingga diperlukan proses *greyscaling* supaya proses ekstraksi fitur bisa berjalan dengan baik. Proses persiapan suatu citra untuk proses ekstraksi fitur seperti ini biasanya disebut preproses. Setelah preproses selesai tahap ekstraksi fitur bisa dilakukan. Proses-proses ekstraksi fitur tersebut dijelaskan di sub-bab berikut ini.

A. Grey Level Cooccurence Matrix

Dalam suatu tekstur, distribusi *grey-level pixel* bisa dideskripsikan sebagai statistik, contohnya kemungkinan dua *pixel* yang berdekatan memiliki *grey-level* yang sama. Informasi *grey-level* ini bisa dijadikan matriks dua dimensi yang disebut dengan *Grey Level Cooccurrence Matrix*. GLCM dihitung berdasarkan berapa sering suatu *pixel* dengan *grey-level* i muncul di horisontal, vertikal, atau diagonal berpasangan dengan *pixel* dengan *grey-level* j.

Beberapa fitur yang bisa diekstraksi dari GLCM adalah energi, kontras, homogenitas, dan korelasi. Energi menghitung total tiap elemen pangkat dua. Kontras menghitung variasi lokal *grey-level* dalam GLCM. Korelasi mengitung kemungkinan kemunculan secara bersamaan sepasang *pixel*. Homogenitas menunjukkan jarak distribusi elemen-elemen dalam GLCM. Dalam suatu tekstur biasanya ada homogenitas yang berulang-ulang, sehingga fitur ini paling cocok digunakan untuk klasifikasi tekstur [4].

Untuk Homogenitas dari GLCM, dihitung untuk empat arah

(θ = 0°, 45°, 90°, dan 135°). Vektor fitur dengan ukuran 4 dibuat untuk masing-masing citra. Homogenitas dihitung berdasarkan persamaan:

$$Homogenitas = \sum_{i,j} \frac{p(i,j)}{1+|i-j|} \quad (1)$$

B. Discrete Wavelet Transform

Wavelet adalah sebuah fungsi matematika yang digunakan untuk merepresentasikan sinyal waktu yang kontinyu ke dalam skala lain. Dengan *wavelet* dapat dilakukan analisis terhadap fungsi baik pada domain waktu maupun frekuensi resolusi yang berbeda [5]. Salah satu jenis *wavelet* yang terkenal adalah *wavelet* Daubachies. *Wavelet* ini banyak digunakan untuk mendefinisikan ruang multiresolusi.

Untuk kasus sinyal pada dimensi 2-D biasanya dilakukan dengan menerapkan struktur bank filter secara terpisah terhadap sinyal citra. Digunakan Low-Pass Filter atau LPF (L) dan High-Pass Filter atau HPF (H). Untuk ekstraksi fitur dalam sistem ini digunakan dekomposisi dengan varian *wavelet* Daubachies. Nilai L untuk Daubechies-4 (D4) yang akan digunakan disini adalah sebagai berikut:

$$\begin{aligned} h_0 &= (1+\sqrt{3})/4\sqrt{2} \\ h_1 &= (3+\sqrt{3})/4\sqrt{2} \\ h_2 &= (3-\sqrt{3})/4\sqrt{2} \\ h_3 &= (1-\sqrt{3})/4\sqrt{2} \end{aligned} \quad (2)$$

Sedangkan nilai H berhubungan dengan nilai L di atas yaitu sebagai berikut:

$$g_0 = h_3\,,\, g_1 = -h_2\,, g_2 = h_1\,,\, g_3 = -h_0 \quad (3)$$

Berdasarkan nilai L dan H di atas dibangun filter-filter DWT berdasarkan operasi matriks yang ditunjukkan di (4).

$$\begin{aligned} H_{ll} &= h^T h \\ H_{lh} &= h^T g \\ H_{hl} &= g^T h \\ H_{hh} &= g^T g \end{aligned} \quad (4)$$

Operasi DWT untuk suatu citra dilakukan berdasarkan algoritma berikut.

1. Nyatakan citra masukan sebagai matriks dua dimensi.
2. Konvolusikan matriks citra dengan filter yang dinginkan.
3. Lakukan *downsampling* matriks hasil dari tahap 2.

Algoritma DWT di atas dilakukan untuk setiap filter yang ada sehingga didapat 4 buah citra yang merupakan sub bidang dari citra awal. Subbidang-subbidang tersebut diberi label Low-Low (LL, didapat dari filter H_{ll}) yang berisi aproksimasi citra awal, Low-High (LH, didapat dari filter H_{lh}) yang berisi detail horizontal, High-Low (HL, didapat dari filter H_{hl}) yang berisi detail vertikal, dan High-High (HH, didapat dari filter H_{hh}) yang berisi detail diagonal. Sub bidang LL akan dijadikan *input* untuk proses dekomposisi berikutnya.

C. Rotated Wavelet Filter

Dalam proses ekstraksi tekstur, data karakteristik dari informasi diagonal memberikan tambahan signifikan dalam performa pengenalan fitur. *Rotated Wavelet Filter* bisa digunakan untuk merotasi *wavelet* filter awal. Hal ini dilakukan karena sub bidang HH di *wavelet* standar mengandung informasi diagonal tetapi susah dibedakan apakah informasi tersebut berdasarkan sudut 45° atau 135°. Apabila dilakukan rotasi sendiri maka bisa diatur arah sudut mana yang dihitung.

Filter yang digunakan dalam RWF didapatkan dengan cara merotasi filter yang bersesuaian dari filter-filter DWT sebesar sudut yang diinginkan. Dalam tugas akhir ini, sudut yang digunakan adalah 45°. Ukuran dari filter RWF adalah (2N-1) x (2N-1) dimana N adalah panjang dari filter 1-Dimensi [6]. Karena filter yang digunakan adalah D4 yang berarti panjang filter 1-D adalah 4 maka ukuran RWF adalah 7x7.

Proses ekstraksi untuk RWF sama dengan yang dilakukan untuk DWT. Tetapi data yang digunakan hanya dari sub bidang LH dan HL karena hanya dua sub bidang ini yang mengandung informasi diagonal.

Fitur yang digunakan dari DWT dan RWF untuk merepresentasikan suatu citra adalah energi. Dasar pemikiran dari penggunaan energi sebagai fitur untuk pembedaan tekstur adalah bahwasanya distribusi energi dalam domain frekuensi mengidentifikasikan sebuah tekstur [7]. Dekomposisi citra menggunakan DWT dan RWT dilakukan sampai 5 level. Energy dari DWT untuk masing-masing level, untuk setiap sub bidang (HH, HL, LH, LL) dihitung dan dijadikan vektor fitur dengan ukuran 5x4=20. Sedangkan dari RWT diambil Energi untuk masing-masing level dari sub bidang LH dan HL yang berarti ada 2x5=10 fitur. Persamaan untuk menghitung energi adalah sebagai berikut:

$$Energi = \frac{1}{MxN} \sum_{i=1}^{M} \sum_{j=1}^{N} |X_{ij}| \quad (5)$$

D. Fuzzy C-Means Clustering

Setelah didapatkan fitur-fitur tekstur berdasarkan proses-proses di atas, maka akan dilakukan proses pengelompokan data (*Data Clustering*). *Data clustering* adalah proses membagi elemen-elemen data ke klaster-klaster sehingga elemen-elemen di klaster yang sama menjadi semirip mungkin, sedangkan elemen di klaster yang berbeda menjadi setidak mirip mungkin. Metode *clustering* yang akan digunakan adalah *Fuzzy C-Means* (FCM) yang telah dimodifikasi.

Algoritma FCM berusaha membagi sekumpulan data *n* ke dalam sejumlah *c* klaster. Hasil dari algoritma akan menghasilkan list pusat klaster $C = \{c_1, \ldots, c_c\}$ dan matriks partisi $U = \mu_{ij}$, i = 1, …, *n*, j = 1, …, *c*. dimana setiap elemen μ_{ij} menunjukkan tingkat kedekatan elemen ke-i dengan klaster ke-j [8]. Persamaan untuk menghitung nilai μ_{ij} adalah sebagai berikut:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{ik}}\right)^{2/(m-1)}} \quad (6)$$

dimana

μ_{ij} : Nilai keanggotaan data ke-i dalam klaster ke-j

d_{ij} : Jarak data ke-i ke pusat klaster ke-j.

m : Parameter yang menunjukkan tingkat *fuzziness*, m>1

c : Jumlah klaster

d_{ik} : Jarak data ke-i ke pusat klaster ke-k.

Algoritma FCM ini punya satu batasan utama, yaitu total keanggotaan suatu data ke semua klaster harus sama dengan satu.

Penghitungan pusat klaster c_j yang baru didasarkan pada nilai keanggotaan masing-masing data. Persamaannya adalah sebagai berikut.

$$c_j = \frac{\sum_{i=1}^{n} {\mu_{ij}}^m \cdot x_i}{\sum_{i=1}^{n} {\mu_{ij}}^m} \quad (7)$$

Secara umum algoritma FCM bisa dinyatakan sebagai berikut:

1. Tentukan nilai keanggotaan suatu data dengan semua klaster secara random
2. Ulangi langkah-langkah berikut ini sampai terjadi konvergensi.
   a) Hitung pusat dari masing-masing klaster c_j berdasarkan (9).
   b) Hitung keanggotaan setiap data untuk masing-masing klaster μ_{ij} berdasarkan (7).

Untuk menghitung jarak antara dua data digunakan metode *Canberra Distance*. Metode ini digunakan karena nilai-nilai dalam satu vektor fitur mempunyai skala yang berbeda-beda sehingga diperlukan normalisasi yang sudah diperhitungkan dalam *Canberra Distance*. Persamaannya adalah sebagai berikut.

$$d_{xy} = \sum_{a=1}^{p} \frac{|x_a - y_a|}{|x_a| + |y_a|} \quad (8)$$

Penghitungan evaluasi untuk klaster hasil dari sistem ini dilihat menggunakan beberapa metode, yaitu *Overall F-Measure* dan penghitungan kualitas pengklasteran.

F-measure adalah suatu tes akurasi yang menggunakan *precision* dan *recall* sebagai tolak ukurnya. *Precision* adalah nilai yang diperoleh dari pembagian hasil data yang benar dengan seluruh hasil data. Karena dalam sistem ini menggunakan *fuzzy clustering* dimana suatu data merupakan anggota lebih dari satu klaster, maka nilai *precision* akan kecil. Sedangkan *recall* diperoleh dari pembagian hasil data yang benar dibagi seluruh data yang seharusnya benar [9]. *F-measure* (*F*) dihitung dari *precision p* dan *recall r* berdasarkan (9).

$$F = 2 \cdot \frac{p \cdot r}{p + r} \quad (9)$$

Dalam sistem ini, untuk masing-masing klaster hasil pengklasteran dihitung nilai *F-Measure*-nya terhadap masing-masing *dataset*. Selanjutnya dilakukan penghitungan *Overall F-Measure* (*F(C)*) sebagai nilai evaluasi sistem secara keseluruhan. Semakin tinggi nilai ini, maka semakin bagus hasil pengklasteran sistem tersebut [10]. *F(C)* dihitung berdasarkan Persamaan (10).

$$F(C) = \sum_{K_i \epsilon K} \frac{|K_i|}{|D|} max_{c_j \epsilon C}(F) \quad (10)$$

dimana

$|D|$: Total data dari *dataset*

C : Label klaster hasil dari sistem

K : Label kelas dari *dataset*

$|K_i|$: Jumlah data dalam kelas K_i

c_j : Label klaster ke-j hasil dari sistem

Sedangkan kualitas pengklasteran (biasa disebut validitas pengklasteran) dihitung berdasarkan rasio antara jarak intra klaster dan antar klaster. Jarak intra klaster diperoleh dengan cara menghitung rata-rata dari jumlah jarak masing-masing data dengan pusat klasternya. Sedangkan jarak antar klaster adalah jarak minimum antara masing-masing klaster. Semakin bagus hasil dari suatu pengklasteran, semakin kecil nilai validasinya.

III. HASIL DAN DISKUSI

Dataset yang digunakan berjumlah sebanyak 182 citra dengan jumlah motif 6. Untuk motif parang ada 29 citra, Ceplok 37 citra, Lereng 16 citra, Semen 13 citra, Lung-lungan 58 citra, dan Buketan 29 citra. *Dataset* ini diambil dari koleksi buku Danar Hadi karangan H. Santoso Doellah.

A. Uji Coba Pengklasteran

Hal pertama yang dilakukan dalam percobaan ini adalah menentukan kelas-kelas dari motif-motif dari *dataset* pengklasteran. Penentuan kelas dipilih berdasarkan persentase terbesar anggota kelompok asli dari *dataset* pengklasteran. Kemudian berdasarkan parameter yang diuji coba yaitu *fuzziness*, dilakukan penghitungan kualitas pengklasteran menggunakan *Overall F-Measure* dan validasi berdasarkan jarak intra klaster dan antar klaster.

Tabel 1.
Hasil Uji Coba Pengaruh *Fuzziness*

Fuzziness	*Overall F-Measure*	Validasi
1.1	0,315	0,007375
1.5	0,348	0,020843
2	0,364	0,005208

Rangkuman data hasil percobaan bisa dilihat pada Tabel 1. Penggunaan *fuzziness* yang berbeda memberikan pengaruh yang cukup signifikan. Nilai yang mendekati satu (yang berarti mirip dengan *k-means*) memberikan *Overall F-Measure* yang paling rendah. Hal ini bisa dimaklumi karena dengan *fuzziness* kecil, maka data-data semakin terikat sehingga jika di awal inisialisasi ada data yang salah masuk klaster, maka sampai selesai proses *clustering* data tersebut tidak akan terpisah dari klaster itu meskipun sebenarnya salah. Peningkatan nilai *fuzziness* memberikan kenaikan *Overall F-*

Measure dan hasil validasi yang lebih baik.

B. *Analisa untuk Batik Multi-Label*

Percobaan hanya dilakukan berdasarkan pengamatan visual saja, sehingga ada kemungkinan hasilnya tidak objektif. Data yang digunakan dalam percobaan ini hanya beberapa buah data. Sebagai contoh akan dilakukan percobaan untuk citra batik pada Gambar 1.

Gambar 1. Batik Patran Kangkung

Dari perhitungan didapatkan citra ini mempunyai label-label Lereng, Parang, dan Buketan. Citra ini mempunyai unsur lereng dengan motif diagonalnya, parang karena selain diagonal salah satu motif daun dianggap motif mlinjon, sedangkan detail daun-daun yang membentuk motif diagonal merupakan motif buketan.

Percobaan ini dilakukan beberapa kali pada citra-citra yang berbeda untuk mendapatkan gambaran mengenai kualitas penggunaan *fuzzy c-means* dalam pengenalan motif *multi-label*. Dari percobaan-percobaan tersebut diambil kesimpulan bahwa metode *fuzzy c-means* cukup bisa digunakan untuk mengenali batik multi label.

IV. KESIMPULAN

Kesimpulan yang bisa diambil dari tugas akhir ini adalah sebagai berikut.

1) *Fuzzy C-Means Clustering* tidak terlalu cocok digunakan dalam pengklasteran data yang terlalu tersebar seperti motif batik ini. Di sisi lain penggunaan FCM untuk pengenalan batik *multi-label* cukup memuaskan.
2) Penentuan nilai *fuzziness* mempengaruhi nilai *Overall F-Measure*. *Overall F-Measure* tertinggi didapatkan saat penggunaan *fuzziness* sebesar 2 dengan nilai 0,364.

DAFTAR PUSTAKA

[1] B. Arisandi, N. Suciati, dan A.Y. Wijaya. *Pengenalan Motif Batik dengan Rotated Wavelet Filter dan Neural Network.* Jurusan Teknik Informatika, ITS, Indonesia. 2011.

[2] Santosa Doellah. *Batik : Pengaruh Zaman dan Lingkungan*. Batik Danar Hadi Solo. 2002.

[3] G. Hoffmann. *Luminance Models for Grayscale Conversions*. 2008.

[4] D. Hazra. "Texture recognition with combined GLCM, wavelet and rotated wavelet features". *International Journal of Computer and Electrical Engineering*, Vol.3, No.1, February, 2011, pp. 1793-8163.

[5] I. Daubechies. *Ten Lectures on Wavelets*. Philadelphia: PA. 1992.

[6] M. Kokare, P. Biswas, dan B. Chatterji. "Texture image retrieval using rotated wavelet filter". *Pattern Recognition Letters* 28, 2007, pp. 1240–1249.

[7] M. Kokare, P. Biswas, dan B. Chatterji. *Wavelet Transform Based Texture Features For Content Based Image Retrieval*. Kharagpur: Electronics and Electrical Communication Engineering Department, Indian Institut of Technology. 2003.

[8] B. Thomas, G. Raju, dan S. Wangmo. "A modified fuzzy c-means algorithm for natural data exploration". *World Academy of Science, Engineering and Technology* 49, 2009.

[9] M.E.S. Mendes, dan L. Sack. "Evaluating fuzzy clustering for relevance-based information". IEEE International Conference on Fuzzy Systems, Mei, 2003, pp. 648 – 653.

[10] B.C. Fung. *Hierarchical Document Clustering Using Frequent Itemsets*. Master Thesis, Simon Fraser University, Computing Science. 2002.

Cluster Analysis Menggunakan Algoritma *Fuzzy C-means* dan *K-means* Untuk Klasterisasi dan Pemetaan Lahan Pertanian di Minahasa Tenggara

Jemaictry Tamaela[1], Eko Sediyono[2], Adi Setiawan[3]
[1,2]Magister SistemInformasiUniversitas Kristen SatyaWacana
[3]Fakultas Sain dan Matematika, Universitas Kristen Satya Wacana
Jl. Diponegoro 52-60 Salatiga, 50711
email: [1]jeyzone77@gmail.com; [2]ekosed1@yahoo.com; [3]adi_setia_03@yahoo.com

Masuk: 13 Mei 2017; Direvisi: 12 Juni 2017; Diterima: 21 Juni 2017

Abstract. *The purpose of this study is to perform cluster analysis and implementation by utilizing fuzzyc-means (FCM) and k-means (KM) to process agricultural data based on the data mining results. The fuzzy c-means (FCM) and k-means (KM) are implemented to find out and form the agricultural land clusters which appropriate the commodity types based on the supporting attributes that are used. The analysis and implementation results could provide some land information such as the number of the clusters, the land areas, the region areas, the locations and the productivity levels. The results of this study could be applied as the suggestion in converting the land functions and structuring the agricultural lands. The utilization of Openstreetmap is an open source solution which is implemented in the application. It could give visual information related to the agricultural land regions based on the clusters which make it easier to comprehend.*
Keywords: *Cluster analysis, C-means, K-means, GIS, Data mining*

Abstrak. *Penelitian ini bertujuan untuk melakukan analisis cluster dan implementasinya dengan menggunakan algoritma fuzzy c-means (FCM) dan k-means (KM) untuk mengelola data pertanian dari hasil data mining yang dilakukan. Fuzzy c-means (FCM) dan k-means (KM) dimplementasikan untuk menemukan dan membentuk klaster-klaster daerah lahan pertanian sesuai dengan jenis komoditi berdasarkan atribut-atribut pendukung yang digunakan. Hasil analisis dan implementasi dapat menyediakan informasi lahan seperti jumlah kluster, luas lahan, luas daerah, letak dan tingkat produktifitas. Hasil yang diperoleh dapat menjadi bahan masukan dalam proses alih fungsi dan penataan lahan pertanian. Penggunaan Openstreetmap merupakan solusi open source yang diimplementasikan pada aplikasi dapat memberikan informasi visual daerah-daerah lahan pertanian berdasarkan klaster yang dihasilkan sehingga lebih mudah untuk dipahami.*
Keywords: *Cluster analysis, c-means, k-means, GIS, Data mining*

1. Pendahuluan

Daerah Minahasa Tenggara (MiTra) merupkan salah satu kabupaten di Sulawesi Utara yang baru dimekarkan pada tahun 2010. Luas daerah MiTra 730,62 km^2 dan didominasi daerah yang bergunung dengan kesuburan tanah yang tinggi. MiTra menjadikan pertanian sebagai salah satu sektor unggulan dalam menunjang perekonomian daerahnya. Sebagai salah satu daerah pertanian di wilayah Sulawesi Utara, lahan-lahan pertanian terutama untuk sektor tanaman pangan perlu diperhatikan sehingga dapat menunjang kebutuhan pangan dan ekonomi daerah.

Analisis dan pengelolaan lahan pangan perlu dilakukan sehingga dapat memaksimalkan fungsi dan alokasi lahan yang ada untuk memperoleh hasil yang maksimal. Sistem informasi geografis (GIS) dapat memberikan keuntungan yaitu gambaran lahan pangan dari sudut yang

berbeda, panduan pengolahan lahan pangan, dan informasi pendukung keputusan yang dapat digunakan untuk tindakan yang akan dilakukan (Kisilevich, dkk., 2013).

Salah satu cara yang dapat dilakukan untuk mengelola data geografis lahan adalah dengan melakukan *clustering* menggunakan algoritma *fuzzy c-means* (FCM) dan *k-means* (KM). *Fuzzy clustering* melakukan pembobotan terhadap tingkat keanggotaan himpunan *fuzzy*-nya (Bezdek & Dunn, 1975). FCM merupakan metode pengelompokan yang memungkinkan satu bagian data memiliki dua atau lebih kelompok (Dunn, 1973). Dengan FCM, objek akan cenderung dibagi menjadi anggota suatu kelompok sesuai derajat keanggotaannya (Jansen, 2007). *K-means* (KM) dikembangkan oleh McQueen pada 1967. KM merupakan algoritma *clustering* yang populer dan bekerja dengan cara membagi data ke dalam partisi data (Juanying, dkk., 2011).

Analisis *clustering* dengan menggunakan algoritma FCM dan KM dapat digunakan untuk mempermudah proses klasterisasi lahan pertanian berdasarkan jenis dan atribut yang telah ditentukan. Selain itu, hasil analisis dan implementasi dapat menjadi bahan pertimbangan dalam proses analisis dan pengambilan keputusan terkait alih fungsi dan alokasi lahan diwaktu yang akan datang. Penggunaan Algoritma FCM dan KM bertujuan untuk membandingkan variasi hasil kluster yang dihasilkan kedua algoritma dalam implementasinya pada aplikasi yang dibuat berdasarkan pada banyaknya jumlah data yang tersedia.

Berdasarakan latar belakang yang telah dijelaskan, maka dilakukan penelitian mengenai *cluster analysis* dan implementasinya menggunakan algoritma FCM dan KM untuk klasterisasi dan pemetaan lahan pertanian untuk lahan pertanian di Minahasa Tenggara. Implementasi dilakukan dalam bentuk aplikasi desktop.

2. Tinjauan Pustaka

2.1. Penelitian Terdahulu

Beberapa penelitian terdahulu berkaitan dengan analisis *clustering* dengan algoritma FCM dan KM diantaranya dilakukan Ali, dkk. (2008) melakukan review dan analisis algoritma-algoritma *clustring* yang efektif dalam Segmentasi citra terutama pada citra berbasis *fuzzy*. Vendruscolo & Kaleita (2011) menerapkan FCM untuk menghitung data spasial pertanian dengan tujuan mengetahui jumlah klaster optimal dari tanaman yang homogen, proses ini divalidasi menggunakan dua atribut yaitu *Partition Coefficient (PC)* dan *Classification Entropy(CE) and Dunn's Index (DI).* Meskipun hanya menggunakan dua atribut tetapi telah menunjukan hasil yang menjanjikan. Martino & Sessa (2009) menerapakan *extended fuzzy c-means* yang merupakan alternatif dari FCM ke dalam GIS untuk menentukan daerah-daerah titik panas. Afifah, dkk. (2016) menerapkan algoritma FCM untuk pengklasteran lahan sawah di Indonesia sebagai evaluasi ketersediaan produksi pangan. Hardiani, dkk. (2014) menggunakan algoritma FCM dan RFM dalam proses *data mining* untuk *customer relationship management* dalam kajian *data mining customer relationship management* pada lembaga keuangan mikro. Rodrigues & Cora (2015) menerapkan algoritma FCM dalam melakukan identifikasi manajemen lahan berdasarkan variabilitas tanah dan lahan jagung.

Miyamoto (2011) melakukan penelitian tentang perbedaan fungsi-fungsi objek dalam algoritma FCM dan *Kernel-Based Clustering.* Ghosh & Dubey (2013) melakukan analisis dan perbandingan performa algoritma KM dan FCM dalam implementasinya pada perangkat lunak. Sankoh, dkk. (2015) menerapkan kombinasi algoritma *Neural Network,* KM dan *Particle Swarm Optimization* untuk melakukan klasterisasi jenis musik. Ong (2013) menerapkan algoritma KM dalam penelitiannya untuk menentukan strategi *marketing* President University. Dalam penelitiannya menggunakan tiga atribut yaitu asal, jurusan dan GPA. Tujuan dan hasil penelitian ini digunakan untuk membantu tim promosi dalam memprediksi strategi promosi yang akan dilakukan oleh President University di Indonesia.

Selain dengan algoritma FCM dan KM, dilakukan proses *data mining* untuk mengelola data-data geografis MiTra yang akan digunakan sebagai basis data pada proses klusterisasi. Lanzi, dkk. (2012) dalam penelitian *Data Mining in GIS: A Novel Context-Based Fuzzy Geographically Weighted Clustering Algorithm* menyebutkan sistem informasi geografis (GIS)

memegang peranan penting dalam berbagai penelitian dan industri. Penelitian dilakukan untuk menemukan aspek-aspek penting *data mining* dalam GIS dengan menggunakan GIS *clustering* dengan analisis *Geo-Demographic*. Penelitian yang dilakukan Kaur, dkk. (2014) dengan menerapkan algortima KM dan *K-Nearest Neighbor* (KNN) dalam *data mining* pertanian untuk memprediksi harga tanaman. Peng, dkk. (2011) dalam penelitian *An incident information management framework based on data integration, data mining, and and multi-criteria decision making*, menyebutkan pentingnya integrasi data suatu kejadian yang sejenis dalam satu format yang sama. Kemudian dengan menggunakan teknik *data mining* dapat diidentifikasi pelayanan yang sesuai untuk keadaan sebelum dan sesudah kejadian dalam sistem manajemen informasinya.

Berdasarkan tinjauan pustaka yang ada maka dalam penelitian ini akan dilakukan *cluster analysis* dengan menggunakan algoritma *fuzzy c-means* (FCM) dan *k-means* (KM) untuk proses klasterisasi dan pemetaan lahan pertanian di kabupaten Minahasa Tenggara.

2.2. Landasan Teori

Cluster analysis merupakan pembelajaran tanpa pengawasan. *Clustering* bekerja dengan menemukan struktur dalam koleksi data yang tidak di label dan mengelompokannya ke dalam kelompok yang saling memiliki kesamaan (Madhulatha, 2012). FCM merupakan metode pengelompokan yang memungkin satu data memiliki dua atau lebih kelompok (Dunn, 1973) dan diperbaiki oleh bezdek tahun 1981. Langkah-langkah dalam FCM dapat dilihat pada Persamaan (1). Untuk setiap nilai m merupakan bilangan asli lebih dari 1, u_{ij} adalah derajat keanggotaan dari x_i di klaster j, c_j merupakan pusat dimensi klaster. Untuk memaksimalkan hasil klaster, dilakukan perulangan dengan memperbarui keanggotaan u_{ij} dan pusat-pusat klaster c_j dengan Persamaan (2). Proses perulangan ini berhenti ketika keanggotan dan pusat-pusat klaster tidak lagi berubah atau berpindah dengan Persamaan (3).

$$J_m = \sum_{i-1}^{N}\sum_{j-1}^{C} u_{ij}^m \|x_i - c_j\|^2 \tag{1}$$

$$u_{ij} = \frac{1}{\sum_{k-1}^{C}\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}} \qquad c_j = \frac{\sum_{i-1}^{N} u_{ij}^m \cdot x_i}{\sum_{i-1}^{N} u_{ij}^m} \tag{2}$$

$$\max_{ij} = \left\{\left|u_{ij}^{(k+1)} - u_{ij}^{(k)}\right|\right\} < \varepsilon \tag{3}$$

Algoritma *K-Means* (KM), dikembangkan oleh MacQueen tahun 1967, merupakan salah satu algoritma sederhana dalam memecahkan masalah pengelompokan. MacQueen menjelaskan dalam algoritma k-means, dataset dibagi ke dalam beberapa klaster, dan satu data hanya dapat menjadi satu anggota klaster tertentu (MacQueen, 1967). *K-means* menentukan terlebih dahulu mean atau titik tengah kluster berdasarkan jumlah kluster yang dibuat, kemudian *k-means* membagi data set ke dalam masing-masing kluster berdasarkan nilai yang sama atau yang mendekati nilai tengah kluster (Zijun, 2012). Algoritma ini bertujuan untuk meminimalisasi suatu fungsi objek. Fungsi objek dalam k-means dapat ditentukan dengan Persamaan (4), dimana $\|x_i^{(j)} - c_j\|^2$ merupakan jarak titik data $x_i^{(j)}$ dan pusat klaster c_j merupakan indikator jarak poin n data dari pusat klaster masing-masing.

$$J = \sum_{j-1}^{k}\sum_{i-i}^{n} \left\|x_i^{(j)} - c_j\right\|^2 \tag{4}$$

Langkah-langkah dalam algoritma k-means: (1) Menentukan titik k awal pada objek yang akan diklaster. (2) Menetapkan setiap objek dalam kelompok yang memiliki nilai objektif terdekat. (3) Ketika semua objek telah ditetapkan, posisi titik k dihitung kembali. (4) Langkah ke-2 dan ke-3 diulangi sampai tidak lagi berubah.

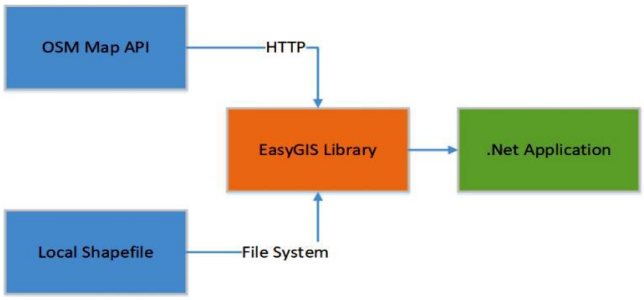
3. Metodologi Penelitian

Penelitian dilakukan melalui beberapa tahap. Tahap pertama yaitu: (1) Tahap studi pendahuluan. Pada tahap ini dilakukan diidentifikasi masalah yang ada. (2) Tahap analisis dan implementasi. Pada tahap ini dilakukan pengumpulan pustaka pendukung dan parameter untuk proses analisis dan implementasi. (3) Tahap kemanfaatan. Pada tahap ini dipelajari manfaat yang dihasilkan dari hasil analisis dan implementasi.

Dalam penelitian ini terdapat beberapa tahapan *data mining* yang dilakukan. Tahapan ini dilakukan untuk memaksimalkan data-data pertanian MiTra yang dimiliki. Tahapan *data mining* yang dilakukan dalam penelitian ini: (1) *Data cleaning*, tahap ini menghapus data gangguan atau yang tidak diperlukan dan mengisi data yang hilang untuk meningkatkan kualitas data (Joshi & Usha, 2015). (2) Data integrasi, tahap ini menggabungkan berbagai sumber data yang saling berhubungan. Modul integrasi data yang heterogen dan terintegrasi harus disajikan dalam format yang seragam (Peng, dkk., 2011). (3) Pemilihan data, tahap memilih data yang relevan dengan dengan tujuan penelitian. (4) Transformasi data, tahap ini mengubah data menjadi format baru untuk diproses dalam *data mining*. (5) *Data mining*, tahap ini menerapkan metode cerdas untuk ekstraksi pola. Pada tahap ini digunakan algoritma FCM dan KM yang di implementasikan dalam aplikasi. (6) Pola presentasi, tahap ini untuk presentasikan pola yang diperoleh.

4. Hasil dan Pembahasan

Aplikasi yang dihasilkan dibuat sendiri dengan teknologi *.NET framework* 4.6 menggunakan bahasa pemrograman C# dan menggunakan basis data MySQL. Teknologi . *NET framework 4.6* digunakan untuk mempermudah pengembangan aplikasi dan sebagai penerjemah agar aplikasi dapat dijalankan oleh sistem. Aplikasi diintegrasikan dengan *OpenStreetMap* (OSM) API. OSM API disediakan *OpenStreetMap* dalam format *javascript* sehingga pengguna dapat melakukan kostumisasi pada peta yang digunakan dan menyediakan akses langsung ke server *OpenStreetMap*. Integrasi OSM API dengan aplikasi yang dibuat dapat dilihat pada Gambar 1.



Gambar 1. Integrasi OSM API dengan aplikasi

Aplikasi yang dihasilkan digunakan untuk mengklasterkan lahan pertanian untuk empat komoditi pangan yaitu padi sawah, padi ladang, jagung dan ubi kayu. Hasil *data mining* dari data-data pertanian MiTra digunakan sebagai basis data untuk proses *mining* dan klasterisasi melalui aplikasi. Tabel data hasil *mining* dari data-data MiTra dapat dilihat pada Gambar 2 dan Gambar 3.

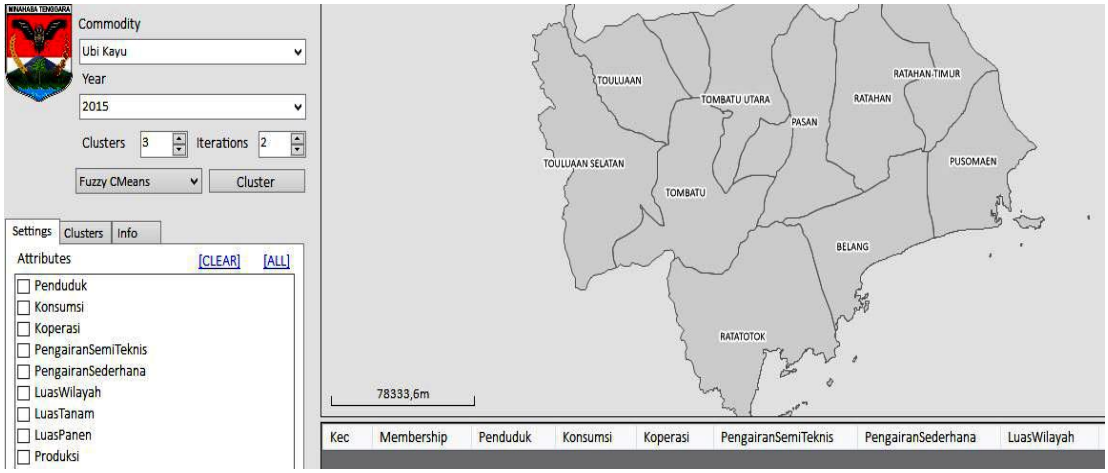
Wilayah	Luas Wilayah (Ha)	Penduduk	Produksi Padi Sawah			Produksi Padi Ladang			Model Pengairan Sawah	
			Luas Tanam (Ha)	Luas Panen (Ha)	Produksi (Ton)	Luas Tanam (Ha)	Luas Panen (Ha)	Produksi (Ton)	Semi Teknis	Sederhana
Ratatotok	10.418	12.696	18	15	69	117	102	317,3	0	9
Posumean	5.362	8.392	352	347	1.792	15	12	37,4	96	150
Belang	7.517	16.290	651	632	3.318	45	42	131	115	130
Ratahan	6.163	12.981	780	753	3.773	15	12	38	276	111
Pasan	4.979	6.546	527	506	2.631	25	22	39	53	120
Ratahan Timur	6.399	5.763	146	138	653	50	48	149	0	1
Tombatu	6.795	8.870	630	604	3.068	110	103	321	184	142
Tombatu Timur	1.881	8.690	852	802	3.984	30	25	78	317	273
Tombatu Utara	3.717	7.773	990	974	4.798	0	0	0	0	366
Touluaan	10.180	6.420	375	356	1.738	0	0	0	0	275
Touluaan Selatan	5.276	3.910	50	42	194	130	125	389	0	222
Silian Raya	4.375	5.487	512	482	2.040	0	0	0	0	241

Gambar 2. Data hasil transformasi

Produksi Jagung			Produksi Ubi Kayu			Koperasi Pemasaran
Luas Tanam (Ha)	Luas Panen (Ha)	Produksi (Ton)	Luas Tanam (Ha)	Luas Panen (Ha)	Produksi (Ton)	
947	920	3.523	25	21	262	0
885	860	3.050	12	8	98	0
1203	1105	3.809	18	15	187	3
852	810	2.904	9	5	61	0
871	842	2.975	12	7	88	0
564	523	1.882	8	5	60	0
387	362	1.282	11	8	98	0
953	915	3.027	15	11	141	4
381	352	1.225	8	5	61	4
320	295	1.023	12	7	88	0
543	512	1.728	15	8	98	0
327	295	1.017	20	12	150	0

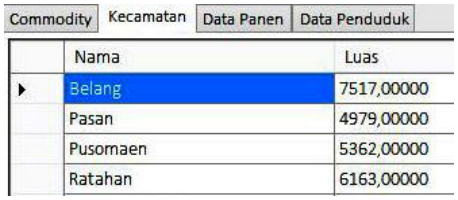
Gambar 3. Data hasil transformasi

Dalam penelitian ini data yang digunakan sebagai data *sample* selain data hasil observasi, data-data diperoleh dari data Indikator Kesejahteraan Rakyat Kabupaten Minahasa Tenggara (Badan Pusat Statistik Kabupaten Minahasa Selatan, 2015a) dan data Kabupaten Minahasa Tenggara Dalam Angka tahun 2015 (Badan Pusat Statistik Kabupaten Minahasa Selatan, 2015b) dan data Kabupaten Minahasa Tenggara Dalam Angka tahun 2016(Badan Pusat Statistik Kabupaten Minahasa Selatan, 2016). Aplikasi dibuat dengan tampilan yang sederhana untuk mempermudah *user* dalam menggunakan aplikasi. Contoh tampilan aplikasi dapat dilihat pada Gambar 4.



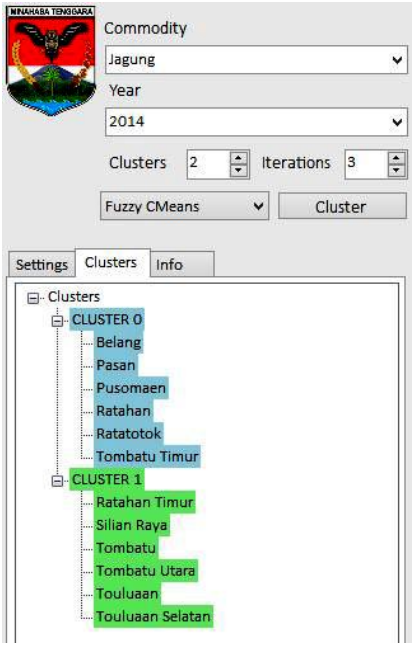
Gambar 4. Tampilan aplikasi

Bagian kiri merupakan fitur utama dari aplikasi yang terdiri dari komoditi, tahun, kluster, iterasi dan algoritma yang diterapkan. *Settings* berisi atribut yang digunakan dalam proses klasterisasi. *Clusters* berisi informasi kluster yang dihasilkan. Info berisi informasi atribut yang digunakan dan nilai masing-masing. Untuk menambah data ke dalam aplikasi, disediakan *master data*. Pada bagian ini, *user* dapat mengisi dan merubah data-data yang diperlukan sesuai dengan jenis datanya. Contoh *master data* dapat dilihat pada Gambar 5.



Gambar 5. *Master Data*

Pengujian aplikasi dilakukan dengan menggunakan data Minahasa Tenggara. Pengujian dilakukan dengan menggunakan data tahun 2014 yang diperoleh dari data Minahasa Tenggara Dalam Angka Tahun 2015 dengan contoh komoditi jagung berdasarkan atribut luas tanam dan luas panen. Proses pengujian pertama menggunakan algoritma FCM. Contoh hasil kluster yang dihasilkan dengan algoritma FCM dapat dilihat pada Gambar 6.

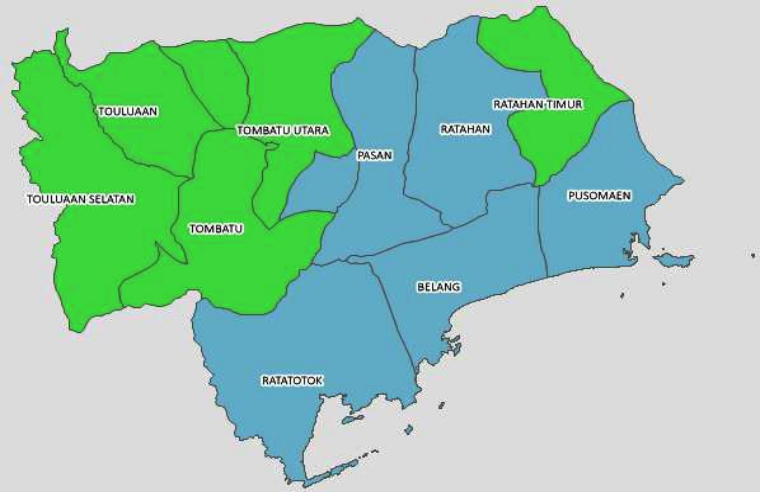


Gambar 6. Hasil kluster FCM

Hasil pada Gambar 6 adalah proses klasterisasi dengan menggunakan dua kluster dengan iterasi sebanyak tiga putaran. Daerah Belang, Pasan, Posumaen, Ratatotok dan Tombatu timur masuk ke dalam klaster nol atau klaster pertama dengan nilai mean klaster untuk luas tanam 793,11 dan luas panen 753,56 ditandai dengan Warna biru. Daerah Ratahan Timur, Silian Raya, Tombatu, Tombatu Utara, Touluaan, Touluaan Selatan merupakan anggota klaster satu atau klaster ke dua dengan nilai *mean* klaster berdasarkan luas tanam 420,33 dan luas panen 389,83 ditandai dengan warna hijau. Contoh informasi hasil kluster yang berisi nilai *membership* dan *mean* masing-masing kecamatan dapat dilihat pada Gambar 7. Dari hasil klusterisasi yang dilakukan, peta hasil kluster yang dihasilkan dapat dilihat pada Gambar 8.

Kec	Membership	LuasTanam	LuasPanen
Ratahan Timur	[0,42] [0,58]	564	523
Silian Raya	[0,35] [0,65]	327	295
Tombatu	[0,28] [0,72]	387	362
Tombatu Utara	[0,29] [0,71]	381	352
Touluaan	[0,35] [0,65]	320	295
Touluaan Selatan	[0,40] [0,60]	543	512
-MEAN-		**420,33**	**389,83**

Gambar 7. Informasi *membership* kluster FCM



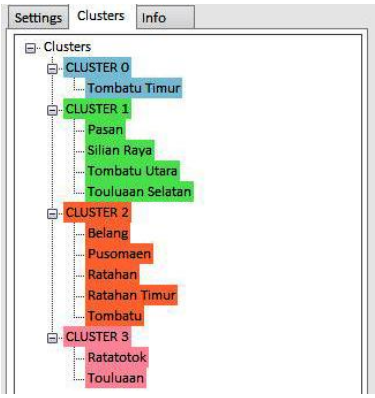
Gambar 8. Peta hasil kluster jagung FCM

Hasil akhir kluster dengan menggunakan algoritma FCM untuk komoditi jagung dengan atribut luas tanam dan luas panen dapat dilihat pada Tabel 1. Dari hasil Tabel 1, karena pada iterasi ke tiga memberikan hasil yang sama dengan iterasi sebelumnya (iterasi ke dua), maka proses iterasi dihentikan pada iterasi ke tiga. Pada percobaan dengan menggunakan tiga kluster, ditemukan adanya kluster kosong atau kluster yang tidak memiliki anggota karena nilai keanggotaan dari masing-masing atribut kecamatan tidak memenuhi kriteria dari klaster tersebut. Dari hasil percobaan dengan menggunkan komoditi jagung berdasarkan luas tanam dan luas panen tahun 2014 maka diperoleh hasil daerah dengan luas lahan pertanian dan panen jagung terbesar berada pada wilayah Belang, Pasan, Posumaen, Ratahan, Ratatotok dan Tombatu Timur, sedangkan wilayah dengan lahan pertanian jagung dan panen terkecil berada di Ratahan Timur, Silian Raya, Tombatu, Tombatu Utara, Touluaan dan Touluaan Selatan.

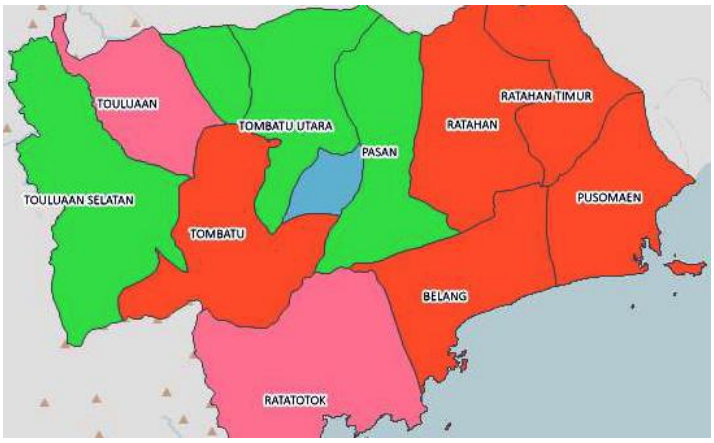
Tabel 1. Hasil kluster FCM

Kecamatan	**Cluster 0**			**Cluster 1**		
	1	2	3	1	2	3
Belang	✓				✓	✓
Pasan	✓				✓	✓
Posumaen	✓				✓	✓
Ratahan	✓				✓	✓
Ratahan Timur	✓	✓	✓			
Ratatotok	✓				✓	✓
Silian Raya		✓	✓	✓		
Tombatu				✓	✓	✓
Tombatu Timur	✓	✓	✓			
Tombatu Utara		✓	✓	✓		
Touluaan	✓	✓	✓			
Touluaan Selatan	✓	✓	✓			

Pengujian dengan menggunakan algoritma KM dengan menggunakan data tahun 2014, menggunakan komoditi jagung dengan atribut yang berbeda dengan pengujian pertama yaitu luas wilayah dan luas tanam. Pengujian ini dilakukan dengan menggunakan empat kluster dengan iterasi sebanyak tiga kali. Hasil kluster dari aplkasi dengan menggunakan algoritma KM menggunakan empat kluster dapat dilihat pada Gambar 9. Peta yang dihasilkan dari hasil klaster dengan algoritma KM dapat dilihat pada Gambar 10.



Gambar 9. Hasil kluster dengan algoritma KM dengan empat kluster



Gambar 10. Peta hasil 4 klaster jagung algoritma KM

Dari peta yang dihasilkan, Tombatu berada dalam kluster sendiri yaitu kluster nol dengan nilai mean kluster berdasarkan luas wilayah 1.881 dan luas tanam 953 ditandai dengan warna biru. Daerah Pasan, Silian Raya, Tombatu Utara dan Touluan berada dalam kluster satu dengan nilai mean kluster berdasarkan luas wilayah 4.568,75 dan luas tanam 530,5 ditandai dengan warna hijau. Daerah Belang, Posumaen, Ratahan, Ratahan Timur dan Tombatu berada dalam kluster dua dengan nilai mean kluster berdasarkan luas wilayah 6.447,2 dan luas tanam 778,2 ditandai dengan warna merah. Daerah Ratatotok dan Touluaan berada dalam kluster tiga dengan nilai mean kluster berdasarkan luas wilayah 10.299 dan luas tanam 633,5 ditandai dengan warna merah muda. Dengan demikian diperoleh hasil kluster lahan pertanian untuk komiditi jagung dengan wilayah terkecil berada di daerah Tombatu dan terbesar berada di daerah Ratatotok dan Touluan. Proses kluster lahan pangan dengan algoritma KM menggunakan empat kluster dengan tiga iterasi dapat dilihat pada Tabel 2.

Tabel 2. Hasil Klaster KM

Kecamatan	Cluster 0			Cluster 1			Cluster 2			Cluster 3		
	1	2	3	1	2	3	1	2	3	1	2	3
Belang								✓	✓			
Pasan				✓	✓	✓						
Posumaen							✓	✓	✓			
Ratahan							✓	✓	✓			
Ratahan Timur							✓	✓	✓			
Ratatotok										✓	✓	✓
Silian Raya				✓	✓	✓						
Tombatu							✓	✓	✓			
Tombatu Timur	✓	✓	✓									
Tombatu Utara				✓	✓	✓						
Touluaan										✓	✓	✓
Touluaan Selatan							✓	✓	✓			

Dari hasil Tabel 2, diketahui karena pada iterasi ketiga memberikan hasil yang sama dengan iterasi sebelumnya (iterasi kedua), maka proses iterasi dihentikan pada iterasi ketiga. Hasil Tabel 2 menunjukan kluster nol merupakan daerah dengan nilai terendah dan kluster tiga merupakan daerah dengan nilai tertinggi.

5. Kesimpulan

Berdasarkan hasil implementasi dan analisis, maka penelitian ini memperoleh kesimpulan, algoritma FCM dan KM yang diimplementasikan pada aplikasi dapat berjalan dengan baik dalam mengklasterkan daerah-daerah lahan pertanian sesuai komoditi berdasarkan pada jenis atribut yang digunakan. Dari berbagai data yang ada, data dalam bentuk angka dipilih untuk digunakan sebagai data pengujian. Hal ini dikarenakan data angka dapat dihitung nilai *mean* atau rata-ratanya. Selain itu pada penelitian FCM dan KM hanya dapat diimplementasikan untuk memproses data dalam bentuk angka.

Pada pengujian, proses iterasi maksimal didapat pada iterasi ketiga karena pada iterasi selanjutnya tidak terjadi perubahan anggota kluster. Hal ini dipengaruhi oleh banyaknya jumlah data yang tersedia. Dengan jumlah data yang terbatas, algoritma KM dapat bekerja lebih baik dibandingkan FCM dengan menghasilkan jumlah kluster yang lebih banyak atau bervariasi.

Aplikasi yang dihasilkan mudah digunakan dalam proses pengisian data, pemilihan komoditi, penentuan banyaknya klaster, jumlah iterasi yang diperlukan, visualisasi dan pemilihan atribut-atribut yang dapat disesuaikan dengan kebutuhan *user*. Hasil yang diperoleh dapat menjadi masukan kepada pemerintah daerah MiTra dalam mengelola dan memaksimalkan potensi lahan pertanian daerah.

Penggunaan *Openstreetmap* merupakan solusi *open source* yang dapat digabungkan dengan aplikasi sehingga dapat memberikan informasi visual daerah-daerah pertanian berdasarkan klaster yang dihasilkan sehingga lebih mudah untuk dipahami. Selain memperoleh data dalam bentuk angka, *user* juga memperoleh gambaran letak daerah berdasarkan peta yang dihasilkan.

6. Saran

Dari keseluruhan pemaparan yang dijelaskan sebelumnya dalam penelitian, dapat disimpulkan bahwa pada penelitian ini dilakukan proses klasterisasi dan pemetaan lahan pertanian pangan daerah MiTra sehingga dapat diketahui daerah-daerah pertanian berdasarkan komoditi dan atributnya. Untuk dapat memaksimalkan hasil penelitian, perlu dilakukan kajian lebih lanjut dengan memperhitungkan kemiringan lahan, struktur tanah, kecocokan komoditi dengan tanah lahan, potensi bencana dan banyaknya daerah aliran sungai (DAS) yang dapat digunakan dalam menunjang lahan pertanian. Hal ini juga perlu dilakukan sehingga alih fungsi lahan menjadi lahan pertanian lain atau pemukiman penduduk dapat dilakukan dengan tepat sehingga daerah yang memiliki potensi lahan pertanian yang tinggi dapat dimanfaatkan dengan maksimal. Kajian ini yang belum dilakukan dalam penelitian ini dikarenakan keterbatasan waktu dan data yang tersedia.

Referensi

Afifah, N., Rini, D.C., Lubab, A. (2016). Pengklasteran Lahan Sawah di Indonesia Sebagai Evaluasi Ketersediaan Produksi Pangan Menggunakan Fuzzy C-Means. *Jurnal Matematika "MANTIK", 2* (1).

Ali, M.A., Karmakar, G.C., & Dooley, L.S. (2008). Review On Fuzzy Clustering Algorithm. *IETECH Journal Of Advance Computations, 2* (3), 169-181.

Badan Pusat Statistik Kabupaten Minahasa Selatan. 2015a. Indikator Kesejahteraan Rakyat Kabupaten Minahasa Tenggara 2015. Badan Pusat Statistik Kabupaten Minahasa Selatan.

Badan Pusat Statistik Kabupaten Minahasa Selatan. 2015b. Minahasa Tenggara Dalam Angka Tahun 2015. Badan Pusat Statistik Kabupaten Minahasa Selatan.

Badan Pusat Statistik Kabupaten Minahasa Selatan. 2016. Minahasa Tenggara Dalam Angka Tahun 2016. Badan Pusat Statistik Kabupaten Minahasa Selatan.

Bezdek, J.C., & Dunn, J.C. (1975). Optimal fuzzy partitions: A heuristic for estimating the parameters in a mixture of normal dustrubutions. *IEEE Transactions on Computers,* 835-838.

Bezdek, J.C. (1981). Pattern Recognition with Fuzzy Objective Function Algorithms, *Plenum Press*, New York.

Dunn, J. C. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics, 3,* 32-57

Ghosh, S., & Dubey, S.K. (2013) Comparative Analysis of K-Means and Fuzzy C-Means Algorithms. *International Journal of Advance Computer Science and Aplications (IJACSA), 4* (4).

Hardiani. T., Sulistyo. S., & Hartanto. R. (2014). Kajian Data Mining Customer Relationship Management Pada Lembaga Keuangan Mikro.S2 Teknik Elektro dan Teknologi Informasi, Universitas Gadjah Mada Yogyakarta. *Jurnal Sistem Informasi Bisnis Undip 3.*

Joshi, S., & Usha, T. (2015). Data Cleaning Using Based Data Mining Technique. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 4* (2).

Juanying, X., Shuai, J., Wexin, X., & Xinbo, G. (2011). An Efficient Global K-means Clustering Algorithm. *Journal of Computer, 6* (2).

Kaur, M., Gulati, H., & Kundra, H. (2014). Data Mining in Agricultur on Crop Price Prediction: Techniques and Application. *International Journal of Computer Applications, 99* (12).

Kisilevich, S., Keim, D., & Rokach, L. (2013). A GIS-based decision support system for hotel room rate estimation and temporal price prediction: The hotel brokers ontext. *Journal Decision Support Syst, 54,* 1119–1133.

Lanzi, P. L., Cuong, B. C., Hung, H. A., & others. (2012). Data Mining in GIS: A Novel Context-Based Fuzzy Geographically Weighted ClusteringAlgorithm. *International Journal of Machine Learning and Computing, 2*(3), 235.

MacQueen, B.J. (1967): "Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability".*

Madhulatha, T.S. (2012). An Overview On Clustering Methods. *IOSR Journal of Engineering, 2* (4), 791-725

Martino, F.D., & Sessa, S. (2009). Implementation of the Extended Fuzzy C-Means Algorithm in Geographic Information System. *Journal of Uncertain System,3* (4), 298-306.

Miyamoto, M. (2011). Different Objective Function Functions in Fuzzy C-means Algorithms and Kernel-Based Clustering. *International Journal of Fuzzy System, 13* (2).

Ong, J.O. (2013). Implementasi algoritma K-means Clustering Untuk Menentukan Strategi Marketing President University. *Jurnal Ilmiah Teknik Industri (JITI) 12* (1), 10-20.

Peng, Y., Zhang, Y., Tang, Y., & Li, S. (2011). An incident information management framework based on data integration, data mining, and multi-criteria decision making. *Decision Support Systems, 51*(2), 316–327.

Rodrigues, E.S., & Cora, J.E. (2015). Managemen Zones Using Fuzzy Clustring Based On Spatial-Temporaal Variability Of Soil And Corn Yield. *Journal of the Brazilian Association of Agricultural Engineering Journal, 3,* 470-483.

Sankoh, A.S., Musthafa, A.R., Rosidi, Imron, M., & Arifin, A.Z. (2015). Klasterisasi Jenis Musik Menggunakan Kombinasi Neural Network, K-Means dan Particle Swarm Optimization. *Journal Buana Informatika, 6* (3).

Vendruscolo, L.G., & Kaleita, A.F. (2011). Modeling Zone Management in Precision Agriculture Through Fuzzy C-means Technique at Spatial Database. *Digital Repository,* Iowa State University.

Zijun, Z. (2012). K-mean algorithm cluster analysis in data mining.

Clustering Data Nilai Siswa SMA Untuk Penentuan Jurusan Menggunakan Algoritma Fuzzy C-Means

Fajar Akbar[1)]

[1]) Program Pascasarjana Magister Ilmu Komputer STMIK Nusa Mandiri
Jl. Salemba Raya No. 5 Jakarta Pusat
Fajar.fkb@bsi.ac.id

ABSTRACT

Determination of the majors a process that occurs at the time the class- X will go up to class- IX, this is because the addition proses is carried out must be based in the interests and talent of each student's.In addition, this process is done by Bp/Bk, witch this assessment process based on value evaluation of each subject's flagship from each Department, to major in Science related subject (biology, physics, chemistry, mathematics), to social scienece Departement (economics, sociology, geography, history), and for the Departement of IPB (Indonesian Language, English, Germany Language).In this reaseach FCM Clustering can assist and improve value for Determination of the majors process 72%.

KeyWords:
Fuzzy C-Means, Department, Clustering

ABSTRAKSI

Penentuan jurusan merupakan proses yang terjadi pada saat siswa kelas-X akan naik ke kelas- IX, ini dikarenakan proses penjurusan ini dilakukan harus berdasarkan minat dan bakat dari setiap siswa yang ada. Proses penjurusan ini dilakukan oleh bagian Bp/Bk, yang mana proses penilaian ini berdasarkan nilai evaluasi dari setiap mata pelajaran unggulan dari tiap-tiap jurusan, untuk jurusan mata pelajaran yang terkait IPA (Biologi, Fisika, Kimia, Matematika), untuk jurusan IPS (Ekonomi, Sosiologi, Geografi, Sejarah), dan untuk jurusanIPB (Bahasa Indonesia, Bahasa Inggris dan Bahasa Jerman). Didalam penellitian ini clustering dapat membantu dan meningkatkat ketepatan pemilihan jurusan atau pemilihan kosentrasi sebesar 72 % .

Kata Kunci:
Fuzzy C-Means, Jurusan, klustering

PENDAHULUAN

Sesuai dengan Permendikbud 81A tahun 2013 tentang implementasi kurikulum 2013 menjelaskan bahwa pendidikan merupakan proses yang sistematik untuk meningkatkan martabat manusia secara holistik yang memungkinkan potensi diri berkembang secara optimal. Peminatan atau yang lebih disebut penjurusan bertujuan untuk lebih memfokuskan atau mengarahkan materi pembelajaran para siswa-siswi sekolah menengah agar sesuai dengan minat dan bakat yang dimiliki oleh para siswa-siswi sekolah. Peminatan atau penjurusan ini juga membantu memberi gambaran akan bidang yang nantinya akan ditekuni oleh para siswa-siswi setelah lulus sekolah menengah nanti. Jadi para siswa-siswi yang telah lulus tidak kesulitan menentukan bidang yang akan ditekuni ketika kuliah atau ketika bekerja nanti.

Dengan metode *clustering* data yang ada akan dikelompokan menjadi beberapa klaster yang masing- masing diwakili oleh pusat klaster. Data yang diklasifikasikan berdasarkan hasil akhir dari evaluasi nilai yang diperoleh siswa dan siswi berdasarkan setiap mata pelajaran yang mewakili setiap jurusan. Sedangkan "Algoritma Genetika banyak digunakan untuk masalah optimalisasi yang sulit dilakukan oleh metode konvensional"(Arochman,1-9). "Metode ini dalam proses pnyelesaiannya memformulasikan terminologi genetika yang diterapkan kedalam permasalahan sehari – hari". (Wijaya, Bima Sukma dan Windarto,189-197).

Tujuan dari penelitian ini diharapkan dapat membantu menyelesaikan permasalahan pemilihan jurusan agar tidak ada kealahan dalam menentukan penjurusan bagi para siswa dan memperoleh hasil pengelompokan jurusan yang sesuai dan tepat dengan minat dan potensi setiap siswa-siswi agar bisa mengembangkan diri sebelum terjun langsung dalam bermasyarakat.

BAHAN DAN METODE

A. Pendekatan penjurusan

Sesuai dengan Pasal 12 ayat 2 Undang –Undang No 20 Tahun 2003 tentang Sistem Pendidikan yang menyatakan bahwa Setiap Peserta didik berhak mendapatkan pelayanan pendidikan sesuai dengan bakat , minat dan kemampuannya. Pasal ini menjelaskan bahwa setiap siswa sekolah, khususnya sekolah menengah atas ataupun sederajat mempunyai hak dalam menentukan pilihan pembelajaran sesusai dengan optensi, minat dan bakat yang dimiliki oleh para siswa-siswi. Pihak sekoah wajib menyediakan jurusan atau peminatan agar siswa- siswi dapat melanjutkan pengembangan potensi yang dimiliki oleh masing – masing individu.

Proses pemilihan pemintaan pada dilakukan pada saat kenaikan kelas dari kelas X ke Kelas XI. Proses peminatan yang dilakukan berdasarakan nilai evaluasi dari setiap mata pelajaran peminatan, selain nilai evaluasi dari setiap mata pelajaran peminatan sebagai bahan pertimbangan juga dilihat dari minat dari siswa/i yang bersangkutan.

B. Konsep *Clustering* Pada *Data Mining*

Proses klastering pada data mining menurut Hermawati(2009a:16) "Proses mempartisi data-set menjadi beberapa sub-set ataukelompok sedemikian

rupa sehingga elemen-elemen dari suatu kelompok tertentu memiliki set properti yang di*share* bersama".

Analisa klastering Hermawati(2009b:123) merupakan analisa yang dilakukan pada sekumpulan objek sehingga objek – objek dalam satu kelompok sama dengan yang lain dan berbeda dengan objek dalam kelompok yang lainnya. Tujuan dari melakukan analisa klaster ialah meminimalkan jarak didalam klaster dan memaksimalkan jarak antara klaster. Tipe-tipe dari Clustering sebagai berikut:

1. *Well –Separated Cluster*
2. *Center Based Cluster*
3. *Contiguous Cluster (nearest neighbor atau Trasnsitive*
4. *Density Based*

C. Algoritma Fuzzy C-Means

Fuzzy C-means merupakan suatu teknik peng-klasteran data yang keberadaan setiap titik data dalam suatu klaster ditentukan oleh derajat keanggotaanya (Purnomo dan Kususmadewi:2010). Konsep dasar FCM yaitu menentukan pusat klaster, yang akan menandai lokasi rata-rata untuk setiap klaster. Dengan caramemperbaiki pusat klaster akan bergerak menuju lokasi yang tepat. Perulangan ini didasarkan pada minimasi fungsi obyektif yang menggambarkan jarak dari titik data yang diberikan kepusat klaster yang berbobot oleh derajat keanggotaan titik data tersebut.

Algoritma FCM didasarkan pada minimasi fungsi objektif yang diformulasikan dalam persamaan:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C}(u_{ij})^m ||z_i - c_j||^2$$

Dimana:

Uij = Level keanggotaan dari Zi dalam cluster j.

Zi = Nilai data ke-I dari d- dimensi data.

Ci = Nilai ke-j dari d-dimensi cluster center

m = Sembarang bilangan real lebih besar dari 1

Tahapan penyelesaian Algoritma FCM disusun dengan langkah sebagai berikut:

Langkah 1: Tentukan himpunan data Z. tentukan jumlah *cluster* yang diharapkan 1< c< N, nilsi pembobot m>1, toleransi penghentian € >0.

Langkah 2: Inisialisasi matriks partisi secara acak, U(0) € M_{fc}

Ulangi untuk 1=3,4,5

Langkah 3: Hitung *cluster center* (*means*).

$$v_i^{(l)} = \frac{\sum_{k=1}^{N}\left(\mu_{ik}^{(l-1)}\right)^m z_k}{\sum_{k=1}^{N}\left(\mu_{ik}^{(l-1)}\right)^m}, , 1 \le i \le c$$

Langkah 4: Hitung jarak.

Langkah 5: Perbaharui matriks partisi

Untuk 1≤k≤N

Jika $D_{ikA} > 0$ untuk semua i=1,2,3,…….,c.

$$\mu_{ik}^{(l)} = \frac{1}{\sum_j^c \left(D_{ikA} \, \overline{D}_{jkA} \right)^{\frac{2}{(m-1)}}}$$

Atau dengan kata lain:

$$\mu_{ik}^{(l)} = 0 \; jika \; D_{ikA} > 0 \; dan \; \mu_{ik}^{(l)} \epsilon [0,1]$$

$$dengan \sum_{i=1}^{c} \mu_{ik}^{(l)} = 1$$

Ulangi sampai dengan ||U(l)-U(l-1)||<Ɛ.

D. Metode Penelitian Eksperimental

Penelitian yang peneliti lakukan merupakan penelitian jenis eksperimental, dimana menurut F. Hair, Jr. , Rolph, E. A., Romald, L.T. dan G.B William (2008) dalam Sangadji etta dan Sopian (2010). Penelitan eksperimental (*Experimental Research*) merupakan penelitian yang subjeknya diberi perlakuan(*treatment*) lalu diukur akibat perlakuan pada diri subjek. tujuan dari penelilian ekperimental untukmelihat pengaruh variable independen terhadap variable dependen.

Peneliti dalam penelitian ini membahas tentang bagaimana algoritma Fuzzy C-Means (FCM) mampu memberikan rekomendasi pengelompokan atau penentuan jurusan yang optimal bagi para siswa-siswi sekolah dan juga bagaimana Algoritma Gentika mendistribusian $D_{ikA}^2 = [(z]_k v_i^{(l)})^T A \left(z_k - v_i^{(l)}\right), 1 \le i \le c, 1 \le k \le N$ siswa kedalam kelas berdaskan kuota yang tersedia. Proses pengelompokan atau penentuan jurusan ini didapat berdasarkan evaluasi nilai setiap mata pelajaran jurusan dan berdasarkan hasil evaluasi nilai permata pelajaran jurusan.

E. Ekperimen

Data populiasi yang peneliti ambil mepakan data populasi sebaran nilai hasil evaluasi siswa-siswi kelas- X semester genap pada tahun 2013. Jumlah populasi didalam data hasil evaluasi nilai siswa-siswi berjumlah 280 Orang, dari jumlah tersebut akan ambil sampel sejumlah 155 Orang. Jumlah ini diperoleh berdasarkan table yang dikembangkan dari *Isaac* dan *Michael* dengan tingkat kesalahan, 5%.

HASIL DAN PEMBAHASAN

A. Kriteria Penjurusan

Pada proses pemilihan jurusan perlu diperhatikan bahwa pemilihan jurusan berdasarkan beberapa criteria sebagai berikut:

1. Pemilihan jurusan IPA berdasarkan hasil evaluasi nilai matakuliah yang menjadi ciri dari jursan IPA yaitu , MATEMATIKA, FISIKA, KIMIA, BIOLOGI.

2. Pemilihan jurussan IPS berdasarkan hasil evaluasi nilai matapelajarn yang menjadi ciri dari jurusan IPS yaitu, GEOGRAFI, SOSIOLOGI, SEJARAH, EKONOMI.

3. Pemilihan jurusan IPB berdasarkan hasil evaluasi nilai matapelajaran yang menjadi cirri dari jurusan IPB yaitu, BAHASA INGGRIS, BAHASA INDONESIA, BAHASA JERMAN.

4. Penentuan jurusan berdasarkan nilai rata-rata dari tiap mata pelajaran yang menjadi ciri dari tiap-tiap jurusan

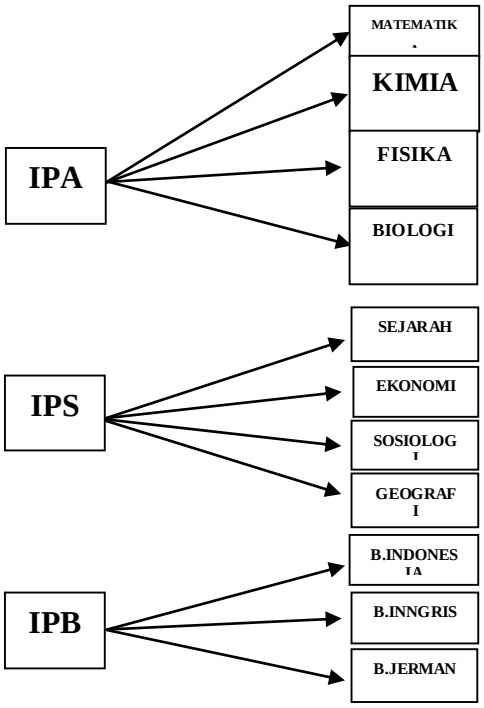
5. Penentuan jurusan juga berdasarkan nilai rata-rata keseluruhan mata pelajaran, apa bila nilai rata-rata kesluruhan >80, maka siswa boleh memilih jurusanyang dikehendaki, jika nilai rata-rata diantara 70-79, maka siswa boleh memilih jurusan hanya IPS dan IPB saja dan apabila nilai rata keseluruhan diantara 0 – 69 hanya dapat memilih jurusan IPB saja.

B. Penerapan Fuzzy C-Means Untuk Penjurusan

Proses penentuan jurusan yang diterapkan menggunakan data hasil perhitungan rata-rata nilai mata pelajaran perjurusan dan rata- rata nilai secara keseluruhan. Nilai hasil evaluasi ini didapat setelah data dirubah menjadi data nilai dengan nilai grade atau pembobotan.

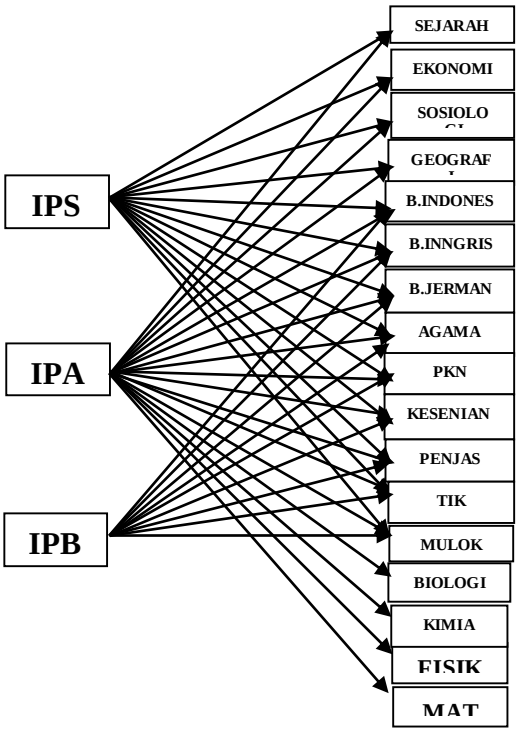
Setelah dibentuk nilai berbentuk grade/bobot maka dihitung nilai rata-rata dari tiap mata pelajaran sesuai dengan gambar korelasi mata pelajaran dibawah ini.

1. Korelasi jurusan dengan mata pelajaran perjurusan



Gambar 1. Korelasi Jurusan dengan Nilai Mata Pelajaran Jurusan

2. Korelasi jurusan dengan mata pelajaran keseluruhan



Gambar 2. Korelasi Jurusan dengan Seluruh Nilai Mata Pelajaran Jurusan

Dengan hasil pengkonversian nilai menjadi grade dan setelah dilakukan perhitungan rata – rata setiap mata pelajaran jurusan dan keseluruhan untuk melakukan pemilihan jurusan dengan Fuzzy C-means(FCM), penentuan penjurusan yang dilakukan menggunakan Fuzzy C-Mean, sebagai berikut:

1. Menetapkan Matriks awal dari U, berupa matrik berukuran n x m (n adalah jumlah sampel=155) dan m adalah atribut/parameter setiap data (3).X_{ij} = data sampel ke-i(i=1,2,3…..n), atribut ke-j(j=1,2,3…….,m).

2. Tentukan parameter awal sebagai berikut:

   a. Jumlah cluster
      = c = 3

   b. Pangkat
      =w = 2

   c. Maximum Iterasi
      = Maxlte = 10

   d. Error terkecil yang diharapkan= ε_0
      = 10-5

   e. Fungsi onjektif awal
      =P_0 =0

   f. Iterasi Awal
      = t =1

3. Bangkitkan bilangan random µik, i=1,2,...,n; k=1,2,...c; sebagai elemen-elemen matriks partisi awal U.
4. Menentukan Pusat Cluster (V)

   Pada bagian ini hasil sebaran dari matrik U, akan dihitung pusat cluster dari tiap-tiap cluster yang ada dengan menggunakan persamaan:

$$v_i^{(l)} = \frac{\sum_{k=1}^{N}\left(\mu_{ik}^{(l-1)}\right)^m z_k}{\sum_{k=1}^{N}\left(\mu_{ik}^{(l-1)}\right)^m}$$

5. Menghitung Fungsi Objektif Pada Iterasi Ke-1

   Perhitungan Fungsi Objektif dilakukan dengan menggunakan persamaan

$$P_t = \sum_{i=1}^{n}\sum_{k=1}^{c}\left(\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right]\left(\mu_{ik}\right)^w\right)$$

Periksa kondisi berhenti dengan persamaan, |Pt-Pt-1|<Ɛ, t>Maxtler,jika kondisi ini belum terpenuhi maka terus dilakukan iterasi sampai kondisi ini terpenuhi. Karena nilai fungsi objektif yang di dapat P_1=6270,4011 artinya kondisi|Pt-Pt-

1|<Ɛ belum tepenuhi maka dilanjutkan ke iterasi selanjutnya sampai kondisi |Pt-Pt-1|<Ɛ terpenuhi.

Berdasarkan hasil penerapan fuzzy C-Means utnuk penentuan jurusan yang ada pada SMAN 84 Jakarta, kondisi fungsi obyektif terpenuhi pada iterasi ke-4 dengan hasil sebaran derajat keanggotaan pada tiap-tiap cluster (jurusan) sebagai berikut:

1. Kelompok pertama (*cluster*-1) / Jurusan IPB akan beranggotakan siswa dengan no urut:

   1, 2, 5, 11, 16, 18, 19, 21, 24, 25, 29, 30, 34, 35, 39, 43, 48, 49,51, 52, 58, 60, 63, 64, 66, 69, 70, 71, 72, 74, 75, 81, 82, 84, 86, 88, 92, 93, 94, 95, 98, 99, 100, 103, 117, 119, 120, 122, 128, 130, 131, 132, 134, 139, 140, 146, 147, 149, 150, 155.

2. Kelompok Kedua (*cluster*-2) / Jurusan IPS akan beranggotakan siswa dengan no urut:

   3, 7, 8, 9, 12, 14, 20, 22, 23, 26, 27, 36, 37, 38, 40, 41, 42, 44, 46, 47, 53, 54, 56, 59, 61, 63, 65, 67, 73, 77, 79, 80, 83, 87, 90, 96, 103, 108, 109, 111, 113, 121, 123, 124, 126, 127, 129, 133, 136, 141, 151

3. Kelompok Ketiga (*cluster*-3) / Jurusan IPA akan beranggotakan siswa denganno urut:

   4, 6, 10, 13, 15, 17, 28, 31, 32, 33, 45, 50, 55, 57, 68, 76, 78, 85, 89, 91, 97,101, 104, 105, 106, 107, 110, 112, 114, 115, 118, 125, 135, 137, 138, 142, 143, 144, 145,148, 152, 153, 154

KESIMPULAN

Penelitian ini dibuat untuk mengolah data nilai hasil evaluasi akademik siswa yang diperuntukan untuk menentukan pemilihan jurusan pada SMAN 84 Jakarta, dimana penentuan jurusan ini harus sesuai dengan hasil evaluasi nilai dan minat dari siswa serta pembagian kelas sesuai dengan kuota kelas jurusanyang tersedia.. Untuk menyelesaikan permasalahan ini penulis menggunakan Fuzy C-Means untuk memnentukan jurusan dan untuk mendistribusikan siswa jurusan sesuai dengan kuota kelas yang ada penulis mengguankan algoritma genetika. Kesimpulan dari penelitian ini adalah:

1. Dengan menggunakan metode FCM pihak sekolah bisa menempatkan siswa dengan jurusan yang lebih akurat.

2. Metode FCM memperoleh tingkat akurasi sebesar 72,26 % dalam penerapan penentuan jurusan pada SMA 84 Jakarta.

3. Metode Algoritma dapat mendistriusikan siswa jurusan kedalam kelas dengan perbandingan yang seimbang sesuai dengan kuota kelas yang disediakan.

UCAPAN TERIMA KASIH

Ucapan terima kasih ingin saya sampaikan kepada Allah SWT, kepada Orang Tua dan Istri Tercinta yang senantiasa mendukung baik

materil maupun moril, serta kepada Bapak Dr. Alimuddin M.Kom, selaku dosen pembimbing Thensis di Program Pasca Sarjana STMIK Nusa Mandiri

DAFTAR PUSTAKA

Asri Yusri (2013). Aplikasi Penentuan Jurusan Mahasiswa baru Jalur PMB Dengan Algoritma Fuzzy C-Mean.Jurnal Informatika, Sains &Teknologi. Bekasi,2-10.

Luthfi Emha (2007). *Fuzzy C-Means untuk Clustering Data*. Seminar Nasional Teknologi.Jogjakarta,D-1-D7.

Munandar Ai, Widyarto Wahyu, Harsiti (2013). Clustering Data Nilai Mahasiswa Untuk Pengelompokan kosentrasi Jurusan Menggunakan *Fuzzy Cluster Means*. Seminar Nasional Aplikasi Teknologi Informasi (SNATI).Jogjakarta, G-30 – G-33

Novamizanti Ledya dan Vimalakirti Retno(2010). Optimasi Logika Fuzzy Menggunakan Algoritma Genetika Pada Identifikasi Pola Tanda Tangan. Konferensi Nasional Sistem dan Informatika.Bali, 151-157.

Sayekti Ratna, Hidayat Nurul,Soebroto Arif (2013).Implemetasi Algoritma Fuzzy C-MEANS untuk Pembangkitan Aturan Fuzzy Pada Pengelompokan Tingkat Risiko Penyakit Kanker Payudara.Jakarta.

Simbolon Cary, Kusumastuti Nilamsari, Irawan Beni (2013). Clustering Lulusan Mahasiswa Matematika FMIPA UNTAN PONTIANAK Menggunakan Algoritma Fuzzy C-Means. Buletin Ilmiah Mat.Stat dan Terapannya (Bimaster). Volume 2, 21-26.

Varghese Bindiya, dkk. Clustering Student Data to Characterize Performance Patterns. International Journal of Advance Computer Science and Application.138-140.

Widyastuti Naniek, Amir Hamzah (2007). Penggunaan Algoritma Genetika alam Peningkatan Kinerja Fuzzy Clustering Untuk Pengenalan Pola. Berkala MIPA(2).Yogyakarta,1-14.

Cluster Analysis Menggunakan Algoritma *Fuzzy C-means* dan *K-means* Untuk Klasterisasi dan Pemetaan Lahan Pertanian di Minahasa Tenggara

Jemaictry Tamaela[1], Eko Sediyono[2], Adi Setiawan[3]
[1,2]Magister SistemInformasiUniversitas Kristen SatyaWacana
[3]Fakultas Sain dan Matematika, Universitas Kristen Satya Wacana
Jl. Diponegoro 52-60 Salatiga, 50711
email: [1]jeyzone77@gmail.com; [2]ekosed1@yahoo.com; [3]adi_setia_03@yahoo.com

Masuk: 13 Mei 2017; Direvisi: 12 Juni 2017; Diterima: 21 Juni 2017

Abstract. *The purpose of this study is to perform cluster analysis and implementation by utilizing fuzzyc-means (FCM) and k-means (KM) to process agricultural data based on the data mining results. The fuzzy c-means (FCM) and k-means (KM) are implemented to find out and form the agricultural land clusters which appropriate the commodity types based on the supporting attributes that are used. The analysis and implementation results could provide some land information such as the number of the clusters, the land areas, the region areas, the locations and the productivity levels. The results of this study could be applied as the suggestion in converting the land functions and structuring the agricultural lands. The utilization of Openstreetmap is an open source solution which is implemented in the application. It could give visual information related to the agricultural land regions based on the clusters which make it easier to comprehend.*
Keywords: *Cluster analysis, C-means, K-means, GIS, Data mining*

Abstrak. *Penelitian ini bertujuan untuk melakukan analisis cluster dan implementasinya dengan menggunakan algoritma fuzzy c-means (FCM) dan k-means (KM) untuk mengelola data pertanian dari hasil data mining yang dilakukan. Fuzzy c-means (FCM) dan k-means (KM) dimplementasikan untuk menemukan dan membentuk klaster-klaster daerah lahan pertanian sesuai dengan jenis komoditi berdasarkan atribut-atribut pendukung yang digunakan. Hasil analisis dan implementasi dapat menyediakan informasi lahan seperti jumlah kluster, luas lahan, luas daerah, letak dan tingkat produktifitas. Hasil yang diperoleh dapat menjadi bahan masukan dalam proses alih fungsi dan penataan lahan pertanian. Penggunaan Openstreetmap merupakan solusi open source yang diimplementasikan pada aplikasi dapat memberikan informasi visual daerah-daerah lahan pertanian berdasarkan klaster yang dihasilkan sehingga lebih mudah untuk dipahami.*
Keywords: *Cluster analysis, c-means, k-means, GIS, Data mining*

1. Pendahuluan

Daerah Minahasa Tenggara (MiTra) merupkan salah satu kabupaten di Sulawesi Utara yang baru dimekarkan pada tahun 2010. Luas daerah MiTra 730,62 km^2 dan didominasi daerah yang bergunung dengan kesuburan tanah yang tinggi. MiTra menjadikan pertanian sebagai salah satu sektor unggulan dalam menunjang perekonomian daerahnya. Sebagai salah satu daerah pertanian di wilayah Sulawesi Utara, lahan-lahan pertanian terutama untuk sektor tanaman pangan perlu diperhatikan sehingga dapat menunjang kebutuhan pangan dan ekonomi daerah.

Analisis dan pengelolaan lahan pangan perlu dilakukan sehingga dapat memaksimalkan fungsi dan alokasi lahan yang ada untuk memperoleh hasil yang maksimal. Sistem informasi geografis (GIS) dapat memberikan keuntungan yaitu gambaran lahan pangan dari sudut yang

berbeda, panduan pengolahan lahan pangan, dan informasi pendukung keputusan yang dapat digunakan untuk tindakan yang akan dilakukan (Kisilevich, dkk., 2013).

Salah satu cara yang dapat dilakukan untuk mengelola data geografis lahan adalah dengan melakukan *clustering* menggunakan algoritma *fuzzy c-means* (FCM) dan *k-means* (KM). *Fuzzy clustering* melakukan pembobotan terhadap tingkat keanggotaan himpunan *fuzzy*-nya (Bezdek & Dunn, 1975). FCM merupakan metode pengelompokan yang memungkinkan satu bagian data memiliki dua atau lebih kelompok (Dunn, 1973). Dengan FCM, objek akan cenderung dibagi menjadi anggota suatu kelompok sesuai derajat keanggotaannya (Jansen, 2007). *K-means* (KM) dikembangkan oleh McQueen pada 1967. KM merupakan algoritma *clustering* yang populer dan bekerja dengan cara membagi data ke dalam partisi data (Juanying, dkk., 2011).

Analisis *clustering* dengan menggunakan algoritma FCM dan KM dapat digunakan untuk mempermudah proses klasterisasi lahan pertanian berdasarkan jenis dan atribut yang telah ditentukan. Selain itu, hasil analisis dan implementasi dapat menjadi bahan pertimbangan dalam proses analisis dan pengambilan keputusan terkait alih fungsi dan alokasi lahan diwaktu yang akan datang. Penggunaan Algoritma FCM dan KM bertujuan untuk membandingkan variasi hasil kluster yang dihasilkan kedua algoritma dalam implementasinya pada aplikasi yang dibuat berdasarkan pada banyaknya jumlah data yang tersedia.

Berdasarakan latar belakang yang telah dijelaskan, maka dilakukan penelitian mengenai *cluster analysis* dan implementasinya menggunakan algoritma FCM dan KM untuk klasterisasi dan pemetaan lahan pertanian untuk lahan pertanian di Minahasa Tenggara. Implementasi dilakukan dalam bentuk aplikasi desktop.

2. Tinjauan Pustaka

2.1. Penelitian Terdahulu

Beberapa penelitian terdahulu berkaitan dengan analisis *clustering* dengan algoritma FCM dan KM diantaranya dilakukan Ali, dkk. (2008) melakukan review dan analisis algoritma-algoritma *clustring* yang efektif dalam Segmentasi citra terutama pada citra berbasis *fuzzy*. Vendruscolo & Kaleita (2011) menerapkan FCM untuk menghitung data spasial pertanian dengan tujuan mengetahui jumlah klaster optimal dari tanaman yang homogen, proses ini divalidasi menggunakan dua atribut yaitu *Partition Coefficient (PC)* dan *Classification Entropy(CE) and Dunn's Index (DI).* Meskipun hanya menggunakan dua atribut tetapi telah menunjukan hasil yang menjanjikan. Martino & Sessa (2009) menerapakan *extended fuzzy c-means* yang merupakan alternatif dari FCM ke dalam GIS untuk menentukan daerah-daerah titik panas. Afifah, dkk. (2016) menerapkan algoritma FCM untuk pengklasteran lahan sawah di Indonesia sebagai evaluasi ketersediaan produksi pangan. Hardiani, dkk. (2014) menggunakan algoritma FCM dan RFM dalam proses *data mining* untuk *customer relationship management* dalam kajian *data mining customer relationship management* pada lembaga keuangan mikro. Rodrigues & Cora (2015) menerapkan algoritma FCM dalam melakukan identifikasi manajemen lahan berdasarkan variabilitas tanah dan lahan jagung.

Miyamoto (2011) melakukan penelitian tentang perbedaan fungsi-fungsi objek dalam algoritma FCM dan *Kernel-Based Clustering.* Ghosh & Dubey (2013) melakukan analisis dan perbandingan performa algoritma KM dan FCM dalam implementasinya pada perangkat lunak. Sankoh, dkk. (2015) menerapkan kombinasi algoritma *Neural Network,* KM dan *Particle Swarm Optimization* untuk melakukan klasterisasi jenis musik. Ong (2013) menerapkan algoritma KM dalam penelitiannya untuk menentukan strategi *marketing* President University. Dalam penelitiannya menggunakan tiga atribut yaitu asal, jurusan dan GPA. Tujuan dan hasil penelitian ini digunakan untuk membantu tim promosi dalam memprediksi strategi promosi yang akan dilakukan oleh President University di Indonesia.

Selain dengan algoritma FCM dan KM, dilakukan proses *data mining* untuk mengelola data-data geografis MiTra yang akan digunakan sebagai basis data pada proses klusterisasi. Lanzi, dkk. (2012) dalam penelitian *Data Mining in GIS: A Novel Context-Based Fuzzy Geographically Weighted Clustering Algorithm* menyebutkan sistem informasi geografis (GIS)

memegang peranan penting dalam berbagai penelitian dan industri. Penelitian dilakukan untuk menemukan aspek-aspek penting *data mining* dalam GIS dengan menggunakan GIS *clustering* dengan analisis *Geo-Demographic*. Penelitian yang dilakukan Kaur, dkk. (2014) dengan menerapkan algortima KM dan *K-Nearest Neighbor* (KNN) dalam *data mining* pertanian untuk memprediksi harga tanaman. Peng, dkk. (2011) dalam penelitian *An incident information management framework based on data integration, data mining, and and multi-criteria decision making*, menyebutkan pentingnya integrasi data suatu kejadian yang sejenis dalam satu format yang sama. Kemudian dengan menggunakan teknik *data mining* dapat diidentifikasi pelayanan yang sesuai untuk keadaan sebelum dan sesudah kejadian dalam sistem manajemen informasinya.

Berdasarkan tinjauan pustaka yang ada maka dalam penelitian ini akan dilakukan *cluster analysis* dengan menggunakan algoritma *fuzzy c-means* (FCM) dan *k-means* (KM) untuk proses klasterisasi dan pemetaan lahan pertanian di kabupaten Minahasa Tenggara.

2.2. Landasan Teori

Cluster analysis merupakan pembelajaran tanpa pengawasan. *Clustering* bekerja dengan menemukan struktur dalam koleksi data yang tidak di label dan mengelompokannya ke dalam kelompok yang saling memiliki kesamaan (Madhulatha, 2012). FCM merupakan metode pengelompokan yang memungkin satu data memiliki dua atau lebih kelompok (Dunn, 1973) dan diperbaiki oleh bezdek tahun 1981. Langkah-langkah dalam FCM dapat dilihat pada Persamaan (1). Untuk setiap nilai m merupakan bilangan asli lebih dari 1, u_{ij} adalah derajat keanggotaan dari x_i di klaster j, c_j merupakan pusat dimensi klaster. Untuk memaksimalkan hasil klaster, dilakukan perulangan dengan memperbarui keanggotaan u_{ij} dan pusat-pusat klaster c_j dengan Persamaan (2). Proses perulangan ini berhenti ketika keanggotan dan pusat-pusat klaster tidak lagi berubah atau berpindah dengan Persamaan (3).

$$J_m = \sum_{i-1}^{N}\sum_{j-1}^{C} u_{ij}^m \|x_i - c_j\|^2 \tag{1}$$

$$u_{ij} = \frac{1}{\sum_{k-1}^{C}\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}} \qquad c_j = \frac{\sum_{i-1}^{N} u_{ij}^m \cdot x_i}{\sum_{i-1}^{N} u_{ij}^m} \tag{2}$$

$$\max_{ij} = \left\{\left|u_{ij}^{(k+1)} - u_{ij}^{(k)}\right|\right\} < \varepsilon \tag{3}$$

Algoritma *K-Means* (KM), dikembangkan oleh MacQueen tahun 1967, merupakan salah satu algoritma sederhana dalam memecahkan masalah pengelompokan. MacQueen menjelaskan dalam algoritma k-means, dataset dibagi ke dalam beberapa klaster, dan satu data hanya dapat menjadi satu anggota klaster tertentu (MacQueen, 1967). *K-means* menentukan terlebih dahulu mean atau titik tengah kluster berdasarkan jumlah kluster yang dibuat, kemudian *k-means* membagi data set ke dalam masing-masing kluster berdasarkan nilai yang sama atau yang mendekati nilai tengah kluster (Zijun, 2012). Algoritma ini bertujuan untuk meminimalisasi suatu fungsi objek. Fungsi objek dalam k-means dapat ditentukan dengan Persamaan (4), dimana $\|x_i^{(j)} - c_j\|^2$ merupakan jarak titik data $x_i^{(j)}$ dan pusat klaster c_j merupakan indikator jarak poin n data dari pusat klaster masing-masing.

$$J = \sum_{j-1}^{k}\sum_{i-i}^{n} \left\|x_i^{(j)} - c_j\right\|^2 \tag{4}$$

Langkah-langkah dalam algoritma k-means: (1) Menentukan titik k awal pada objek yang akan diklaster. (2) Menetapkan setiap objek dalam kelompok yang memiliki nilai objektif terdekat. (3) Ketika semua objek telah ditetapkan, posisi titik k dihitung kembali. (4) Langkah ke-2 dan ke-3 diulangi sampai tidak lagi berubah.

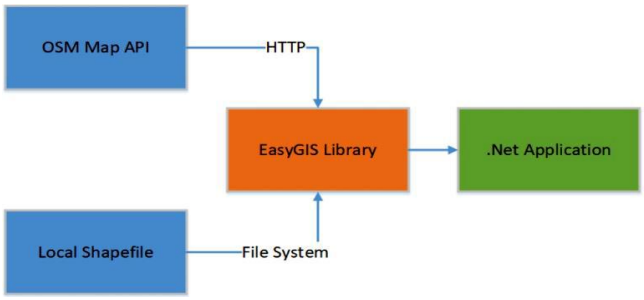
3. Metodologi Penelitian

Penelitian dilakukan melalui beberapa tahap. Tahap pertama yaitu: (1) Tahap studi pendahuluan. Pada tahap ini dilakukan diidentifikasi masalah yang ada. (2) Tahap analisis dan implementasi. Pada tahap ini dilakukan pengumpulan pustaka pendukung dan parameter untuk proses analisis dan implementasi. (3) Tahap kemanfaatan. Pada tahap ini dipelajari manfaat yang dihasilkan dari hasil analisis dan implementasi.

Dalam penelitian ini terdapat beberapa tahapan *data mining* yang dilakukan. Tahapan ini dilakukan untuk memaksimalkan data-data pertanian MiTra yang dimiliki. Tahapan *data mining* yang dilakukan dalam penelitian ini: (1) *Data cleaning*, tahap ini menghapus data gangguan atau yang tidak diperlukan dan mengisi data yang hilang untuk meningkatkan kualitas data (Joshi & Usha, 2015). (2) Data integrasi, tahap ini menggabungkan berbagai sumber data yang saling berhubungan. Modul integrasi data yang heterogen dan terintegrasi harus disajikan dalam format yang seragam (Peng, dkk., 2011). (3) Pemilihan data, tahap memilih data yang relevan dengan dengan tujuan penelitian. (4) Transformasi data, tahap ini mengubah data menjadi format baru untuk diproses dalam *data mining*. (5) *Data mining*, tahap ini menerapkan metode cerdas untuk ekstraksi pola. Pada tahap ini digunakan algoritma FCM dan KM yang di implementasikan dalam aplikasi. (6) Pola presentasi, tahap ini untuk presentasikan pola yang diperoleh.

4. Hasil dan Pembahasan

Aplikasi yang dihasilkan dibuat sendiri dengan teknologi *.NET framework* 4.6 menggunakan bahasa pemrograman C# dan menggunakan basis data MySQL. Teknologi . *NET framework 4.6* digunakan untuk mempermudah pengembangan aplikasi dan sebagai penerjemah agar aplikasi dapat dijalankan oleh sistem. Aplikasi diintegrasikan dengan *OpenStreetMap* (OSM) API. OSM API disediakan *OpenStreetMap* dalam format *javascript* sehingga pengguna dapat melakukan kostumisasi pada peta yang digunakan dan menyediakan akses langsung ke server *OpenStreetMap*. Integrasi OSM API dengan aplikasi yang dibuat dapat dilihat pada Gambar 1.



Gambar 1. Integrasi OSM API dengan aplikasi

Aplikasi yang dihasilkan digunakan untuk mengklasterkan lahan pertanian untuk empat komoditi pangan yaitu padi sawah, padi ladang, jagung dan ubi kayu. Hasil *data mining* dari data-data pertanian MiTra digunakan sebagai basis data untuk proses *mining* dan klasterisasi melalui aplikasi. Tabel data hasil *mining* dari data-data MiTra dapat dilihat pada Gambar 2 dan Gambar 3.

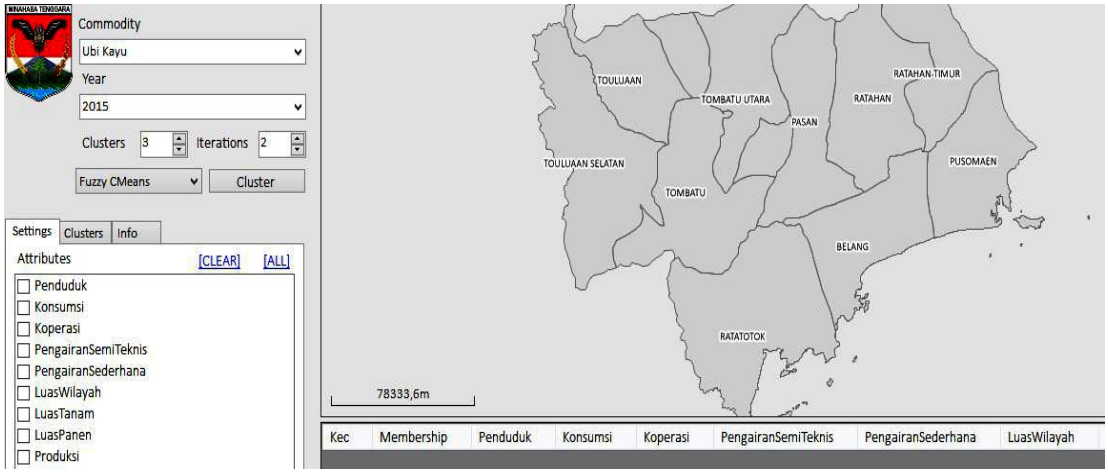
Wilayah	Luas Wilayah (Ha)	Penduduk	Produksi Padi Sawah			Produksi Padi Ladang			Model Pengairan Sawah	
			Luas Tanam (Ha)	Luas Panen (Ha)	Produksi (Ton)	Luas Tanam (Ha)	Luas Panen (Ha)	Produksi (Ton)	Semi Teknis	Sederhana
Ratatotok	10.418	12.696	18	15	69	117	102	317,3	0	9
Posumean	5.362	8.392	352	347	1.792	15	12	37,4	96	150
Belang	7.517	16.290	651	632	3.318	45	42	131	115	130
Ratahan	6.163	12.981	780	753	3.773	15	12	38	276	111
Pasan	4.979	6.546	527	506	2.631	25	22	39	53	120
Ratahan Timur	6.399	5.763	146	138	653	50	48	149	0	1
Tombatu	6.795	8.870	630	604	3.068	110	103	321	184	142
Tombatu Timur	1.881	8.690	852	802	3.984	30	25	78	317	273
Tombatu Utara	3.717	7.773	990	974	4.798	0	0	0	0	366
Touluaan	10.180	6.420	375	356	1.738	0	0	0	0	275
Touluaan Selatan	5.276	3.910	50	42	194	130	125	389	0	222
Silian Raya	4.375	5.487	512	482	2.040	0	0	0	0	241

Gambar 2. Data hasil transformasi

Produksi Jagung			Produksi Ubi Kayu			Koperasi Pemasaran
Luas Tanam (Ha)	Luas Panen (Ha)	Produksi (Ton)	Luas Tanam (Ha)	Luas Panen (Ha)	Produksi (Ton)	
947	920	3.523	25	21	262	0
885	860	3.050	12	8	98	0
1203	1105	3.809	18	15	187	3
852	810	2.904	9	5	61	0
871	842	2.975	12	7	88	0
564	523	1.882	8	5	60	0
387	362	1.282	11	8	98	0
953	915	3.027	15	11	141	4
381	352	1.225	8	5	61	4
320	295	1.023	12	7	88	0
543	512	1.728	15	8	98	0
327	295	1.017	20	12	150	0

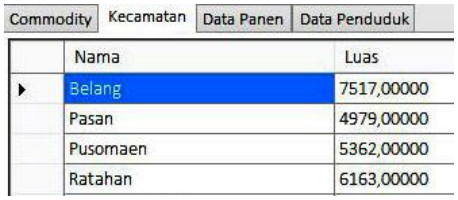
Gambar 3. Data hasil transformasi

Dalam penelitian ini data yang digunakan sebagai data *sample* selain data hasil observasi, data-data diperoleh dari data Indikator Kesejahteraan Rakyat Kabupaten Minahasa Tenggara (Badan Pusat Statistik Kabupaten Minahasa Selatan, 2015a) dan data Kabupaten Minahasa Tenggara Dalam Angka tahun 2015 (Badan Pusat Statistik Kabupaten Minahasa Selatan, 2015b) dan data Kabupaten Minahasa Tenggara Dalam Angka tahun 2016(Badan Pusat Statistik Kabupaten Minahasa Selatan, 2016). Aplikasi dibuat dengan tampilan yang sederhana untuk mempermudah *user* dalam menggunakan aplikasi. Contoh tampilan aplikasi dapat dilihat pada Gambar 4.



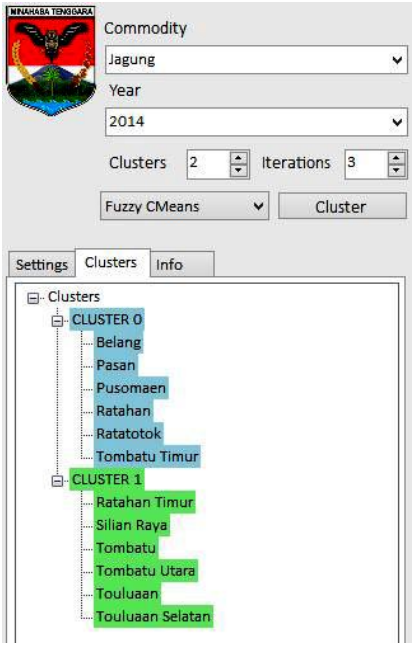
Gambar 4. Tampilan aplikasi

Bagian kiri merupakan fitur utama dari aplikasi yang terdiri dari komoditi, tahun, kluster, iterasi dan algoritma yang diterapkan. *Settings* berisi atribut yang digunakan dalam proses klasterisasi. *Clusters* berisi informasi kluster yang dihasilkan. Info berisi informasi atribut yang digunakan dan nilai masing-masing. Untuk menambah data ke dalam aplikasi, disediakan *master data*. Pada bagian ini, *user* dapat mengisi dan merubah data-data yang diperlukan sesuai dengan jenis datanya. Contoh *master data* dapat dilihat pada Gambar 5.



Gambar 5. *Master Data*

Pengujian aplikasi dilakukan dengan menggunakan data Minahasa Tenggara. Pengujian dilakukan dengan menggunakan data tahun 2014 yang diperoleh dari data Minahasa Tenggara Dalam Angka Tahun 2015 dengan contoh komoditi jagung berdasarkan atribut luas tanam dan luas panen. Proses pengujian pertama menggunakan algoritma FCM. Contoh hasil kluster yang dihasilkan dengan algoritma FCM dapat dilihat pada Gambar 6.

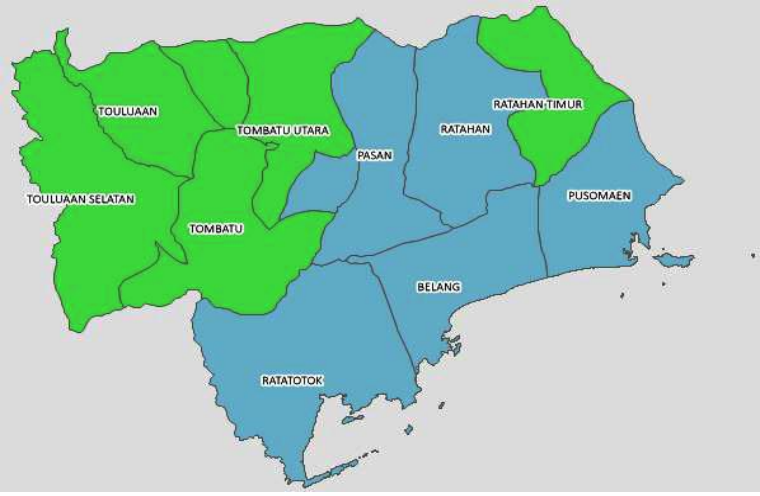


Gambar 6. Hasil kluster FCM

Hasil pada Gambar 6 adalah proses klasterisasi dengan menggunakan dua kluster dengan iterasi sebanyak tiga putaran. Daerah Belang, Pasan, Posumaen, Ratatotok dan Tombatu timur masuk ke dalam klaster nol atau klaster pertama dengan nilai mean klaster untuk luas tanam 793,11 dan luas panen 753,56 ditandai dengan Warna biru. Daerah Ratahan Timur, Silian Raya, Tombatu, Tombatu Utara, Touluaan, Touluaan Selatan merupakan anggota klaster satu atau klaster ke dua dengan nilai *mean* klaster berdasarkan luas tanam 420,33 dan luas panen 389,83 ditandai dengan warna hijau. Contoh informasi hasil kluster yang berisi nilai *membership* dan *mean* masing-masing kecamatan dapat dilihat pada Gambar 7. Dari hasil klusterisasi yang dilakukan, peta hasil kluster yang dihasilkan dapat dilihat pada Gambar 8.

Kec	Membership	LuasTanam	LuasPanen
Ratahan Timur	[0,42] [0,58]	564	523
Silian Raya	[0,35] [0,65]	327	295
Tombatu	[0,28] [0,72]	387	362
Tombatu Utara	[0,29] [0,71]	381	352
Touluaan	[0,35] [0,65]	320	295
Touluaan Selatan	[0,40] [0,60]	543	512
-MEAN-		**420,33**	**389,83**

Gambar 7. Informasi *membership* kluster FCM



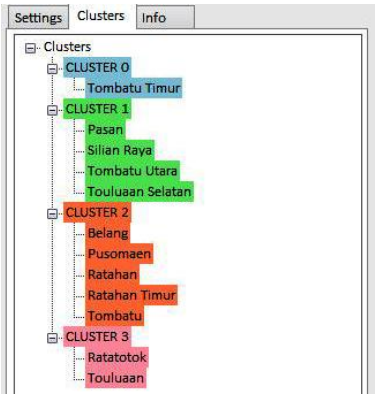
Gambar 8. Peta hasil kluster jagung FCM

Hasil akhir kluster dengan menggunakan algoritma FCM untuk komoditi jagung dengan atribut luas tanam dan luas panen dapat dilihat pada Tabel 1. Dari hasil Tabel 1, karena pada iterasi ke tiga memberikan hasil yang sama dengan iterasi sebelumnya (iterasi ke dua), maka proses iterasi dihentikan pada iterasi ke tiga. Pada percobaan dengan menggunakan tiga kluster, ditemukan adanya kluster kosong atau kluster yang tidak memiliki anggota karena nilai keanggotaan dari masing-masing atribut kecamatan tidak memenuhi kriteria dari klaster tersebut. Dari hasil percobaan dengan menggunkan komoditi jagung berdasarkan luas tanam dan luas panen tahun 2014 maka diperoleh hasil daerah dengan luas lahan pertanian dan panen jagung terbesar berada pada wilayah Belang, Pasan, Posumaen, Ratahan, Ratatotok dan Tombatu Timur, sedangkan wilayah dengan lahan pertanian jagung dan panen terkecil berada di Ratahan Timur, Silian Raya, Tombatu, Tombatu Utara, Touluaan dan Touluaan Selatan.

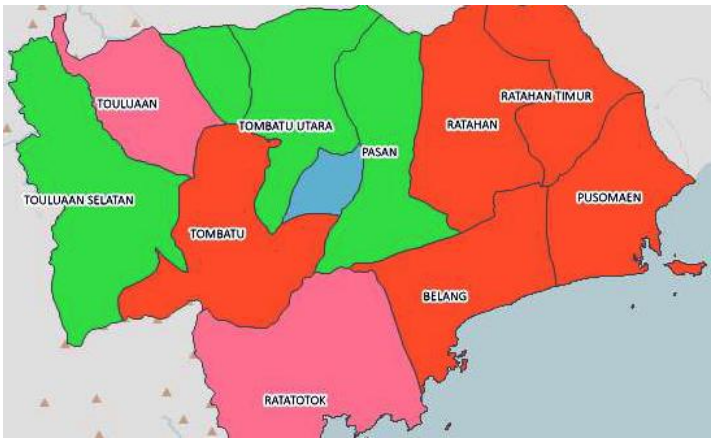
Tabel 1. Hasil kluster FCM

Kecamatan	**Cluster 0**			**Cluster 1**		
	1	2	3	1	2	3
Belang	✓				✓	✓
Pasan	✓				✓	✓
Posumaen	✓				✓	✓
Ratahan	✓				✓	✓
Ratahan Timur	✓	✓	✓			
Ratatotok	✓				✓	✓
Silian Raya		✓	✓	✓		
Tombatu				✓	✓	✓
Tombatu Timur	✓	✓	✓			
Tombatu Utara		✓	✓	✓		
Touluaan	✓	✓	✓			
Touluaan Selatan	✓	✓	✓			

Pengujian dengan menggunakan algoritma KM dengan menggunakan data tahun 2014, menggunakan komoditi jagung dengan atribut yang berbeda dengan pengujian pertama yaitu luas wilayah dan luas tanam. Pengujian ini dilakukan dengan menggunakan empat kluster dengan iterasi sebanyak tiga kali. Hasil kluster dari aplkasi dengan menggunakan algoritma KM menggunakan empat kluster dapat dilihat pada Gambar 9. Peta yang dihasilkan dari hasil klaster dengan algoritma KM dapat dilihat pada Gambar 10.



Gambar 9. Hasil kluster dengan algoritma KM dengan empat kluster



Gambar 10. Peta hasil 4 klaster jagung algoritma KM

Dari peta yang dihasilkan, Tombatu berada dalam kluster sendiri yaitu kluster nol dengan nilai mean kluster berdasarkan luas wilayah 1.881 dan luas tanam 953 ditandai dengan warna biru. Daerah Pasan, Silian Raya, Tombatu Utara dan Touluan berada dalam kluster satu dengan nilai mean kluster berdasarkan luas wilayah 4.568,75 dan luas tanam 530,5 ditandai dengan warna hijau. Daerah Belang, Posumaen, Ratahan, Ratahan Timur dan Tombatu berada dalam kluster dua dengan nilai mean kluster berdasarkan luas wilayah 6.447,2 dan luas tanam 778,2 ditandai dengan warna merah. Daerah Ratatotok dan Touluaan berada dalam kluster tiga dengan nilai mean kluster berdasarkan luas wilayah 10.299 dan luas tanam 633,5 ditandai dengan warna merah muda. Dengan demikian diperoleh hasil kluster lahan pertanian untuk komiditi jagung dengan wilayah terkecil berada di daerah Tombatu dan terbesar berada di daerah Ratatotok dan Touluan. Proses kluster lahan pangan dengan algoritma KM menggunakan empat kluster dengan tiga iterasi dapat dilihat pada Tabel 2.

Tabel 2. Hasil Klaster KM

Kecamatan	Cluster 0			Cluster 1			Cluster 2			Cluster 3		
	1	2	3	1	2	3	1	2	3	1	2	3
Belang								✓	✓			
Pasan				✓	✓	✓						
Posumaen							✓	✓	✓			
Ratahan							✓	✓	✓			
Ratahan Timur							✓	✓	✓			
Ratatotok										✓	✓	✓
Silian Raya				✓	✓	✓						
Tombatu							✓	✓	✓			
Tombatu Timur	✓	✓	✓									
Tombatu Utara				✓	✓	✓						
Touluaan										✓	✓	✓
Touluaan Selatan							✓	✓	✓			

Dari hasil Tabel 2, diketahui karena pada iterasi ketiga memberikan hasil yang sama dengan iterasi sebelumnya (iterasi kedua), maka proses iterasi dihentikan pada iterasi ketiga. Hasil Tabel 2 menunjukan kluster nol merupakan daerah dengan nilai terendah dan kluster tiga merupakan daerah dengan nilai tertinggi.

5. Kesimpulan

Berdasarkan hasil implementasi dan analisis, maka penelitian ini memperoleh kesimpulan, algoritma FCM dan KM yang diimplementasikan pada aplikasi dapat berjalan dengan baik dalam mengklasterkan daerah-daerah lahan pertanian sesuai komoditi berdasarkan pada jenis atribut yang digunakan. Dari berbagai data yang ada, data dalam bentuk angka dipilih untuk digunakan sebagai data pengujian. Hal ini dikarenakan data angka dapat dihitung nilai *mean* atau rata-ratanya. Selain itu pada penelitian FCM dan KM hanya dapat diimplementasikan untuk memproses data dalam bentuk angka.

Pada pengujian, proses iterasi maksimal didapat pada iterasi ketiga karena pada iterasi selanjutnya tidak terjadi perubahan anggota kluster. Hal ini dipengaruhi oleh banyaknya jumlah data yang tersedia. Dengan jumlah data yang terbatas, algoritma KM dapat bekerja lebih baik dibandingkan FCM dengan menghasilkan jumlah kluster yang lebih banyak atau bervariasi.

Aplikasi yang dihasilkan mudah digunakan dalam proses pengisian data, pemilihan komoditi, penentuan banyaknya klaster, jumlah iterasi yang diperlukan, visualisasi dan pemilihan atribut-atribut yang dapat disesuaikan dengan kebutuhan *user*. Hasil yang diperoleh dapat menjadi masukan kepada pemerintah daerah MiTra dalam mengelola dan memaksimalkan potensi lahan pertanian daerah.

Penggunaan *Openstreetmap* merupakan solusi *open source* yang dapat digabungkan dengan aplikasi sehingga dapat memberikan informasi visual daerah-daerah pertanian berdasarkan klaster yang dihasilkan sehingga lebih mudah untuk dipahami. Selain memperoleh data dalam bentuk angka, *user* juga memperoleh gambaran letak daerah berdasarkan peta yang dihasilkan.

6. Saran

Dari keseluruhan pemaparan yang dijelaskan sebelumnya dalam penelitian, dapat disimpulkan bahwa pada penelitian ini dilakukan proses klasterisasi dan pemetaan lahan pertanian pangan daerah MiTra sehingga dapat diketahui daerah-daerah pertanian berdasarkan komoditi dan atributnya. Untuk dapat memaksimalkan hasil penelitian, perlu dilakukan kajian lebih lanjut dengan memperhitungkan kemiringan lahan, struktur tanah, kecocokan komoditi dengan tanah lahan, potensi bencana dan banyaknya daerah aliran sungai (DAS) yang dapat digunakan dalam menunjang lahan pertanian. Hal ini juga perlu dilakukan sehingga alih fungsi lahan menjadi lahan pertanian lain atau pemukiman penduduk dapat dilakukan dengan tepat sehingga daerah yang memiliki potensi lahan pertanian yang tinggi dapat dimanfaatkan dengan maksimal. Kajian ini yang belum dilakukan dalam penelitian ini dikarenakan keterbatasan waktu dan data yang tersedia.

Referensi

Afifah, N., Rini, D.C., Lubab, A. (2016). Pengklasteran Lahan Sawah di Indonesia Sebagai Evaluasi Ketersediaan Produksi Pangan Menggunakan Fuzzy C-Means. *Jurnal Matematika "MANTIK", 2* (1).

Ali, M.A., Karmakar, G.C., & Dooley, L.S. (2008). Review On Fuzzy Clustering Algorithm. *IETECH Journal Of Advance Computations, 2* (3), 169-181.

Badan Pusat Statistik Kabupaten Minahasa Selatan. 2015a. Indikator Kesejahteraan Rakyat Kabupaten Minahasa Tenggara 2015. Badan Pusat Statistik Kabupaten Minahasa Selatan.

Badan Pusat Statistik Kabupaten Minahasa Selatan. 2015b. Minahasa Tenggara Dalam Angka Tahun 2015. Badan Pusat Statistik Kabupaten Minahasa Selatan.

Badan Pusat Statistik Kabupaten Minahasa Selatan. 2016. Minahasa Tenggara Dalam Angka Tahun 2016. Badan Pusat Statistik Kabupaten Minahasa Selatan.

Bezdek, J.C., & Dunn, J.C. (1975). Optimal fuzzy partitions: A heuristic for estimating the parameters in a mixture of normal dustrubutions. *IEEE Transactions on Computers,* 835-838.

Bezdek, J.C. (1981). Pattern Recognition with Fuzzy Objective Function Algorithms, *Plenum Press*, New York.

Dunn, J. C. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics, 3,* 32-57

Ghosh, S., & Dubey, S.K. (2013) Comparative Analysis of K-Means and Fuzzy C-Means Algorithms. *International Journal of Advance Computer Science and Aplications (IJACSA), 4* (4).

Hardiani. T., Sulistyo. S., & Hartanto. R. (2014). Kajian Data Mining Customer Relationship Management Pada Lembaga Keuangan Mikro.S2 Teknik Elektro dan Teknologi Informasi, Universitas Gadjah Mada Yogyakarta. *Jurnal Sistem Informasi Bisnis Undip 3.*

Joshi, S., & Usha, T. (2015). Data Cleaning Using Based Data Mining Technique. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 4* (2).

Juanying, X., Shuai, J., Wexin, X., & Xinbo, G. (2011). An Efficient Global K-means Clustering Algorithm. *Journal of Computer, 6* (2).

Kaur, M., Gulati, H., & Kundra, H. (2014). Data Mining in Agricultur on Crop Price Prediction: Techniques and Application. *International Journal of Computer Applications, 99* (12).

Kisilevich, S., Keim, D., & Rokach, L. (2013). A GIS-based decision support system for hotel room rate estimation and temporal price prediction: The hotel brokers ontext. *Journal Decision Support Syst, 54,* 1119–1133.

Lanzi, P. L., Cuong, B. C., Hung, H. A., & others. (2012). Data Mining in GIS: A Novel Context-Based Fuzzy Geographically Weighted ClusteringAlgorithm. *International Journal of Machine Learning and Computing, 2*(3), 235.

MacQueen, B.J. (1967): "Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability".*

Madhulatha, T.S. (2012). An Overview On Clustering Methods. *IOSR Journal of Engineering, 2* (4), 791-725

Martino, F.D., & Sessa, S. (2009). Implementation of the Extended Fuzzy C-Means Algorithm in Geographic Information System. *Journal of Uncertain System,3* (4), 298-306.

Miyamoto, M. (2011). Different Objective Function Functions in Fuzzy C-means Algorithms and Kernel-Based Clustering. *International Journal of Fuzzy System, 13* (2).

Ong, J.O. (2013). Implementasi algoritma K-means Clustering Untuk Menentukan Strategi Marketing President University. *Jurnal Ilmiah Teknik Industri (JITI) 12* (1), 10-20.

Peng, Y., Zhang, Y., Tang, Y., & Li, S. (2011). An incident information management framework based on data integration, data mining, and multi-criteria decision making. *Decision Support Systems, 51*(2), 316–327.

Rodrigues, E.S., & Cora, J.E. (2015). Managemen Zones Using Fuzzy Clustring Based On Spatial-Temporaal Variability Of Soil And Corn Yield. *Journal of the Brazilian Association of Agricultural Engineering Journal, 3,* 470-483.

Sankoh, A.S., Musthafa, A.R., Rosidi, Imron, M., & Arifin, A.Z. (2015). Klasterisasi Jenis Musik Menggunakan Kombinasi Neural Network, K-Means dan Particle Swarm Optimization. *Journal Buana Informatika, 6* (3).

Vendruscolo, L.G., & Kaleita, A.F. (2011). Modeling Zone Management in Precision Agriculture Through Fuzzy C-means Technique at Spatial Database. *Digital Repository,* Iowa State University.

Zijun, Z. (2012). K-mean algorithm cluster analysis in data mining.

ISSN:2442-5567 | E-ISSN:2443-289X

PENENTUAN PENERIMA BEASISWA DENGAN ALGORITMA *FUZZY C-MEANS* DI UNIVERSITAS MEGOW PAK TULANG BAWANG

Muhardi[1] Nisar[2]

[1]Universitas Megow Pak [2]MTI Institut Informatika & Bisnis Darmajaya
Email: ardyardy68@yahoo.co.id

ABSTRAK

Beasiswa adalah pemberian berupa bantuan keuangan yang diberikan kepada perorangan yang bertujuan untuk digunakan demi keberlangsungan pendidikan yang ditempuh. Penentuan beasiswa dapat dikelompokan berdasarkan kriteria Indeks Prestasi Kumulatif (IPK), tingkat kemiskinan, tanggungan orang tua dan prestasi mahasiswa dalam proses rekruitmen beasiswa. Algoritma *Fuzzy C-Means* merupakan satu algoritma yang mudah dan sering digunakan dalam pengelompokan data karena membuat suatu perkiraan yang efisien dan tidak memerlukan banyak parameter. Beberapa penelitian telah menghasilkan kesimpulan bahwa metode *Fuzzy C-Means* dapat digunakan untuk mengelompokkan data berdasarkan atribut-atribut tertentu. Penerapan algoritma *Fuzzy C-Means* dalam penentuan beasiswa di kelompokan menjadi tiga *cluster* yaitu menerima, dipertimbangkan dan tidak berhak menerima beasiswa , sempel data sebenyak 75 data mahasiswa diperoleh tiga *cluster* berdasarkan nilai rata-rata penentuan beasiswa kemudian setiap *cluster* diklasifikasikan berdasarkan kriteria mana yang lebih diprioritaskan dengan nilai terbesar pada jarak akhir merupakan *cluster* yang menerima beasiswa, sedangkan *cluster* dengan nilai terkecil merupakan *cluster* yang tidak berhak menerima beasiswa.

Kata Kunci: *Beasiswa, clustering, Fuzzy C- Means.*

ABSTRACT

The scholarship is the grant in the form of financial assistance given to an individual who intends to use for the sake of the continuity of education traveled. Determination of the scholarship can be grouped based on the criteria of the cumulative Achievement Index (IPK), the level of poverty, dependent parent and student achievement in the process of recruitment of scholarship. Fuzzy C-Means algorithm is an algorithm that is easily and often used in grouping data because making an estimate of efficient and doesn't require a lot of parameters. Some research has resulted in the conclusion that the Fuzzy C-Means method can be used to classify data based on certain attributes. Application of Fuzzy C-Means algorithm in determination of scholarship in kelompokan into three clusters, namely receiving, considering and not eligible to receive the scholarship, sempel 75 student data sebenyak data retrieved three clusters based on the average value of the determination of a scholarship then each cluster are classified according to criteria which is more prioritized with the greatest value at the end of the cluster is receiving a scholarship, whereas the smallest value cluster is a cluster that is not eligible to receive the scholarship.

Keywords*: scholarship, clustering, Fuzzy C- Means*

ISSN:2442-5567 | E-ISSN:2443-289X

1. PENDAHULUAN

Beasiswa adalah pemberian berupa bantuan keuangan yang diberikan kepada perorangan yang diberikan kapada perorangan yang bertujuan untuk digunakan demi keberlansungan pendidikan yang ditempuh. Pemberian beasiswa merupakan program kerja yang ada di setiap universitas atau perguruan tinggi. Program beasiswa diadakan untuk meringankan beban mahasiswa dalam menempuh masa studi kuliah khususnya dalam masalah biaya. Pemberian beasiswa kepada mahasiswa dilakukan secara selektif sesuai dengan jenis beasiswa yang diadakan. Univeritas Megow Pak Tulang Bawang menyediakan beberpa program beasiswa, sebagai contoh yaitu beasiswa Peningkatan Akademik (PPA), Beasiswa Bantuan Belajar Mahasiswa (BBM) dan lain sebagainya. Indeks prestasi kumulatif, Tingkat Kemiskinan, tanggungan orang tua dan prestasi menjadi kriteria dalam proses rekruitmen beasiswa.

Proses seleksi penerimaan beasiswa secara manual yaitu dengan menginputkan satu persatu data mahasiswa ke dalam file *excel* kemudian melakukan *sorting* data mahasiswa seringkali melakukan beberapa permasalahan, antara lain membutuhkan waktu yang lama dan ketelitian yang tinggi. Selain itu, transparansi serta ketidak jelasan metodologi yang digunakan dalam proses komputasi penerimaan beasiswa juga menjadi salah satu permasalahan, sehingga dibutuhkan suatu sistem yang dapat membantu dalam proses pengambilan keputusan siapa saja mahasiswa yang menerima beasiswa berdasarkan kriteria-kriteria yang telah ditentukan secara capat dan tepat sasaran.

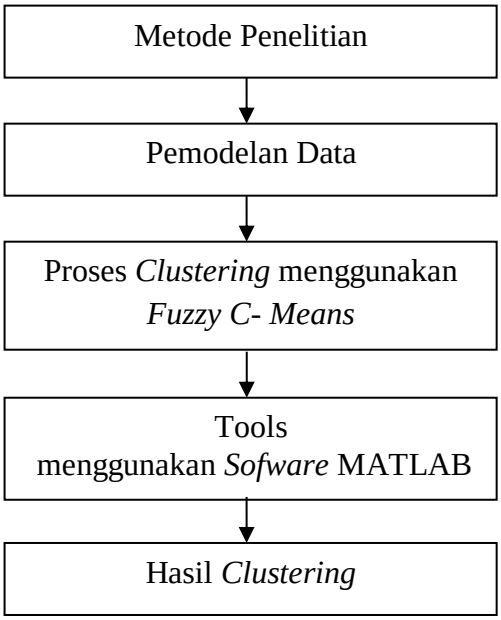
Salah satu metode yang dapat digunakan untuk menyelesaikan permasalahan tersebut adalah metode *Fuzzy C-Means.* Algoritma *Fuzzy C-Means* merupakan satu algoritma yang mudah dan sering digunakan dalam pengelompokan data kerana membuat suatu perkiraan yang efsien dan tidak memerlukan banyak parameter. Beberapa penelitian telah menghasilkan kesimpulan bahwa metode *Fuzzy CMeans* dapat digunakan untuk mengelompokkan data berdasarkan atribut-atribut tertentu. Pada kasus penelitian ini akan menganalisis penerapan metode *Fuzzy C-Means* untuk mengelompokkan data mahasiswa berdasarkan kemampuan mahasiswa dibidang akademik untuk proses penentuan beasiswa.
Pada penelitian sebelumnya, Bahari (2011) melakukan Penerapan metode *Fuzzy C-Means* untuk mengelompokan data siswa dalam penentuan jurusan di Sekolah Menengah Atas pada 81 sampel data siswa yang diuji dalam penelitian menunjukan bahwa metode *Fuzzy C-Means* memiliki tingkat akurasi yang lebih tinggi (rata-rata 78,39%), jika dibandingkan dengan metode penentuan jurusan secara manual yang di lakukan (hanya memiliki tingkat akurasi rata-rata 56,17 %).

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitan

Penelitian ini akan dilaksanakan berdasarkan rancangan penelitian seperti yang di tunjukan pada Gambar 1.

ISSN:2442-5567 | E-ISSN:2443-289X



Gambar 1. Alur Kerangka Penelitian

Penelitian yang dilaksanakan adalah jenis penelitian eksprimen, yaitu melakukan pengujian tingkat akurasi algoritma *Fuzzy C-Means clustering* sebagai model untuk menentukan penerima beasiswa. Data eksperimen diambil dari tempat penelitian yaitu Universitas Megow Pak Tulang Bawang.

2.2 Pengumpulan Data

Pengumpulan data digunakan untuk mengumpulkan data-data dan informasi-informasi yang diperlukan dalam pembuatan sistem pendukung keputusan. Pengumpulan data pada penelitian ini menggunakan metode pengumpulan data *study literature* dan telaah dokumen.

a. *Study Literature*

Study literature dilakukan dengan cara mencari bahan materi yang berhubungan dengan permasalahan, algoritma *Fuzzy C-Means*, sistem pendukung keputusan dan beasiswa, guna mempermudah proses implementasi sistem. Pencarian materi dilakukan melalui pencarian di buku panduan dan internet.

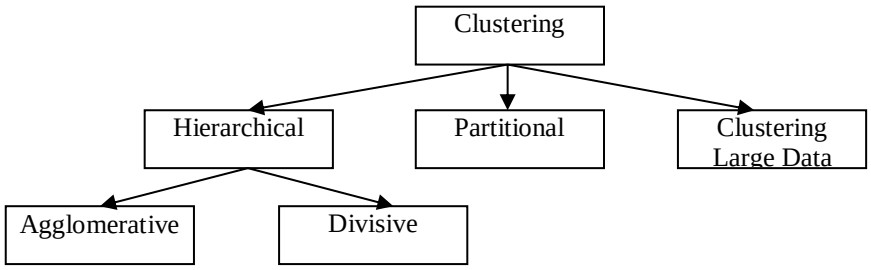
b. Telaah Dokumen

Telaah dokumen adalah pengumpulan data dengan cara mengumpulkan dan mempelajari dokumen-dokumen yang didapatkan dari pihak Jurusan Informaitka FMIPA Universitas Megow Pak Tulang Bawang. Dari metode pengumpulan data ini diperoleh 75 data mahasiswa.

2.3 Algoritma *Clustering*

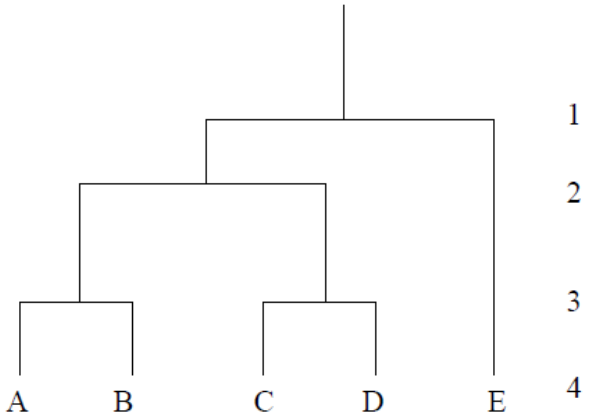
Secara umum pembagian algoritma *clustering* dapat digambarkan sebagai berikut:

ISSN:2442-5567 | E-ISSN:2443-289X



Gambar 2. Kategori Algoritma *Clustering*

Hierarchical clustering menentukan sendiri jumlah *cluster* yang dihasilkan. Hasil dari metode ini adalah suatu struktur data berbentuk pohon yang disebut dendogram dimana data dikelompokkan secara bertingkat dari yang paling bawah dimana tiap *instance* data merupakan satu *cluster* sendiri, hingga tingkat paling atas dimana keseluruhan data membentuk satu *cluster* besar berisi *cluster-cluster* seperti gambar 2.



Gambar 3. *Dendogram*

Divisive hierarchical clustering mengelompokkan data dari kelompok yang terbesar hingga ke kelompok yang terkecil, yaitu masing-masing *instance* dari kelompok data tersebut. Sebaliknya, *agglomerative hierarchical clustering* mulai mengelompokkan data dari kelompok yang terkecil hingga kelompok yang terbesar. Beberapa algoritma yang menggunakan metode ini adalah: *RObust Clustering Using LinKs* (ROCK), *Chameleon, Cobweb, Shared Nearest Neighbor* (SNN).

Partitional clustering yang mengelompokkan data ke dalam k *cluster* dimana k adalah banyaknya *cluster* dari input user. Kategori ini biasanya memerlukan pengetahuan yang cukup mendalam tentang data dan proses bisnis yang memanfaatkannya untuk mendapatkan kisaran nilai input yang sesuai. Beberapa algoritma yang masuk dalam kategori ini antara lain: *K-Means, Fuzzy C-Means, Clustering Large Aplications* (CLARA), *Expectation Maximation* (EM), *Bond Energy Algorithm* (BEA), *algoritma Genetika*, Jaringan Saraf Tiruan.

ISSN:2442-5567 | E-ISSN:2443-289X

a. Algoritma *Fuzzy Clustering C-Means* (FCM)

Pada proses pengklasteran (*clustering*) secara klasik (misalnya pada algoritma *Clustering K-Means*), pembentukan partisi dilakukan sedemikian rupa sehingga setiap obyek berada tepat pada satu partisi. Namun, adakalanya tidak dapat menempatkan suatu obyek tepat pada suatu partisi, karena sebenarnya obyek tersebut terletak di antara 2 atau lebih partisi yang lain. Pada logika *fuzzy*, metode yang dapat digunakan untuk melakukan pengelompokan sejumlah data dikenal dengan nama *fuzzy clustering*. *Fuzzy Clustering* lebih alami jika dibandingkan dengan pengklasteran secara klasik. Suatu algoritma *clustering* dikatakan sebagai *fuzzy clustering* jika algoritma tersebut menggunakan parameter strategi adaptasi secara *soft competitive*. Sebagian besar algoritma *fuzzy clustering* didasarkan atas optimasi fungsi obyektif atau modifikasi dari fungsi obyektif tersebut.

Salah satu teknik *fuzzy clustering* adalah *Fuzzy C-Means* (FCM). FCM adalah suatu teknik pengklasteran data yang keberadaan tiap-tiap data dalam suatu *cluster* ditentukan oleh nilai/derajat keanggotaan tertentu. Teknik ini pertama kali diperkenalkan oleh Jim Bezdek pada tahun 1981. Berbeda dengan teknik pengklasteran secara klasik (dimana suatu obyek hanya akan menjadi anggota suatu klaster tertentu), dalam FCM setiap data bisa menjadi anggota dari beberapa *cluster*.Batas-batas *cluster* dalam FCM adalah lunak (*soft*). Konsep dasar FCM, pertama kali adalah menentukan pusat *cluster* yang akan menandai lokasi rata-rata untuk tiaptiap *cluster*. Pada kondisi awal, pusat *cluster* ini masih belum akurat. Tiap-tiap data memiliki derajat keanggotaan untuk tiap-tiap *cluster*. Dengan cara memperbaiki pusat *cluster* dan nilai keanggotaan tiap-tiap data secara berulang, maa akan terlihat bahwa pusat *cluster* akan bergerak menuju lokasi yang tepat. Perulangan ini didasarkan pada minimasi fungsi obyektif. Fungsi Obyektif yang digunakan pada FCM adalah:

$$J_w(U,V;X) = \sum_{k-1}^{n} \sum_{i-=1}^{c} (\mu_{ik})^w (d_{ik})^2 \quad (1)$$

dengan + ∈ ,1,∞),

$$d_{ik} = d(x_k - v_i) = [\sum_{j=1}^{m} (x_{kj} - x_{ij})]^{\frac{1}{2}} \quad (2)$$

x adalah data yang akan diklaster:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix} \quad (3)$$

dan v adalah matriks pusat *cluster* :

$$V = \begin{bmatrix} v_{11} & \cdots & v_{1m} \\ \vdots & & \vdots \\ v_{m1} & \cdots & v_{mm} \end{bmatrix} \quad (4)$$

nilai Jw terkecil adalah yang terbaik, sehingga:

$$J_w^* (U^*, U^*; x) = \min J(U, U; x) \quad (5)$$

Jika $d_{ik} > 0, \forall i, k$; $w > 1$ dan X setidaknya memiliki m elemen, maka (u, v) ∈ $M_{fm} > 0, \forall i, k$; w dapat meminimasi Jw hanya jika:

$$\mu_{ik}=\frac{[\sum_{j=1}^{m}(x_{ij}-v_{kj})^{2}]^{\frac{-1}{w-1}}}{\sum_{i=1}^{n}[\sum_{j=1}^{m}(x_{ij}-v_{kj})^{2}]^{\frac{-1}{w-1}}}; \quad (6)$$

dan

$$v_{kj}=\frac{\sum_{i=1}^{n}((\mu_{ik})^{w}*X_{ij})}{\sum_{i=1}^{n}(\mu_{ik})^{w}};1 \leq i \leq m;1 \leq J \leq m \quad (7)$$

Algoritma *Fuzzy C-Means* (FCM) diberikan sebagai berikut :

1. Menentukan data yang akan di cluster X, berupa matriks berukuran n x m (n=jumlah sampel data, m = atribut setiap data). Xij=data sampel ke-I (i=1,2,…,n), atribut ke-j (j=1,2,…,m).
2. Menentukan :
   - Jumlah cluster = c
   - Pangkat = w
   - Maksimum interasi = MaxIter
   - Error terkecil yang diharapkan = ξ
   - Fungsi objektif awal = $P_0 = 0$
   - Interasi awal = t = 1

3. Membangkitkan bilangan random μik, i=1,2,3 …,n; k=1,2,3…c; sebagai elemen-elemen matriks partisi awal U.
   Menghitung jumlah setiap kolom :

$$Q_i=\sum_{k=1}^{c}\mu_{ik} \quad (8)$$

dengan j=1,2,…n.

Menghitung :

$$\mu_{ik}=\frac{\mu_{ik}}{Q_i} \quad (9)$$

4. Menghitung pusat cluster ke-k: V_{kj}, dengan k=1,2,…c; dan j=1,2,…m

$$V_{kj}=\frac{\sum_{i=1}^{n}((\mu_{ik})^{w}*X_{ij}}{\sum_{i=1}^{n}((\mu_{ik})^{w}} \quad (10)$$

5. Menghitung fungsi objektif pada interasi ke-t :

$$P_t=\sum_{i=1}^{n}\sum_{k}^{c}([\sum_{j=1}^{m}(X_{ij}-V_{kj})^{2}](\mu_{ik})^{w}) \quad (11)$$

6. Menghitung perubahan matriks partisi :

$$\mu_{ik}=\frac{[\sum_{j=1}^{m}(X_{ij}-V_{kj})^{2}]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}[\sum_{j=1}^{m}(X_{ij}-V_{kj})^{2}]^{\frac{-1}{w-1}}} \quad (12)$$

b. Proses *Clustering* menggunakan *Fuzzy C-Means*

Tahap ini akan diterapkan metode *Fuzzy C-Means* untuk mengelompokkan data. Hasil pengelompokkan ini kemudian akan digunakan untuk pertimbangan menentukan

ISSN:2442-5567 | E-ISSN:2443-289X

mahasiswa yang berhak menerima beasiswa. Adapun algoritma *C-Means Clustering* pada penerima beasiswa adalah sebagai berikut :

1. Menetapkan matriks partisi awal U berupa matriks berukuran n x m (n adalah jumlah sampel data, yaitu=75, dan m adalah parameter/atribut setiap data, yaitu=2). X_{ij}=data sampel ke-i (i=1,2,…,n), atribut ke-j (j=1,2,..,m).
2. Menentukan Nilai Parameter Awal :
   - Jumlah *cluster* (c) = 3
   - Pangkat/bobot (w) = 2
   - Maksimum interasi (MaxIter) = 100
   - Error terkecil yang diharapkan (ξ)= 10^{-5}
   - Fungsi Objektif awal (Po) = 0
   - Interasi awal (t) = 1
3. Membangkitkan bilangan random μik, i=1,2,…c; sebagai elemen-elemen matriks partisi awal (U).

3. HASIL DAN PEMBAHASAN

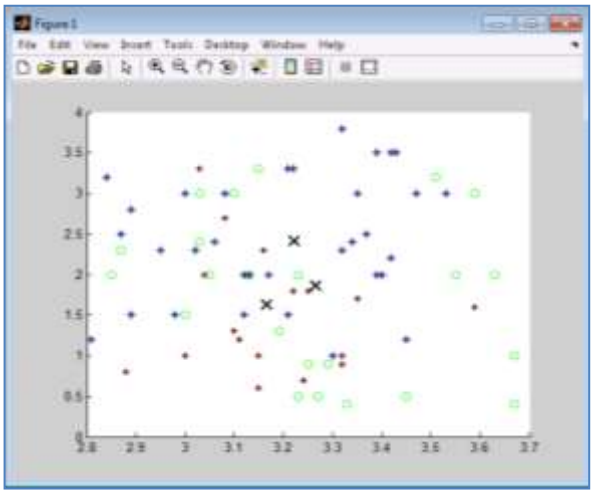
3.1 Tools menggunakan *Sofware* MATLAB

Data yang digunakan dalam penelitian ini adalah data penentuan beasiswa sebanyak 75 mahasiswa. Dalam proses *clustering*, IPK dijadikan sebagai Xi1, penghasilan orang tua dijadikan Xi2, tanggungan orang tua dijadikan Xi3 dan prestasi dijadikan Xi4. Pengolahan data pada penelitian ini menggunakan bantuan program yaitu Matlab. Dari *clustering* yang dilakukan diperoleh hasil yaitu nilai fungsi obyektif selama iterasi, pusat *cluster* atau *center* serta derajat keanggotaan lulusan untuk setiap *cluster* pada iterasi terakhir. Dalam penelitian ini, proses iterasinya berhenti pada iterasi ke-37 karena nilai |Pt - Pt-1|< ξ. Nilai fungsi obyektif pada iterasi terakhir yang diperoleh adalah 111.949781. Untuk penelitian ini *clustering* menggunakan *sofware* Matlab denga menjalankan di *Command Window* yaitu: >> fcm_k3.m

```
Iteration count = 1, obj. fcn = 185.869420
Iteration count = 2, obj. fcn = 146.530736
Iteration count = 3, obj. fcn = 141.726376
Iteration count = 4, obj. fcn = 134.833704
Iteration count = 5, obj. fcn = 130.961108
Iteration count = 6, obj. fcn = 129.145894
Iteration count = 7, obj. fcn = 127.219322
Iteration count = 8, obj. fcn = 124.903487
Iteration count = 9, obj. fcn = 122.880711
Iteration count = 10, obj. fcn = 121.359551
Iteration count = 11, obj. fcn = 119.898753
Iteration count = 12, obj. fcn = 118.236843
Iteration count = 13, obj. fcn = 116.491898
Iteration count = 14, obj. fcn = 114.963865
Iteration count = 15, obj. fcn = 113.865257
Iteration count = 16, obj. fcn = 113.180881
Iteration count = 17, obj. fcn = 112.770379
Iteration count = 18, obj. fcn = 112.513345
Iteration count = 19, obj. fcn = 112.342297
Iteration count = 20, obj. fcn = 112.223512
Iteration count = 21, obj. fcn = 112.139405
```

```
Iteration count = 22, obj. fcn = 112.079712
Iteration count = 23, obj. fcn = 112.037682
Iteration count = 24, obj. fcn = 112.008476
Iteration count = 25, obj. fcn = 111.988487
Iteration count = 26, obj. fcn = 111.975011
Iteration count = 27, obj. fcn = 111.966053
Iteration count = 28, obj. fcn = 111.960170
Iteration count = 29, obj. fcn = 111.956348
Iteration count = 30, obj. fcn = 111.953887
Iteration count = 31, obj. fcn = 111.952314
Iteration count = 32, obj. fcn = 111.951315
Iteration count = 33, obj. fcn = 111.950683
Iteration count = 34, obj. fcn = 111.950285
Iteration count = 35, obj. fcn = 111.950035
Iteration count = 36, obj. fcn = 111.949879
Iteration count = 37, obj. fcn = 111.949781
C =
   3.2649   1.8625   2.2484   4.2701
   3.1653   1.6271   5.1550   1.1770
   3.2225   2.4133   3.0759   0.7967
```

Penyebaran masing-masing anggota *cluster* pada iterasi terakhir dapat dilihat pada *cluster interface* gambar dibawah ini :



Gambar 4. Posisi *Claster* Iterasi terakhir

Keterangan :
Tanda * = *cluster* 1
Tanda + = *cluster* 2
Tanda ○ = *cluster* 3
Tanda X = pusat ke 3 *cluster*

Detail Matriks partisi U yang dihasilkan pada iterasi terakhir (iterasi ke-37) menggunakan fungsi Matlab adalah :

ISSN:2442-5567 | E-ISSN:2443-289X

Columns 1 through 10

0.1634	0.5724	0.0523	0.0467	0.9198	0.0643	0.0525	0.1063	0.0438	0.0605
0.1810	0.1653	0.0954	0.3021	0.0345	0.7542	0.2756	0.1726	0.3300	0.1892
0.6557	0.2623	0.8522	0.6512	0.0457	0.1815	0.6719	0.7211	0.6262	0.7503

Columns 11 through 20

0.0507	0.0576	0.6670	0.0923	0.6497	0.1132	0.6893	0.0108	0.7094	0.0427
0.7378	0.2660	0.1091	0.0971	0.1801	0.6406	0.1642	0.9393	0.0919	0.3048
0.2115	0.6763	0.2239	0.8107	0.1701	0.2463	0.1465	0.0500	0.1987	0.6525

Columns 21 through 30

0.0463	0.0299	0.8531	0.3603	0.0327	0.0592	0.0435	0.0470	0.8602	0.0845
0.7249	0.0572	0.0618	0.3797	0.8855	0.1855	0.3987	0.3511	0.0572	0.7507
0.2288	0.9129	0.0852	0.2600	0.0818	0.7553	0.5578	0.6018	0.0826	0.1648

Columns 31 through 40

0.8104	0.0560	0.1747	0.0183	0.3397	0.6931	0.0611	0.0556	0.2106	0.6219
0.0792	0.8198	0.6184	0.0430	0.3074	0.1693	0.1484	0.1314	0.5727	0.1510
0.1105	0.1242	0.2069	0.9386	0.3530	0.1376	0.7904	0.8130	0.2167	0.2271

Columns 41 through 50

0.0513	0.1739	0.5558	0.8307	0.8829	0.1085	0.0982	0.0307	0.0732	0.0897
0.8324	0.6221	0.2029	0.0788	0.0535	0.2561	0.1566	0.0896	0.4973	0.0934
0.1163	0.2040	0.2413	0.0905	0.0636	0.6354	0.7452	0.8797	0.4295	0.8170

Columns 51 through 60

0.0050	0.1026	0.0297	0.0466	0.0378	0.0673	0.0446	0.0707	0.5855	0.5157
0.9734	0.1039	0.0565	0.2871	0.1144	0.1508	0.3947	0.5757	0.1587	0.2346
0.0216	0.7934	0.9139	0.6663	0.8478	0.7819	0.5607	0.3536	0.2558	0.2497

Columns 61 through 70

0.0825	0.0826	0.0433	0.1208	0.6158	0.0877	0.1063	0.7764	0.0346	0.8865
0.2752	0.2753	0.3423	0.1954	0.1421	0.1292	0.1073	0.0887	0.0653	0.0543
0.6423	0.6421	0.6144	0.6838	0.2421	0.7831	0.7864	0.1349	0.9000	0.0592

Columns 71 through 75

0.8488	0.0548	0.7033	0.6666	0.7703
0.0635	0.4938	0.0943	0.1236	0.1006
0.0877	0.4514	0.2024	0.2097	0.1291

Pada iterasi terakhir (iterasi ke-37), pusat *cluster Vkj* yang dihasilkan dengan k=1,2,3,4 dan j=1,2 adalah:

$$V_{kj} = \begin{pmatrix} 3.2649 & 1.8625 & 2.2484 & 4.2701 \\ 3.1653 & 1.6271 & 5.1550 & 1.1770 \\ 3.2225 & 2.4133 & 3.0759 & 0.7967 \end{pmatrix}$$

ISSN:2442-5567 | E-ISSN:2443-289X

Nilai ini merupakan nilai dari koordinat ketiga titik pusat *cluster* dan memberikan garis besar tiap *cluster* yaitu:

1. Kelompok pertama (*cluster* ke-1), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.2649; memiliki rata-rata penghasilan orang tua sekitar 1.8625; tanggungan orang tua 2.2484 dan memiliki prestasi 4.2701.
2. Kelompok kedua (*cluster* ke-2), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.1653; memiliki rata-rata penghasilan orang tua sekitar 1.6271; tanggungan orang tua 5.1550 dan memiliki prestasi 1.1770.
3. Kelompok ketiga (*cluster* ke-3), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.2225; memiliki rata-rata penghasilan orang tua sekitar 2.4133; tanggungan orang tua 3.0759 dan memiliki prestasi 0.7967.

Derajat keanggotaan penentuan beasiswa untuk sertiap *cluster* pada iterasi terakhir (iterasi ke-37) dapat dilihat pada Tabel 1. di bawah ini:

Tabel 1. Derajat Keanggotaan Penentuan Beasiswa Untuk Setiap *Cluster* Pada Iterasi Terakhir

No	Derajat Keanggotaan (μ) pada Iterasi Terakhir			Kecenderungan data masuk pada *cluster*		
	(μ_1)	(μ_2)	(μ_3)	C1	C2	C3
1	0.1634	0.1810	0.6557			*
2	0.5724	0.1653	0.2623	*		
3	0.0523	0.0954	0.8522			*
4	0.0467	0.3021	0.6512			*
5	0.9198	0.0345	0.0457	*		
6	0.0643	0.7542	0.1815		*	
7	0.0525	0.2756	0.6719			*
8	0.1063	0.1726	0.7211			*
9	0.0438	0.3300	0.6262			*
10	0.0605	0.1892	0.7503			*
11	0.0507	0.7378	0.2115		*	
12	0.0576	0.2660	0.6763			*
13	0.6670	0.1091	0.2239	*		
14	0.0923	0.0971	0.8107			*
15	0.6497	0.1801	0.1701	*		
16	0.1132	0.6406	0.2463		*	
17	0.6893	0.1642	0.1465	*		
18	0.0108	0.9393	0.0500		*	
19	0.7094	0.0919	0.1987	*		
20	0.0427	0.3048	0.6525			*
21	0.0463	0.7249	0.2288		*	
22	0.0299	0.0572	0.9129			*
23	0.8531	0.0618	0.0852	*		
24	0.3603	0.3797	0.2600		*	
25	0.0327	0.8855	0.0818		*	
26	0.0592	0.1855	0.7553			*

ISSN:2442-5567 | E-ISSN:2443-289X

Nilai ini merupakan nilai dari koordinat ketiga titik pusat *cluster* dan memberikan garis besar tiap *cluster* yaitu:

1. Kelompok pertama (*cluster* ke-1), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.2649; memiliki rata-rata penghasilan orang tua sekitar 1.8625; tanggungan orang tua 2.2484 dan memiliki prestasi 4.2701.
2. Kelompok kedua (*cluster* ke-2), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.1653; memiliki rata-rata penghasilan orang tua sekitar 1.6271; tanggungan orang tua 5.1550 dan memiliki prestasi 1.1770.
3. Kelompok ketiga (*cluster* ke-3), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.2225; memiliki rata-rata penghasilan orang tua sekitar 2.4133; tanggungan orang tua 3.0759 dan memiliki prestasi 0.7967.

Derajat keanggotaan penentuan beasiswa untuk sertiap *cluster* pada iterasi terakhir (iterasi ke-37) dapat dilihat pada Tabel 2. di bawah ini:

Tabel 2. Derajat Keanggotaan Penentuan Beasiswa Untuk Setiap *Cluster* Pada Iterasi Terakhir

No	Derajat Keanggotaan (μ) pada Iterasi Terakhir			Kecenderungan data masuk pada *cluster*		
	(μ_1)	(μ_2)	(μ_3)	C1	C2	C3
1	0.1634	0.1810	0.6557			*
2	0.5724	0.1653	0.2623	*		
3	0.0523	0.0954	0.8522			*
4	0.0467	0.3021	0.6512			*
5	0.9198	0.0345	0.0457	*		
6	0.0643	0.7542	0.1815		*	
7	0.0525	0.2756	0.6719			*
8	0.1063	0.1726	0.7211			*
9	0.0438	0.3300	0.6262			*
10	0.0605	0.1892	0.7503			*
11	0.0507	0.7378	0.2115		*	
12	0.0576	0.2660	0.6763			*
13	0.6670	0.1091	0.2239	*		
14	0.0923	0.0971	0.8107			*
15	0.6497	0.1801	0.1701	*		
16	0.1132	0.6406	0.2463		*	
17	0.6893	0.1642	0.1465	*		
18	0.0108	0.9393	0.0500		*	
19	0.7094	0.0919	0.1987	*		
20	0.0427	0.3048	0.6525			*
21	0.0463	0.7249	0.2288		*	
22	0.0299	0.0572	0.9129			*
23	0.8531	0.0618	0.0852	*		
24	0.3603	0.3797	0.2600		*	
25	0.0327	0.8855	0.0818		*	
26	0.0592	0.1855	0.7553			*

Nilai ini merupakan nilai dari koordinat ketiga titik pusat *cluster* dan memberikan garis besar tiap *cluster* yaitu:

1. Kelompok pertama (*cluster* ke-1), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.2649; memiliki rata-rata penghasilan orang tua sekitar 1.8625; tanggungan orang tua 2.2484 dan memiliki prestasi 4.2701.
2. Kelompok kedua (*cluster* ke-2), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.1653; memiliki rata-rata penghasilan orang tua sekitar 1.6271; tanggungan orang tua 5.1550 dan memiliki prestasi 1.1770.
3. Kelompok ketiga (*cluster* ke-3), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.2225; memiliki rata-rata penghasilan orang tua sekitar 2.4133; tanggungan orang tua 3.0759 dan memiliki prestasi 0.7967.

Derajat keanggotaan penentuan beasiswa untuk sertiap *cluster* pada iterasi terakhir (iterasi ke-37) dapat dilihat pada Tabel 3. di bawah ini:

Tabel 3. Derajat Keanggotaan Penentuan Beasiswa Untuk Setiap *Cluster* Pada Iterasi Terakhir

No	Derajat Keanggotaan (μ) pada Iterasi Terakhir			Kecenderungan data masuk pada *cluster*		
	(μ_1)	(μ_2)	(μ_3)	C1	C2	C3
1	0.1634	0.1810	0.6557			*
2	0.5724	0.1653	0.2623	*		
3	0.0523	0.0954	0.8522			*
4	0.0467	0.3021	0.6512			*
5	0.9198	0.0345	0.0457	*		
6	0.0643	0.7542	0.1815		*	
7	0.0525	0.2756	0.6719			*
8	0.1063	0.1726	0.7211			*
9	0.0438	0.3300	0.6262			*
10	0.0605	0.1892	0.7503			*
11	0.0507	0.7378	0.2115		*	
12	0.0576	0.2660	0.6763			*
13	0.6670	0.1091	0.2239	*		
14	0.0923	0.0971	0.8107			*
15	0.6497	0.1801	0.1701	*		
16	0.1132	0.6406	0.2463		*	
17	0.6893	0.1642	0.1465	*		
18	0.0108	0.9393	0.0500		*	
19	0.7094	0.0919	0.1987	*		
20	0.0427	0.3048	0.6525			*
21	0.0463	0.7249	0.2288		*	
22	0.0299	0.0572	0.9129			*
23	0.8531	0.0618	0.0852	*		
24	0.3603	0.3797	0.2600		*	
25	0.0327	0.8855	0.0818		*	
26	0.0592	0.1855	0.7553			*

ISSN:2442-5567 | E-ISSN:2443-289X

1. Kelompok pertama (*cluster* ke-1), akan berisi data mahasiswa ke : 2, 5, 13, 15, 17, 19, 23, 29, 31, 36, 40, 43, 44, 45, 59, 60, 65, 68, 70, 71, 73, 74 dan 75.
2. Kelompok kedua (*cluster* ke-2), akan berisi data mahasiswa ke : 6, 11, 16, 18, 21, 24, 25, 30, 32, 33, 39, 41, 42, 49, 51, 58 dan 72.
3. Kelompok ketiga (*cluster* ke-3), akan berisi data mahasiswa ke : 1, 3, 4, 7, 8, 9, 10, 12, 14, 20, 22, 26, 27, 28, 34, 35, 37, 38, 46, 47, 48, 50, 52, 53, 54, 55, 56, 57, 61, 62, 63, 64, 66, 67 dan 69.

3.2 PEMBAHAAN

Proses *Clustering* memerlukan klasifikasi untuk menentukan kelompok (*cluster*) mana yang berhak untuk menerima beasiswa. Dalam penelitian akan mengelompokkan mahasiswa mejadi tiga (3) *cluster* yaitu :

1. *Cluster* yang menerima beasiswa
2. *Cluster* yang dipertimbangkan menerima beasiswa.
3. *Cluster* yang tidak menerima beasiswa.

Kemudian setiap *cluster* dibagi berdasarkan kriteria mana yang lebih diprioritaskan (berdasarkan IPK atau PO (Penghasilan Orang tua), Tanggungan Orang Tua (TO), Prestasi mahasiswa selama menjadi mahasiswa).

Iterasi pada percobaan ini berhenti pada iterasi ke-37. Hasil akhir *clutering* yang di perolah adalah :

1. *Cluster* pertama memiliki pusat *cluster* (3.2649; 1.8625; 2.2484; 4.2701)
2. *Cluster* kedua memiliki pusat *cluster* (3.1653; 1.6271; 5.1550; 1.1770)
3. *Cluster* ketiga memiliki pusat *cluster* (3.2225; 2.4133; 3.0759; 0.7967)

Hasil klasifikasi *clustering* dapat dilihat pada Tabel 4.

Tabel 4. Hasil Kalasifikasi

Proritas IPK		
Menerima	Dipertimbangkan	Tidak Menerima
Cluster 1	*Cluster* 3	*Cluster* 2
3.2649	3.2225	3.1653
Prioritas TK		
Menerima	Dipertimbangkan	Tidak Menerima
Cluster 3	*Cluster* 1	*Cluster* 2
2.4133	1.8625	1.6271
Prioritas TO		
Menerima	Dipertimbangkan	Tidak Menerima
Cluster 2	*Cluster* 3	*Cluster* 1
5.1550	3.0759	2.2484
Prioritas Prestasi		
Menerima	Dipertimbangkan	Tidak Menerima
Cluster 1	*Cluster* 2	*Cluster* 3
4.2701	1.1770	0.7967

ISSN:2442-5567 | E-ISSN:2443-289X

Hasil klasifikasi mahasiswa berdasarkan Tabel 4 untuk 75 data dapat dilihat pada Tabel 5 di bawah ini.

Tabel 5. Hasil klasifikasi prioritas mahasiswa

No	IPK	PO	TO	Prestasi	*C*	IPK	PO	TO	Prestasi
1	2.81	1.2	2	1	3	DP	DP	M	TM
2	2.85	2	3	3	1	M	TM	DP	M
3	2.84	3.2	3	1	3	DP	DP	M	TM
4	2.87	2.5	4	1	3	DP	DP	M	TM
5	2.87	2.3	2	5	1	M	TM	DP	M
6	2.88	0.8	6	0	2	TM	M	TM	DP
7	2.89	2.8	4	1	3	DP	DP	M	TM
8	2.89	1.5	2	0	3	DP	DP	M	TM
9	2.95	2.3	4	1	3	DP	DP	M	TM
10	2.98	1.5	3	0	3	DP	DP	M	TM
11	3	1	5	0	2	TM	M	TM	DP
12	3	3	4	1	3	DP	DP	M	TM
13	3	1.5	1	3	1	M	TM	DP	M
14	3.02	2.3	2	1	3	DP	DP	M	TM
15	3.03	3	4	5	3	DP	DP	M	TM
16	3.03	3.3	7	1	2	TM	M	TM	DP
17	3.03	2.4	4	5	1	M	TM	DP	M
18	3.04	2	5	1	2	TM	M	TM	DP
19	3.05	2	2	3	1	M	TM	DP	M
20	3.06	2.4	4	1	3	DP	DP	M	TM
21	3.08	2.7	5	1	2	TM	M	TM	DP
22	3.08	3	3	1	3	DP	DP	M	TM
23	3.1	3	2	5	1	M	TM	DP	M
24	3.1	1.3	4	3	2	TM	M	TM	DP
25	3.11	1.2	6	1	2	TM	M	TM	DP
26	3.12	1.5	3	0	3	DP	DP	M	TM
27	3.12	2	4	1	3	DP	DP	M	TM
28	3.13	2	4	0	1	M	TM	DP	M
29	3.13	2	1	5	2	TM	M	TM	DP
30	3.15	1	7	1	2	TM	M	TM	DP
31	3.15	3.3	2	5	1	M	TM	DP	M
32	3.15	0.6	6	1	2	TM	M	TM	DP
33	3.16	2.3	6	3	2	TM	M	TM	DP
34	3.17	2	3	1	3	DP	DP	M	TM
35	3.22	3.3	4	3	3	DP	DP	M	TM
36	3.19	1.3	4	5	1	M	TM	DP	M
37	3.21	1.5	3	1	3	DP	DP	M	TM

ISSN:2442-5567 | E-ISSN:2443-289X

38	3.21	3.3	3	0	3	DP	DP	M	TM
39	3.22	1.8	5	3	2	TM	M	TM	DP
40	3.23	0.5	2	3	1	M	TM	DP	M
41	3.24	0.7	6	1	2	TM	M	TM	DP
42	3.25	1.8	7	3	2	TM	M	TM	DP
43	3.25	0.9	3	3	1	M	TM	DP	M
44	3.27	0.5	2	5	1	M	TM	DP	M
45	3.29	0.9	2	5	1	M	TM	DP	M
46	3.3	1	3	1	3	DP	DP	M	TM
47	3.32	3.8	3	1	3	DP	DP	M	TM
48	3.32	2.3	3	0	3	DP	DP	M	TM
49	3.32	0.9	4	0	2	TM	M	TM	DP
50	3.34	2.4	2	1	3	DP	DP	M	TM
51	3.35	1.7	5	1	2	TM	M	TM	DP
52	3.35	3	2	1	3	DP	DP	M	TM
53	3.35	3	3	1	3	DP	DP	M	TM
54	3.37	2.5	4	0	3	DP	DP	M	TM
55	3.39	2	3	0	3	DP	DP	M	TM
56	3.39	3.5	3	0	3	DP	DP	M	TM
57	3.4	2	4	1	3	DP	DP	M	TM
58	3.32	1	4	1	2	TM	M	TM	DP
59	3.23	2	3	3	2	TM	M	TM	DP
60	3.33	0.4	3	3	1	M	TM	DP	M
61	3.42	3.5	4	1	3	DP	DP	M	TM
62	3.43	3.5	4	1	3	DP	DP	M	TM
63	3.42	2.2	4	1	3	DP	DP	M	TM
64	3.45	1.2	2	0	3	DP	DP	M	TM
65	3.45	0.5	1	3	1	M	TM	DP	M
66	3.47	3	2	0	3	DP	DP	M	TM
67	3.53	3	2	1	3	DP	DP	M	TM
68	3.51	3.2	1	5	1	M	TM	DP	M
69	3.53	3	3	1	3	DP	DP	M	TM
70	3.55	2	3	5	1	M	TM	DP	M
71	3.59	3	2	5	1	M	TM	DP	M
72	3.59	1.6	4	1	2	TM	M	TM	DP
73	3.63	2	2	3	1	M	TM	DP	M
74	3.67	1	2	3	1	M	TM	DP	M
75	3.67	0.4	1	5	1	M	TM	DP	M

Keterangan :
M : menerima.
DP : dipertimbangkan.

TM : tidak menerima.

Berdasarkan tabel prioritas di atas terdapat 3 *cluster* yaitu menerima, dipertimbangkan dan tidak menerima yang di hasilkan dari tabel kalsifikasi.

4. KESIMPULAN DAN SARAN

4.1 KESIMPULAN

Berdasarkan pembahasan maka dapat diberikan beberapa simpulan sebagai berikut:

1. Algoritma *Fuzzy C-means* dapat digunakan untuk mengelompokan data mahasiswa lebih halus dengan menerap- kan besarnya derajat keanggotaan setiap elemen untuk masuk ke dalam kelom- pok-kelompok yang ada.
2. Pengujian data dilakukan sebanyak 37 iterasi, diperoleh tiga kelompok berdasarkan nilai rata-rata penentuan beasiswa, yaitu :
   a. Kelompok pertama (*cluster* ke-1), berisi mahasiswa yang memiliki rata-rata ipk sekitar 3.2649; memiliki rata-rata pendapatan orang tua sekitar 1.8625; tanggungan orang tua 2.2484 dan memiliki prestasi 4.2701.
   b. Kelompok kedua (*cluster* ke-2), berisi mahasiswa yang memiliki rata-rata ipk sekitar 3.1653; memiliki rata-rata pendapatan orang tua sekitar 1.6271; tanggungan orang tua 5.1550 dan memiliki prestasi 1.1770.
   c. Kelompok ketiga (*cluster* ke-3), berisi mahasiswa yang memiliki rata-rata ipk sekitar 3.2225; memiliki rata-rata pendapatan orang tua sekitar 2.4133; tanggungan orang tua 3.0759 dan memiliki prestasi 0.7967.
3. Kemudian setiap *cluster* diklasifikasikan berdasarkan kriteria mana yang lebih diprioritaskan dengan nilai terbesar pada jarak akhir merupakan *cluster* yang menerima beasiswa, sedangkan *cluster* dengan nilai terkecil merupakan *cluster* yang tidak berhak menerima beasiswa.

4.2 SARAN

Saran yang dapat digunakan untuk mendapat tingkat akurasi yang lebih baik dalam penerapan algoritma *Fuzzy C-Means* untuk penentuan beasiswa di Univeristas Megou Pak Tulang Bawang, serta untuk memberikan konstribusi yang lebih besar di dunia riset, disarankan agar hasil penelitian ini dikembangkan dengan cara memodifikasi/*updating* algoritma *Fuzzy C-means* yang digunakan saat ini, atau dengan menggabungkan algoritma *Fuzzy C-means* dengan algoritma lain.

REFERENSI

[1] Ahmad Mauliyadi M, dkk. (2013) , " Perbandingan Metode *Fuzzy C-Means* (Fcm) Dan *Fuzzy Gustafson-Kessel* (Fgk) Menggunakan Data Citra Satelit *Quickbird", Matematika 00 (2013) 01–05,*

[2] Ali Keshavarzi, Fereydoon Sarmadian (2012)," Fuzzy Clustering Analysis For Modeling Of Soil Cation Exchange Capacity", AJAE 3(1):27-33 (2012), ISSN:1836-9448.

[3] Apriansyah Putra, Dinna Yunika Hardiyanti (2011) "Penentuan Penerima Beasiswa Dengan Menggunakan *Fuzzy Multiple Atribute Decission Makin ",Vol. 3, No. 1, April 2011, Halaman 286-293.*

ISSN:2442-5567 | E-ISSN:2443-289X

[4] Bahar, Vincent Suhartono, dkk (2011), "Penentuan Jurusan Sekolah Menengah Atas Dengan Algoritma Fuzzy C-Means", Jurnal Teknologi Informasi, Issn 1414-9999.

[5] Dimas Wahyu Wibowo, dkk (2013) " Perhitungan Jumlah dan Jenis Kendaraan Menggunakan Metode Fuzzy C-means dan Segmentasi Deteksi Tepi Canny" , *Vol. 7, No. 2, Desember 2013, Halaman 103-110.*

[6] Eko Sediyono, dkk (2006), " Penentuan Lokasi Fasilitas Gudang Menggunakan *Fuzzy C – Means* (FCM)", Vol.2, No. 2, Desember 2006: 155 – 166.

[7] Ivan Hardiyanto, dkk (2012), " Implementasi Segmentasi Citra Dengan Menggunakan Metode Generalized Fuzzy Cmeans Clustering Algorithm With Improved Fuzzy Partitions", *Vol. 1, No. 1, (2012) 1-5*

[8] Heru Dibyo Laksono, Abraham Arief, (2012)," Perkiraan Kebutuhan Energi Listrik Jangka Panjang Di Provinsi Sumatera Barat Sampai Tahun 2021 Dengan Logika Fuzzy Clustering ",Vol. 5 No. 2 September 2012, Issn : 2086 – 4981.

[9] Kusumadewi, S., Purnomo, H., (2010), *Aplikasi Fuzzy Untuk Pendukung Keputusan*, Graha Ilmu, Jakarta

[10] Kusrini, (2006), Algoritma Data Mining, Penerbit ANDI, Yogyakarta.

[11] Noor Fitriana Hastuti, (2010), Pemanfaatan Metode *K-Means Clustering* dalam Penentuan Penerima Beasiswa, Jurusan Informatika. Fakultas MIPA. Universitas Sebelas Maret.

[12] Prabowo Pudjo Widodo, dkk (2013), " Penerapan *Data Mining* dengan *Matlab",* Penerbit Rekayasa Sains, Bandung.

[13] Turban, Efraim, et al . (2005). Decision Support Systems and Intelligent Systems. Yogyakarta : Penerbit Andi

Kumpulan jurnaL Ilmu Komputer (KLIK)
Volume 01, No.01 September 2014
ISSN: 2406-7857

IMPLEMENTASI METODE FUZZY C-MEANS PADA SISTEM CLUSTERINGDATA VARIETAS PADI

Nurjanah[1], Andi Farmadi[2], Fatma Indriani[2]
[1,2,3] Prodi Ilmu Komputer Fakultas MIPA Universitas Lambung Mangkurat

Jl. A. Yani Km 36 Banjarbaru, Kalimantan Selatan
Email: nurjanah_abs@yahoo.co.id

Abstract

Mutations with gamma rays conducted on five local rice varieties tidal South Kalimantan produce a lot of data availability. In order for these data not only become a graveyard of useless data required a method that could be used to probe the hidden information from the data. The method known as data mining. Data mining is a technique to gain knowledge from the data by looking for certain patterns or rules of a number of large amounts of data. One method of data mining is clustering, where clustering is usually used to group objects that are similar in the same class or segment. By utilizing the data of local rice varieties tidal South Kalimantan mutated by gamma rays, data mining process is done by grouping the data based on the harvest age, productive tillers, and weight of 1000 seeds into 4 groups using fuzzy c-means algorithm. From that cluster information, carried ranking using the Simple Additive Weighting method and acquired knowledge about improved varieties by harvest age, productive tillers, and a weight of 1000 is kuatek with a dose of 30 krad.

Keywords *: Data Mining, Cluster, Fuzzy C-Means, Local Rice Varieties*

Abstrak

Mutasi dengan sinar gamma yang dilakukan terhadap lima varietas padi lokal pasang surut kalimantan selatan menghasilkan tersedianya banyak data. Agar data-data tersebut tidak hanya menjadi kuburan data yang tidak berguna dibutuhkan sebuah metode yang bisa digunakan untuk menggali informasi–informasi tersembunyi dari data tersebut. Metode tersebut dikenal dengan data mining. Data mining merupakan suatu teknik untuk menggali pengetahuan dari data dengan mencari pola atau aturan tertentu dari sejumlah data dalam jumlah besar. Salah satu metode data mining adalah klastering, dimana klastering biasanya digunakan untuk mengelompokan objek-objek yang memiliki kemiripan dalam kelas atau segmen yang sama. Dengan memanfaatkan data varietas padi hasil mutasi dengan sinar gamma dilakukan proses penggalian data dengan cara mengelompokkan data umur panen, anakan produktif, dan bobot 1000 menjadi 4 klaster menggunakan algoritma fuzzy c-means. Dari informasi 4 klaster tersebut, dilakukan perangkingan klaster menggunakan metode Simple Additive Weighting dan diperoleh pengetahuan tentang varietas unggul berdasarkan umur panen, anakan produktif, dan bobot 1000 yaitu kuatek dengan dosis 30 krad.

Kata kunci *: Data Mining, Klaster, Fuzzy C-Means, Varietas Padi Lokal*

1. PENDAHULUAN

Varietas padi lokal Kalimantan Selatan masih menjadi pilihan para petani untuk ditanam, yang dalam penanamannya tidak terlalu bergantung pada input pertanian. Selain itu, varietas padi lokal juga bersifat adaptif dengan lingkungan Kalimantan Selatan yang didominasi oleh lahan rawa, meskipun berumur panjang dan produktivitas relatif rendah daripada varietas padi unggul. Akan tetapi sifat "pera" nasi yang dihasilkan membuat harga jual varietas padi lokal lebih tinggi dibandingkan dengan varietas padi unggul [3].

Perjuangan dalam peningkatan produksi pangan untuk memenuhi kebutuhan masyarakat Kalimantan Selatan merupakan tantangan bagi masyarakat pertanian yang dapat melahirkan peluang berupa pengembangan teknologi alternatif untuk menerobos kebuntuan dan stagnasi produksi beras lokal. Alternatif itu dapat diarahkan pada pengembangan varietas lokal yang dapat beradaptasi pada lingkungan produksi yang buruk dengan produktivitas yang tinggi dan umur panen yang pendek. Peningkatan keragaman populasi dapat dilakukan salah satunya melalui induksi mutasi secara fisik dengan iradiasi sinar gamma.

Salah satu upaya untuk meningkatkan produksi pangan varietas padi lokal pasang surut kalimantan selatan yaitu dengan cara meningkatkan keragaman populasi, dilakukan radiasi dengan sinar gamma terhadap lima varietas padi. Dari upaya tersebut dihasilkan banyak data salah satunya yaitu data varietas padi hasil mutasi dengan sinar gammamaka dibutuhkan sebuah metode yang bisa digunakan untuk menggali informasi–informasi tersembunyi dari data tersebut. Metode tersebut dikenal dengan *data mining*.

Data mining adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar. Salah satu tugas dalam data mining adalah klastering. Tujuan utama dari klastering adalah pengelompokan sejumlah data/obyek ke dalam klaster sehingga dalam setiap klaster akan berisi data yang semirip mungkin [1].

Penggalian suatu data menjadi pengetahuan adalah kegiatan yang sangat berguna, karena pengetahuan tersebut bisa digunakan untuk memecahkan suatu masalah yang ada. *Data mining* merupakan salah satu teknik untuk menggali pengetahuan dari data. Agar data-data tersebut tidak menjadi kuburan data yang tidak bergunamaka dibuat sistem yang mampu menggali data varietas padi yang dimutasi dengan sinar gamma yang dapat mengorganisir data varietas padi yang dimutasi dengan sinar gamma dalam suatu database serta dapat menghasilkan pengelompokan sejumlah data/obyek ke dalam klaster yang memberikan pengetahuan mengenai varietas unggulmenggunakan metode *Fuzzy C-Means* (FCM).

2. METODOLOGI PENELITIAN

2.1. Analisa Data

Pada tahap ini akan dilakukan analisis data dengan KDD (*knowledge Discovery in Databases*). Keseluruhan proses KDD dari mulai memasukan data mengalami 3 tahap secara global sebelum menjadi hasil yang berupa informasi, yaitu:

(a) *Preprocessing*

Pada tahap ini struktur basis data akan dipersiapkan sehingga mempermudah proses mining. Proses *preprocessing* ini mencakup tiga hal utama yaitu :

1) seleksi data
   untuk memilih data yang akan digunakan dalam proses data mining. Dalam penelitian ini dipilih 3 data varietas padi yaitu umur panen, anakan produktif, dan bobot 1000 biji;
2) *Cleaning*
   Memastikan kualitas data yang telah dipilih pada tahap seleksi data, pada tahap ini masalah yang harus dihadapi adalah noisy data dan missing values. Proses cleaning data perlu dilakukan agar data bersih dari duplikasi data, data yang tidak konsisten, atau kesalahan cetak. Sehingga data yang telah melewati proses ini siap untuk di proses di data mining. Pada penelitian ini proses pembersihan data hanya dilakukan pada data yang missing value;
3) integrasi data
   Setelah melakukan pembersihan data yang menyeluruh, maka akan dilakukan tahapan selanjutnya yaitu integrasi data, karena data lebih dari satu tabel maka dilakukan tahap penggabungan data dari berbagai tabel. Kemudian data padi diberi kode agar tidak ada data yang memiliki nama yang sama. Kode padi terdiri dari nomor padi, baris padi, dosis radiasi, dan varietas padi.

(b) Data Mining

Proses data mining yang dilakukan pada penelitian ini adalah menerapkan teknik *clustering* dengan algoritma *Fuzzy C-Means* untuk mendapatkan karakteristik dari data. *Cluster* yang digunakan pada penelitian ini sebanyak empat klaster.

Konsep dasar FCM (*Fuzzy C-Means*) pertama kali adalah menentukan pusat *cluster* yang akan menandai lokasi rata-rata untuk tiap *cluster*. Pada kondisi awal, pusat *cluster* ini masih belum akurat. Tiap-tiap titik data memiliki derajat keanggotan untuk tiap-tiap *cluster*. Dengan cara memperbaiki pusat *cluster* dan derajat keanggotaan tiap-tiap titik data secara berulang, maka akan dapat dilihat bahwa pusat *cluster* akan bergerak menuju lokasi yang tepat. Perulangan ini didasarkan pada derajat keanggotaan yang menggambarkan jarak dari titik data yang diberikan ke pusat *cluster* yang terbobot oleh derajat keanggotaan titik data tersebut [6].

Pada perhitungan menggunakan FCM dengan menggunakan data yang sama tetapi diolah dengan jumlah *cluster* yang berbeda, maka hasil pengelompokannya akan sedikit berbeda, karena data tidak diolah dengan satu variabel saja tetapi dengan semua variabel. Perbedaan hasil pengelompokan itu dikarenakan data pada kelompok tertentu kemungkinan akan berpindah pada kelompok lain jika diolah dengan jumlah *cluster* yang berbeda, ini menunjukkan bahwa sistem aplikasi sudah berjalan dengan benar [6]

Output dari FCM bukan merupakan fuzzy inference sistem, namun merupakan derajat pusat *cluster* dan beberapa derajat keanggotaan untuk

tiap-tiap titik data. Algoritma *Fuzzy C-Means* disusun dengan langkah sebagai [6]:

(1) Input data
input data yang akan di*cluster*yaitu data varietas padi (X) berupa matrik berukuran n × m (n = jumlah data, m = atribut setiap data). X_{ij} = data ke- i (i = 1, 2, …, n), atribut ke- j (j = 1, 2, …, m).

(2) Batasan
(a) Jumlah *cluster* = c = 4
(b) Pangkat = w = 2
(c) Maksimum iterasi = Maxit = 100
(d) Error terkecil yang diharapkan = ξ = 10^{-3}
(e) Fungsi obyektif awal = Po = 0
(f) Iterasi awal = t = 1

(3) Membangkitkan bilangan random μ_{ik}, i = 1, 2, …, n; k = 1, 2, …,c; sebagai elemen-elemen matrik partisi awal U, dengan jumlah setiap nilai elemen kolom dalam satu baris adalah 1 (satu)

$$\sum_{i=1}^{c} \mu_{ci} = 1 \qquad \dots (1)$$

(4) Menghitung pusat *cluster* ke- k : V_{kj} dengan k = 1, 2, …, c; dan j = 1, 2, …,m.

$$V_{kj} = \frac{\sum_{i=1}^{n}\left((\mu_{ik})^{w} * X_{ij}\right)}{\sum_{i-1}^{n}(\mu_{ik})^{w}} \qquad \dots (2)$$

(5) Menghitung fungsi obyektif pada iterasi ke- t, P_t :

$$P_t = \sum_{i=1}^{n}\sum_{k=1}^{c}\left(\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right](\mu_{ik})^{w}\right) \qquad \dots (3)$$

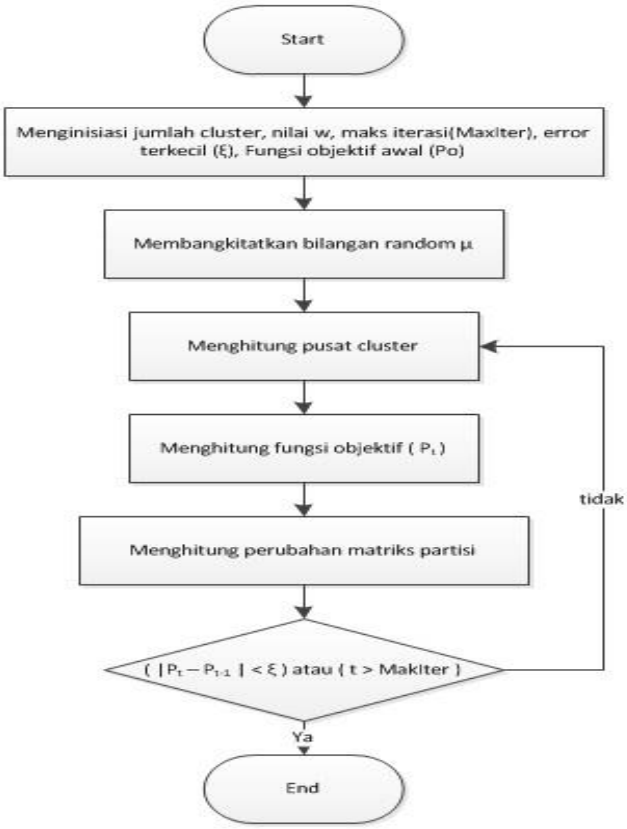
(6) Menghitung perubahan matrik partisi:

$$\mu_{ik} = \frac{\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right]^{\frac{-1}{w-1}}} \dots (4)$$

dengan i = 1, 2, …, n; dan k = 1, 2, …, c

(7) Mengecek kondisi berhenti:
1. Jika : ($|P_t - P_{t-1}| < \xi$) atau (t > MakIter) maka berhenti;
2. Jika tidak : t = t + 1, ulangi langkah ke- d (menghitung V_{kj}).

Bagan alir untuk metode *Fuzzy C-Means* yaitu sebagai berikut :



Gambar 1. Bagan Alir Metode *Fuzzy C-Means*

(c) ***Postprocessing***

Postprocessing bertujuan untuk membantu pengguna dalam memahami informasi, Karena pusat *cluster* masih tidak bisa dienterpretasikan secara langsung, maka diperlukan metode *Simple Additive Weighting* yang bisa melakukan perangkingan nilai pusat *cluster* terakhir tersebut.

Metode SAW (*Simple Additive Weighting*) sering juga dikenal istilah metode penjumlahan terbobot. Konsep dasar metode SAW adalah mencari penjumlahan terbobot dari rating kinerja pada setiap alternatif pada semua atribut. Langkah-langkah metode dalam metode SAW adalah [7] :

1. Membuat matriks keputusan Z berukuran m x n, dimana m = alternatif yang akan dipilih dan n = kriteria .
2. Memberikan nilai x setiap alternatif (i) pada setiap kriteria (j) yang sudah ditentukan, dimana, i=1,2,…m dan j=1,2,…n pada matriks keputusan Z,

$$\mathbf{Z} = \begin{vmatrix} x_{11} & x_{y12} & \dots & x_{1j} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ x_{i1} & x_{i2} & & x_{ij} \end{vmatrix} \dots (5)$$

3. Memberikan nilai bobot preferensi (W) oleh pengambil keputusan untuk masing-masing kriteria yang sudah ditentukan.

$$W = [W_1\ W_2 W_3 \dots\ W_j] \dots(6)$$

4. Melakukan normalisasi matriks keputusan Z dengan cara menghitung nilai rating kinerja ternormalisasi (r_{ij}) dari alternatif A_i pada atribut C_j

$$r_{ij} = \begin{cases} \frac{x_{ij}}{\underset{i}{MAX}(x_{ij})} & \text{jika j adalah atribut keuntungan} \\ \frac{\underset{i}{MIN}(x_{ij})}{x_{ij}} & \text{jika j adalah atribut biaya} \end{cases} \dots(7)$$

Dengan ketentuan :

a. Dikatakan atribut keuntungan apabila atribut banyak memberikan keuntungan bagi pengambil keputusan, sedangkan atribut biaya merupakan atribut yang banyak memberikan pengeluaran jika nilainya semakin besar bagi pengambil keputusan.
b. Apabila berupa atribut keuntungan maka nilai (x_{ij}) dari setiap kolom atribut dibagi dengan nilai (MAX x_{ij}) dari tiap kolom, sedangkan untuk atribut biaya, nilai (MIN x_{ij}) dari tiap kolom atribut dibagi dengan nilai (x_{ij}) setiap kolom.

5. Hasil dari nilai rating kinerja ternormalisasi (r_{ij}) membentuk matriks ternormalisasi (N)

$$\mathbf{N} = \begin{vmatrix} r_{11} & r_{y12} & \dots & r_{1j} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ r_{i1} & r_{i2} & & r_{ij} \end{vmatrix} \dots (8)$$

6. Melakukan proses perankingan dengan cara mengalikan matriks ternormalisasi (N) dengan nilai bobot preferensi (W).
7. Menentukan nilai preferensi untuk setiap alternatif (V_i) dengan cara menjumlahkan hasil kali antara matriks ternormalisasi (N) dengan nilai bobot preferensi (W).

$$V_i = \sum_{j=1}^{n} w_j r_{ij} \dots (9)$$

Nilai Vi yang lebih besar mengindikasikan bahwa alternatif Ai merupakan alternatif terbaik.

2.3. Implementasi

Implementasi dilakukan dengan metode Fuzzy C-Means (FCM) pada sistem clustering data varietas padilokal pasang surut kalimantan selatan yang dimutasi dengan sinar gamma menggunakan delphi XE2

2.4. Pengujian

Untuk memastikan apakah perhitungan dengan metode fuzzy c-means sudah benar dilakukan uji coba perhitungan secara manual meggunakan MS. Excel dan Matlab.Pengujian dilakukan untuk menghindari kesalahan dari sistem yang dibuat. Apabila terjadi kesalahan, maka sistem akan diperbaiki kembali sampai hasil proses sesuai dengan apa yang diharapkan.

3. HASIL DAN PEMBAHASAN

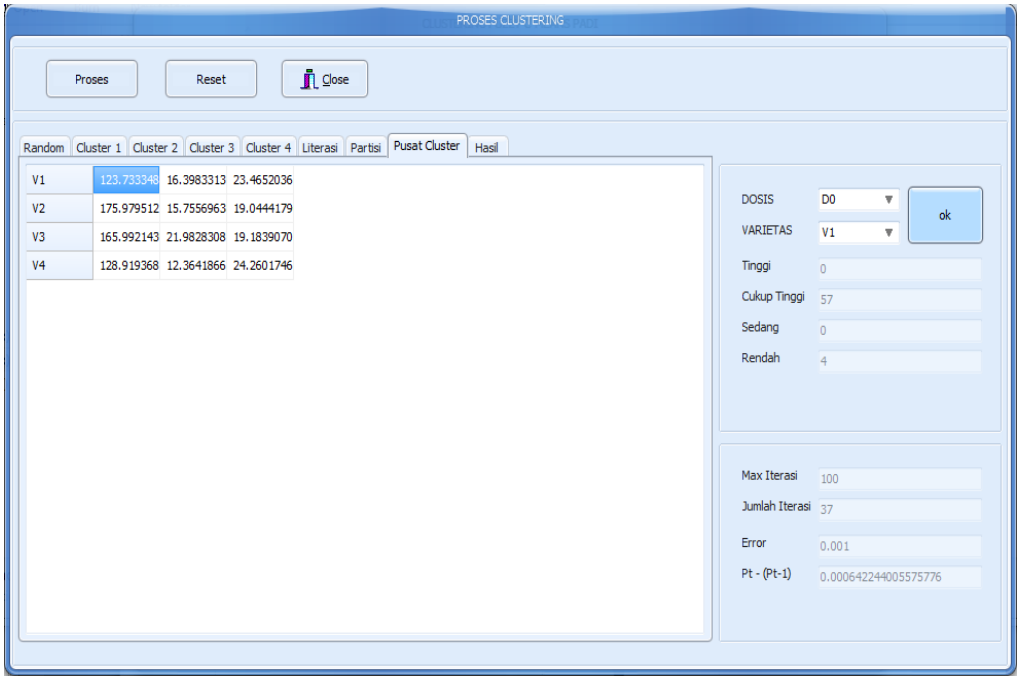
Dariproses clustering Fuzzy C-Means dengan batasan jumlah cluster sebanyak 4 cluster, pangkat 2, masimal iterasi 100, error 0,001 proses berhenti pada iterasi ke-37 dengan error 0,000642 dan diperoleh pusat cluster iterasi terakhir yaitu pada Tabel 1 atau pada Gambar 2 dengan menggunakan rumus dibawah ini (persamaan 2):

$$V_{kj} = \frac{\sum_{i=1}^{2420}\left((\mu_{ik})^2 * X_{ij}\right)}{\sum_{i=1}^{2420}(\mu_{ik})^2}$$

Keterangan :
V_{kj}: pusat cluster dengan k=1,2,...,c dan j = 1,2,...,n
X_{ij} : data ke-i (i=1,2,...m), atribut ke-j (j=1,2,...n)
μ_{ik} : bilangan acak dengan i=1,2,...m dan k=1,2,...c

Tabel 1. Pusat Cluster Delphi

	Umur Panen	Anakan Produktif	Bobot 1000
Klaster 1	165.992143	21.9828308	19.1839070
Klaster 2	123.733348	16.3983313	23.4652036
Klaster 3	128.919368	12.3641866	24.2601746
Klaster 4	175.979512	15.7556963	19.0444179



Gambar 2. Form Proses Fuzzy C-Means

Dari Tabel 1 dapat dilihat bahwa klaster 1 memiliki umur panen terendah ketiga, anakan produktif terbanyak pertama, dan bobot 1000 terberat ketiga. Klaster 2 memiliki umur panen terendah pertama, anakan produktif terbanyak kedua, dan bobot 1000 terberat kedua. Klaster 3 memiliki umur panen terendah kedua, anakan produktif terbanyak keempat, dan bobot 1000 terberat pertama.

Klaster 4 memiliki umur panen terendah keempat, anakan produktif terbanyak ketiga, dan bobot 1000 terberat keempat.

Dilihat dari umur panen klaster 2 lebih unggul karena memiliki umur panen terendah dibandingkan klaster 1, klaster 3, dan klaster 4. Untuk anakan produktif klaster 1 lebih unggul karena memiliki anakan produktif terbanyak. Dan untuk bobot 1000 klaster 3 lebih unggul karena memiliki bobot seribu terberat apabila dibandingkan dengan klaster lainnya.

Dari hasil pusat klaster (Tabel 1) tersebut dilakukan perangkingan dengan metode *Simple Additive Weighting* (SAW) untuk melihat klaster mana yang memiliki umur panen, anakan produktif dan bobot 1000 terbaik. Perhitungan nilai prefrensi dilakukan menggunakan rumus dibawah ini (persamaam 9) :

$$V_i = \sum_{j=1}^{n} w_j r_{ij}$$

Keterangan :
V_i : nilaipreferensiuntuksetiapalternatif
w_j : nilai bobot preferensi
r_{ij} : matriks ternormalisasi

Berikut tabel nilai prefrensi dan tabel hasil pengelompokan per varietas :

Tabel 2. Nilai Prefensi Klaster

Klaster	Nila Prefensi	Kategori
Klaster 1	2.535898966	Cukup Tinggi
Klaster 2	2.712760587	Tinggi
Klaster 3	2.53314209	Sedang
Klaster 4	2.209208036	Rendah

Tabel 3. Hasil Pengelompokan Varietas Siam Harli

Kategori	Varietas Siam Harli (V1)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	-	-	-	46
Cukup Tinggi	57	77	32	-
Sedang	-	-	-	44
Rendah	4	2	3	-

Tabel 4. Hasil Pengelompokan Varietas Siam Unus

Kategori	Varietas Siam Unus (V2)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	-	1	-	-
Cukup Tinggi	48	205	48	86
Sedang	-	1	4	-
Rendah	4	-	5	8

Tabel 5. Hasil Pengelompokan Varietas Kuatek

Kategori	Varietas Kuatek (V3)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	-	-	-	371
Cukup Tinggi	7	152	18	176
Sedang	-	-	-	-
Rendah	41	5	132	-

Tabel 6. Hasil Pengelompokan Varietas Siam Sebelas

Kategori	Varietas Siam Sebelas (V4)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	5	-	-	25
Cukup Tinggi	1	28	9	-
Sedang	1	-	-	22
Rendah	25	158	78	-

Tabel 7. Hasil Pengelompokan Varietas Siam Gumpal

Kategori	Varietas Siam Gumpal (V5)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	-	-	-	13
Cukup Tinggi	-	-	-	-
Sedang	-	-	-	116
Rendah	54	173	105	-

Pengetahuan yang didapat dari hasil penambangan data/klastering ini berupa varietas unggul yaitu varietas yang masuk kategori tinggi. Varietas yang masuk kategori tinggi terdiri atas rata-rata umur panen sebesar 123,7665737; rata-rata anakan produktif sebesar 16,2834365; dan rata-rata bobot 1000 sebesar 23,46984343 dengan nilai preferensi sebesar 2,712760587. Diantara kelima varietas tersebut yang paling banyak masuk kategori tinggi yaitu varietas kuatek dengan dosis 30 krad yakni sebanyak 371 data padi atau sebesar 67,82 % data padi masuk kategori tinggi

Tabel 8. Presentase Data Padi Yang Termasuk Kategori Tinggi

Varietas	Dosis			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Siam Harli (V1)	-	-	-	51,11%
Siam Unus (V2)	-	0,93%	-	-
Kuatek (V3)	-	-	-	67,82%
Siam Sebelas (V4)	15,61%	-	-	53,19%
Siam Gumpal (V5)	-	-	-	10,08%

4. KESIMPULAN

Dari hasil penelitian dan pengamatan dari sistem yang telah dibuat, maka dapat diperoleh kesimpulan sebagai berikut :

1. Metode Fuzzy C-Means berhasil diimplemantasikan pada clustering data varietas padi lokal pasang surut kalimantan selatan yang dimutasi dengan sinar gamma.
2. Dilihat dari umur panen klaster 2 lebih unggul karena memiliki umur panen terendah, untuk anakan produktif klaster 1 lebih unggul karena memiliki anakan produktif terbanyak, dan untuk bobot 1000 klaster 3 lebih unggul karena memiliki bobot seribu terberat apabila dibandingkan dengan klaster lainnya.
3. Pengetahuan yang didapat mengenai varietas unggul berdasarkan umur panen, anakan produktif dan bobot 1000 terbaik menggunakan SAW memiliki nilai preferensi sebesar 23,46984343 yaitu kategori tinggi.
4. Diantara kelima varietas yang paling banyak masuk kategori tinggi yaitu varietas kuatek dengan dosis 30 krad yakni sebanyak 371 data padi atau sebesar 67,82 % data padi masuk kategori tinggi.

DAFTAR PUSTAKA

[1]. Budi Santosa, "**Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis***",* Graha Ilmu, Yogyakarta, 2007.

[2]. Iko Pramudiono. "***Pengantar Data Mining : Menambang Permata Pengetahuan di Gunung Data***" 2003.

[3]. Noor, M., Y. Rina dan Noorginayuwati, "**Persepsi petani tentang lahan gambut dan pengelolaannya",** Laporan Penelitian, Balai Penelitian Pertanian Lahan Rawa, Banjarbaru, 2007.

[4]. Nurjanah, "**Clustering Varietas Padi Lokal Pasang Surut Kalimantan Selatan Yang Dimutasi Dengan Sinar Gamma Menggunakan Metode Fuzzy C-Means**" Naskah Skripsi Universitas Lambung Mangkurat, Banjarbaru, 2014.

[5]. Nuqson Masykur Huda, "***Aplikasi Data Mining Untuk Menampilkan Informasi Tingkat Kelulusan Mahasiswa (Studi Kasus Di Fakultas MIPA Universitas Diponegoro)",*** Naskah Skripsi Universitas Diponegoro, Semarang, 2010.

[6]. Sri Kusumadewi and Hari Purnomo, "**Aplikasi Logika Fuzzy Untuk Pendukung Keputusan***",* Yogyakarta, Graha Ilmu, 2010.

[7]. Wibowo S, Henry., Amalia, Riska., Fadlun M, Andi., Arivanty, Kurnia, "**Sistem Pendukung Keputusan Untuk MenentukanPenerima Beasiswa Bank BRI Menggunakan FMADM (Studi Kasus: Mahasiswa Fakultas Teknologi Industri Unversitas Islam Indonesia)***",* Seminar Nasional Aplikasi Teknologi Informasi, Yogyakarta, 2008.

Fuzzy C- Means Algorithm- A Review

R.Suganya, R.Shanthi

Department of CS, Dr.SNS.Rajalakshmi College of Arts & Science

Abstract- Clustering is a task of assigning a set of objects into groups called clusters. In general the clustering algorithms can be classified into two categories. One is hard clustering; another one is soft (fuzzy) clustering. Hard clustering, the data's are divided into distinct clusters, where each data element belongs to exactly one cluster. In soft clustering, data elements belong to more than one cluster, and associated with each element is a set of membership levels. In this paper we represent a survey on fuzzy c means clustering algorithm. These algorithms have recently been shown to produce good results in a wide variety of real world applications.

Index Terms- Soft clustering, hard clustering, FCM.

I. Introduction

Fast and robust clustering algorithms play an important role in extracting useful information in large databases. The aim of cluster analysis is to partition a set of N object into C clusters such that objects within cluster should be similar to each other and objects in different clusters are should be dissimilar with each other[1]. Clustering can be used to quantize the available data, to extract a set of cluster prototypes for the compact representation of the dataset, into homogeneous subsets.

Clustering is a mathematical tool that attempts to discover structures or certain patterns in a dataset, where the objects inside each cluster show a certain degree of similarity. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Cluster analysis is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization. It will often necessary to modify preprocessing and parameter until the result achieves the desired properties.

In Clustering, one of the most widely used algorithms is fuzzy clustering algorithms. Fuzzy set theory was first proposed by Zadeh in 1965 & it gave an idea of uncertainty of belonging which was described by a membership function. The use of fuzzy set provides imprecise class membership function. Applications of fuzzy set theory in cluster analysis were early proposed in the work of Bellman, Zadeh, and Ruspini This paper opens door step of fuzzy clustering [2]. Integration of fuzzy logic with data mining techniques has become one of the key constituents of soft computing in handling challenges posed by massive collections of natural data. The central idea in fuzzy clustering is the non-unique partitioning of the data into a collection of clusters. The data points are assigned membership values for each of the clusters and fuzzy clustering algorithm allow the clusters to grow into their natural shapes [3]. The fuzzy clustering algorithms can be divided into two types 1) Classical fuzzy clustering algorithms 2) Shape based fuzzy clustering algorithms. Classical fuzzy clustering algorithms can be divided into three types.1) The Fuzzy C-Means algorithm 2) The Gustafson-Kessel algorithm 3) The Gath-Geva algorithm. Shape based fuzzy clustering algorithm can be divided into 1) Circular shape based clustering algorithm 2) Elliptical shape based clustering algorithm 3) Generic shape based clustering algorithm. In this paper, represent a review on fuzzy c means, and extended version of fcm such as pcm, fpcm and their advantages and disadvantages of real time applications.

II. Fuzzy C Means Algorithm

Fuzzy clustering is a powerful unsupervised method for the analysis of data and construction of models. In many situations, fuzzy clustering is more natural than hard clustering. Objects on the boundaries between several classes are not forced to fully belong to one of the classes, but rather are assigned membership degrees between 0 and 1 indicating their partial membership. Fuzzy c-means algorithm is most widely used. Fuzzy c-means clustering was first reported in the literature for a special case (m=2) by Joe Dunn in 1974. The general case (for any m greater than 1) was developed by Jim Bezdek in his PhD thesis at Cornell University in 1973. It can be improved by Bezdek in 1981. The FCM employs fuzzy partitioning such that a data point can belong to all groups with different membership grades between 0 and 1.

Algorithm

1. Initialize $U=[u_{ij}]$ matrix, $U^{(0)}$
2. At k-step: calculate the centers vectors $C^{(k)}=[c_j]$ with $U^{(k)}$

$$C_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m}$$

3. Update$U^{(k)}$,$U^{(k+1)}$
4. $d_{ij} = \sqrt{\sum_{i=1}^{n}(x_i - c_i)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}}$$

5. $if \parallel U(k+1) - U(k) \parallel < \epsilon$ $then$ STOP; otherwise return to step 2.

Here m is any real number greater than 1,
u_{ij} is the degree of membership of x_i in the cluster j,
x_i is the ith of d-dimensional measured data,
c_j is the d-dimension center of the cluster,

This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center. Clearly, summation of membership of each data point should be equal to one. After each iteration membership and cluster centers are updated according to the formula.

Advantages
1) Unsupervised
2) Converges

Limitations:
1) Long computational time
2) Sensitivity to the initial guess (speed, local minima)
3) Sensitivity to noise and One expects low (or even no) membership degree for outliers (noisy points).

III. Possibilistic C-Means (PCM)

To overcome difficulties of the fcm, Krishnapuram and keller proposed a new clustering model named Possibilistic c-Means (PCM).

Algorithm

Fix the number of clusters C; fix m,1<m<∞;\
Set iteration counter l=1;
Intialize the possiblistic C-parttion $U^{(0)}$;
Estimate η_i
Repeat
Update the prototypes using $U^{(l)}$, as indicated bel ow;
Compute $U^{(l+1)}$
Increment l;
Until ($\| U^{(l-1)}-U^{(l)}\|<\varepsilon$);

{ The remaining part of algorithm is optional and to be used only when the actul shape of the generated possibility distribution is important }

Set iteration counter l=1;
Reestimate η_i
Repeat protypes using $U^{(l)}$, as indicated below;
Compute $U^{(l+1)}$
Increment l;
Until ($\| U^{(l-1)}-U^{(l)}\|<\varepsilon$);

η_i –determines distance at which the membership value of a point in a cluster becomes 0.5.

$$\eta_i = K\frac{\sum_{j=1}^{N} u_{ij}^m d_{ij}^2}{\sum_{j=1}^{N} u_{ij}^m}$$

Advantage
Clustering noisy data samples

Disadvantages
1) Very sensitive to good initialization
2) Coincident clusters may result
Because the columns and rows of the typicality matrix are independent of each other
Sometimes this could be advantageous (start with a large value of c and get less distinct clusters)

IV. Fuzzy Possibilistic C Means Algorithm (FPCM)

To overcome difficulties of the pcm, Pal defines a clustering technique that integrates the features of both Fuzzy a Possibilistic c-means called Fuzzy Possibilistic c-Means (FPCM). Membership and Typicality's are very significant for the accurate characteristic of data substructure in clustering difficulty. An objective function in the fpcm depending on both membership and typicality's are represented as::

Memberships and topicalities is represented as:

$$J_{FPCM}(U,T,V) = \sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij}^m + t_{ij}^\eta)\, d^2(x_j, v_i)$$

Which of the following constraints

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j \in \{1,\dots,n\}$$
$$\sum_{i=1}^{c} t_{ij} = 1, \forall i \in \{1,\dots,c\}$$

FPCM generates Memberships and possibilities at the same time, together with the usual point prototypes or cluster center for each cluster.

Advantage
1) Ignores the noise sensitivity deficiency of FCM
2) Overcomes the coincident clusters problem of PCM.

Disadvantages
1) The row sum constraints must be equal to one

V. Possibilistic Fuzzy C Means Algorithm (PFCM)

In fpcm, the constraint corresponding to the sum of all typicality values of all data to a cluster must be equal to one cause problems particularly for a big data set. In order to avoid this problem pal et al propose a new algorithm called Possibilistic Fuzzy c means algorithm (pfcm). The objective function is defined by

$$J_{PFCM}(U,T,V;Z) = \sum_{i=1}^{c}\sum_{j=1}^{n}(au_{ik}^m + bt_{ik}^\eta) \times \|z_k - v_i\|^2 + \sum_{i=1}^{c}\delta_i \sum_{k=1}^{N}(1 - t_{ik})^\eta$$

Subject to
$\sum_{i=1}^{c}\mu_{ik} = 1\ \forall k, 0 < \mu_{ik}, t_{ik} < 1, a > 0, b < 0, m > 1, n > 1$
, a&b define the relative importance between the membership degrees and typicality values. The objective function J_{MN} can be

minimized if $D_{ikA} = \|z_k - v_i\|_A > 0$, for every i and k,m,$\eta$>1 as well as z contains a minimum of c different data with these conditions we have (U,T^T,V)$^{\varepsilon M_{fc} \times M_{pc} \times \Re^p}$.The membership degree calculated with

$$J_{FC}(Z,U,V) = \sum_{i=1}^{c}\sum_{j=1}^{m}(u_{ij})^m \|z_k - v_i\|^2$$

$$t_{ik}=1/1+(b(\|z_k - v_i\|)^2/\delta_i)^{(1/(\eta-1)}, \ 1<i<c$$

$$v_i = (\sum_{k=1}^{N}(au_{ik}^m + bt_{ik}^\eta)Zk)/(\sum_{k=1}^{N}(au_{ik}^m + bt_{ik}^\eta)), \ 1 < i < c$$

Advantage:

1) Ignores the noise sensitivity deficiency of FCM
2) Overcomes the coincident clusters problem of PCM.
3) Eliminates the row sum constraints of FPCM

VI. CONCLUSION AND FUTURE WORK

FCM algorithm is a distinctive clustering algorithm, has been exploited in extensive range of engineering and scientific disciplines, for instance, medicine imaging, pattern detection , data mining and bioinformatics. In view of the fact, the initially developed FCM makes use of the squared-norm to determine the similarity between prototypes and data points, and it performs well only in the case of clustering spherical clusters. Furthermore, several algorithms are developed by numerous authors based on the FCM with the aim of clustering more general dataset. During the survey, we also find some points that can be further improvement in the future using advanced clustering technique to achieve more efficient accuracy in the result and reduce the time taken for data and/or information retrieval from large dataset.

REFERENCES

[1] M.S.Yang," A Survey of fuzzy clustering" Mathl. Comput. Modelling Vol. 18, No. 11, pp. 1-16, 1993.

[2] A. vathy-Fogarassy, B.Feil, J.Abonyi"Minimal Spanning Tree based Fuzzy clustering" Proceedings of World academy of Sc., Eng & Technology, vol-8, Oct-2005, 7-12.

[3] Pal N.R, Pal K, Keller J.M. and Bezdek J.C, "A Possibilistic Fuzzy c-Means Clustering Algorithm", IEEE Transactions on Fuzzy Systems, Vol. 13, No. 4, Pp. 517–530, 2005.

[4] R. Krishnapuram amd J.M. Keller, "A possibilistic approach to clustering", IEEE Trans. Fuzzy Systems, Vol. 1, Pp. 98-110, 1993.

[5] Vuda Sreenivasarao and Dr.S. Vidyavathi, "Comparative Analysis of Fuzzy C- Mean and Modified Fuzzy Possibilistic C -Mean Algorithms in Data Mining", IJCST Vol. 1, No. 1, Pp. 104-106, 2010.

[6] J. C. Dunn (1973): "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", Journal of Cybernetics 3: 32-57

[7] J. C. Bezdek (1981): "Pattern Recognition with Fuzzy Objective Function Algorithms", Plenum Press, New York.

[8] Mohamed Fadhel Saad and Adel M. Alimi," Modified Fuzzy Possibilistic C-means," Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I IMECS 2009, March 18 -20, 2009, Hong Kong.

AUTHORS

First Author – R.Suganya M.Sc., M.Phil., Assistant Professor, Dr.SNS.Rajalakshmi College of Arts & Science, Chinnavedampatti, Coimbatore., Email-id: mail2cinna.rs@gmail.com

Second Author – R.Shanthi, Research Scholar, Dr.SNS.Rajalakshmi College of Arts & Science, Chinnavedampatti, Coimbatore. , Email-id: shanthisangi@gmail.com

Journal of Information Engineering and Applications
ISSN 2224-5782 (print) ISSN 2225-0506 (online)
Vol 2, No.6, 2012

Fuzzy k-c-means Clustering Algorithm for Medical Image Segmentation

Ajala Funmilola A*, Oke O.A, Adedeji T.O, Alade O.M, Adewusi E.A

Department of Computer Science and Engineering, LAUTECH Ogbomoso, Oyo state, Nigeria.

*E-mail of the corresponding author: funfaith2003@yahoo.co.uk.

Abstract

Medical image segmentation is an initiative with tremendous usefulness. Biomedical and anatomical information are made easy to obtain as a result of success achieved in automating image segmentation. More research and work on it has enhanced more effectiveness as far as the subject is concerned. Several methods are employed for medical image segmentation such as Clustering methods, Thresholding method, Classifier, Region Growing, Deformable Model, Markov Random Model etc. This work has mainly focused attention on Clustering methods, specifically k-means and fuzzy c-means clustering algorithms. These algorithms were combined together to come up with another method called fuzzy k-c-means clustering algorithm, which has a better result in terms of time utilization. The algorithms have been implemented and tested with Magnetic Resonance Image (MRI) images of Human brain. Results have been analyzed and recorded. Some other methods were reviewed and advantages and disadvantages have been stated as unique to each. Terms which have to do with image segmentation have been defined along side with other clustering methods.

Keywords: Clustering algorithms, Fuzzy c-means, K-means, Segmentation.

1 Introduction

Diagnostic imaging is an invaluable tool in medicine today. Magnetic Resonance Imaging (MRI), Computed Tomography, Digital Mammography, and other imaging modalities provide effective means for non-invasively mapping the an atomy of a subject. These technologies have greatly increased knowledge of normal and diseased anatomy for medical research and serves as a critical component in diagnosis and treatment planning. (Dzung et. al).

Computer algorithms for the delineation of anatomical structures and other regions of interest are a key components assisting and automating specific radiological tasks. These algorithms are otherwise known as image segmentation algorithms. They are of great importance in biomedical imaging applications like tissue volume quantification, diagnosis, localization pathology, study of anatomical structures, treatment planning, partial volume correction of functional imaging data and computer integrated surgery.

2. Related work

Numerous methods are available in medical image segmentation. These methods are chosen based on the specific applications and imaging modalities. Imaging artifacts such as noise, partial volume effects, and motion can also have significant consequences on the performance of segmentation algorithms. Some of these methods with their idiosyncrasies were described below

Thresholding Method

Thresholding is the most basic of the medical image segmentation techniques. It is based on separating pixels in different classes depending on their gray level. Thresholding approaches segment scalar images by creating a binary partitioning of the image intensities. A thresholding procedure attempts to determine an intensity value, called the *threshold*, which separates the desired classes. The segmentation is then achieved by grouping all pixels with intensity greater than the threshold into one class, and all other pixels into another class. Determination of more than one threshold value is a process called multi-thresholding. Its main limitations are that in its simplest form only two classes are generated and it can not be applied to multi-channel images. In addition, thresholding typically does not take into account the spatial characteristics of an image. This causes it to be sensitive to noise and intensity in homogeneities, which can occur in magnetic resonance images.

Classifiers

Classifier methods are used in pattern recognition they seek to partition a feature space derived from the image using data with known labels. A feature space is the range space of any function of the image, with the most common feature space being the image intensities themselves.

Classifiers are known as *supervised* methods since they require training data that are manually segmented and then used as references for automatically segmenting new data.

Markov Random Field Models

Markov random field (MRF) is not a method but a statistical model that can be used within segmentation methods. MRFs are often incorporated into clustering segmentation algorithms such as the K -means algorithm under a Bayesian prior model (Pham and et al, 1998). The segmentation is then obtained by maximizing “a posteriori” probability of the segmentation given the image data using iterative methods such as iterated conditional modes or simulated annealing. A difficulty associated with MRF models is proper selection of the parameters controlling the strength of spatial interactions. Too high a setting can result in an excessively smooth segmentation and a loss of important structural details.

Artificial Neural Networks

Artificial neural networks (ANNs) are massively parallel networks of processing elements or nodes that simulate biological learning. Each node in an ANN is capable of performing elementary computations. Learning is achieved through the adaptation of weights assigned to the connections between nodes.

Because of the many interconnections used in a neural network, spatial information can easily be incorporated into its classification procedures. Although ANNs are inherently parallel, their processing is usually simulated on a standard serial computer, thus reducing its potential computational advantage (Pham and et al, 1998).

Atlas-Guided Approaches

Atlas-guided approach uses standard atlas or template is available. This it does by bringing together information about the anatomy that requires segmenting. This atlas is then used as a reference frame for segmenting new images. Conceptually, atlas-guided approaches are similar to classifiers except they are implemented in the spatial domain of the image rather than in a feature space (Dzung L. Pham and et al, 1998).

Deformable Models

Deformable models are model-based techniques which are used for delineating region boundaries by the use of closed parametric curves or surfaces. This curves or surfaces are deformed under the influence of internal or external forces. Deformable Models are physically motivated techniques. Delineation of an object boundary in an image is done by placing a closed curve or surface near the desired boundary then an iterative relaxation process is allowed to be undergone. Internal forces are computed from within the curve or surface to keep it smooth throughout the deformation. External forces are usually derived from the image to drive the curve or surface towards the desired feature of interest (Pham and et al, 1998).

Clustering Analysis

Cluster analysis or clustering is the assignment of a set of observations into subsets (called *clusters*) so that observations in the same cluster are similar in some sense (Wikipedia, 2009). Clustering is a method of unsupervised learning, and a common technique for statistical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis, information retrieval, and bioinformatics. Clustering algorithms and the classifier method are likely in function but clustering does not use training data instead they iterate between segmenting the image and characterizing the properties of each class. Consequently they are otherwise termed unsupervised methods. In a sense, clustering methods train themselves using the available data (Dzung L. Pham and et al, 1998).

Three commonly used clustering algorithms are the K-means, the fuzzy C-means algorithm, and the expectation-maximization (EM) algorithm. The K-means clustering algorithm clusters data by iteratively computing a mean intensity for each class and segmenting the image by classifying each pixel in the class with the closest mean (Dzung L. Pham and et al, 1998).

Fuzzy C-Means Clustering

Because of the advantages of magnetic resonance imaging (MRI) over other diagnostic imaging, the majority of researches in medical image segmentation pertain to its use for MR images, and there are a lot of methods available for MR image segmentation. Among them, fuzzy segmentation methods are of considerable benefits, because they could retain much more information from the original image than hard segmentation methods. In particular, the fuzzy C-means (FCM) algorithm, assign pixels to fuzzy clusters without labels. Unlike the hard clustering methods otherwise known as k-means clustering which force pixels to belong exclusively to one class, FCM allows pixels to belong to multiple clusters with varying degrees of membership. Because of the additional flexibility, The Fuzzy

C-means clustering algorithm (FCM) is a soft segmentation method that has been used extensively for segmentation of MR images applications recently. However, its main disadvantages include its computational complexity and the fact that the performance degrades significantly with increased noise (NG and et al, 2006).

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. In other word, each point has a degree of belonging to clusters, as in fuzzy logic, rather than belonging completely to one cluster. Thus, points on the edge of a cluster may be in the cluster to a lesser degree than points in the center of cluster. Fuzzy c-means has been a very important tool for image processing in clustering objects in an image. In the 70's, mathematicians introduced the spatial term into the FCM algorithm to improve the accuracy of clustering under noise. (Wikipedia 2009)

K-Means Clustering

K-means (MacQueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. K-means clustering algorithm is a simple clustering method with low computational complexity as compared to FCM. The clusters produced by K-means clustering do not overlap.

The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early grouping is done. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

K-means clustering algorithm is an unsupervised method. It is used because it is simple and has relatively low computational complexity. In addition, it is suitable for biomedical image segmentation as the number of clusters (K) is usually known for images of particular regions of human anatomy. For example a MR image of the head generally consists of regions representing the bone, soft tissue, fat and background. Since the regions are 4 in number then K will be 4. Finally, this algorithm aims at minimizing an objective function, in this case a squared error function. The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2$$

where $||x_i^{(j)} - c_j||^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the *n* data points from their respective cluster centres. K-means is a simple algorithm that has been adapted to many problem domains. It is a good candidate for extension to work with fuzzy feature vectors.

3 METHODOLOGY

For both clustering methods chosen in this project algorithms and flowcharts have been provided for the proper implementation. These algorithms have been further combined to formulate another called fuzzy k-c-means algorithm. The clustering methods have been compared on the bases of the time it takes each to segment a given image, the number of iteration, and as well as how accurate the result is.

Fuzzy C-Means Algorithm and Flowchart

Fuzzy c-means algorithm allows data to belong to two or more clusters with different membership coefficient. Fuzzy C-Means clustering is an iterative process. First, the initial fuzzy partition matrix is generated and the initial fuzzy cluster centers are calculated. In each step of the iteration, the cluster centers and the membership grade point are updated and the objective function is minimized to find the best location for the clusters. The process stops when the maximum number of iterations is reached, or when the objective function improvement between two consecutive iterations is less than the minimum amount of improvement specified.

Moreover the update in the iteration is done using the membership degree as well as the centre of the cluster that is the two parameter change as the steps are being repeated until a set point called the threshold is reached or the process stops when the maximum number of iterations is reached, or when the objective function improvement between two consecutive iterations is less than the minimum amount of improvement specified. In addition a fuzziness coefficient 'm' is chosen which may be any real number greater than 1.

The algorithm comprises of the following steps:

1. Read the image into the Matlab environment
2. Try to identify the number of iteration it might possibly do within a given period of time.
3. Get the size of the image.
4. Calculate the distance possible size using repeating structure.
5. Concatenate the given dimension for the image size
6. Repeat the matrix to generate large data items in carrying out possibly distance calculation.
7. Begin Iterations by identifying large component of data vis - a - vis the value of the pixel.
8. Stop Iteration when possible identification elapses.
9. Generate the time taken to segment.

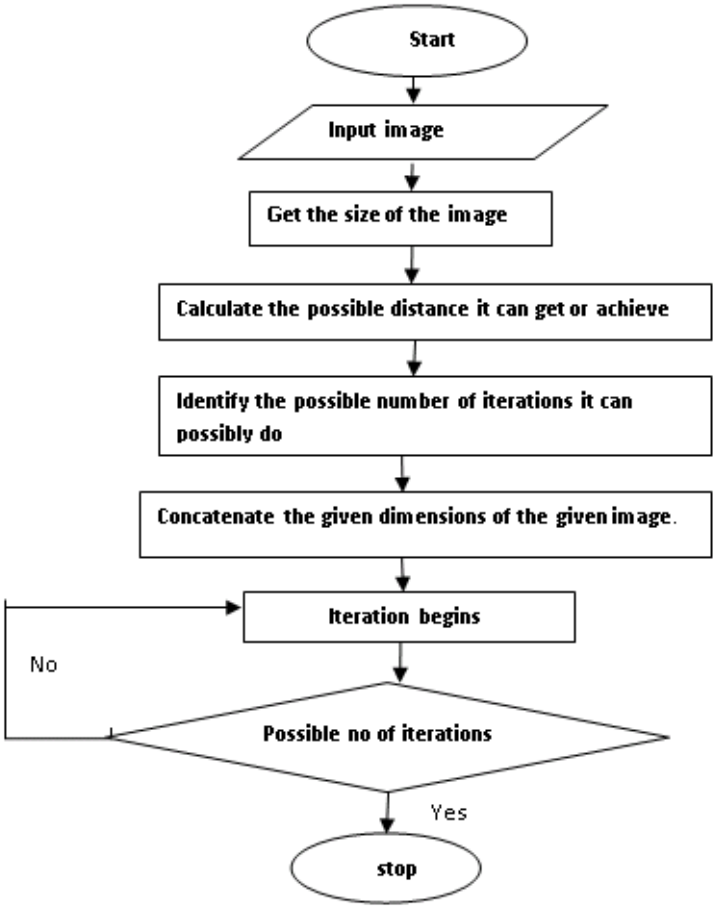


Figure 1: FCM Flowchart

K-Means Algorithm and Flowchart

The k-means clustering also known as hard c-means clustering provides an algorithm used for partitioning a set of N vectors into C groups. The algorithm computes the cluster centers (centroids) for each group. This algorithm minimizes a dissimilarity function. The image to work with is first imputed into the MATLAB work area with the use of the function called imread. This is followed by the calculation of the colour space by the use of the L*b*a* colour space derived from the CIE XYZ tri-stimulus values. The L*a*b* space consists of a luminosity layer 'L*', chromaticity-layer 'a*' indicating where colour falls along the red-green axis, and chromaticity-layer 'b*' indicating where the colour falls along the blue-yellow axis. All of the colour information is in the 'a*' and 'b*' layers. You can measure the difference between two colours using the Euclidean distance metric.

Classification of the colours generated in a*b* space is also a very important part of the implementation, k-means clustering makes this possible. Clustering is a way to separate groups of objects. K-means clustering treats each object as having a location in space. It finds partitions such that objects within each cluster are as close to each other as possible, and as far from objects in other clusters as possible. K-means clustering requires that you specify the number

of clusters to be partitioned and a distance metric to quantify how close two objects are to each other. Moreover, an image is made up of pixels. These pixels are labeled using the result from k-means. For every object in an input, K-means returns an index corresponding to a cluster. The cluster center output from K-means will be used later in the demo. Label every pixel in the image with its cluster index. Finally, the images generated through the segmentation of the original image are created for analysis.

The algorithm has the following steps:

1. Read the image into the MATLAB environment using the imread function
2. Convert the image to L*a*b* colour space using make form and apply form
3. Classify the Colours in 'a*b*' Space Using K-Means Clustering
4. Label every pixel in the Image using the results from K –means
5. Create Images that Segment the H&E Image by colour using clusters.

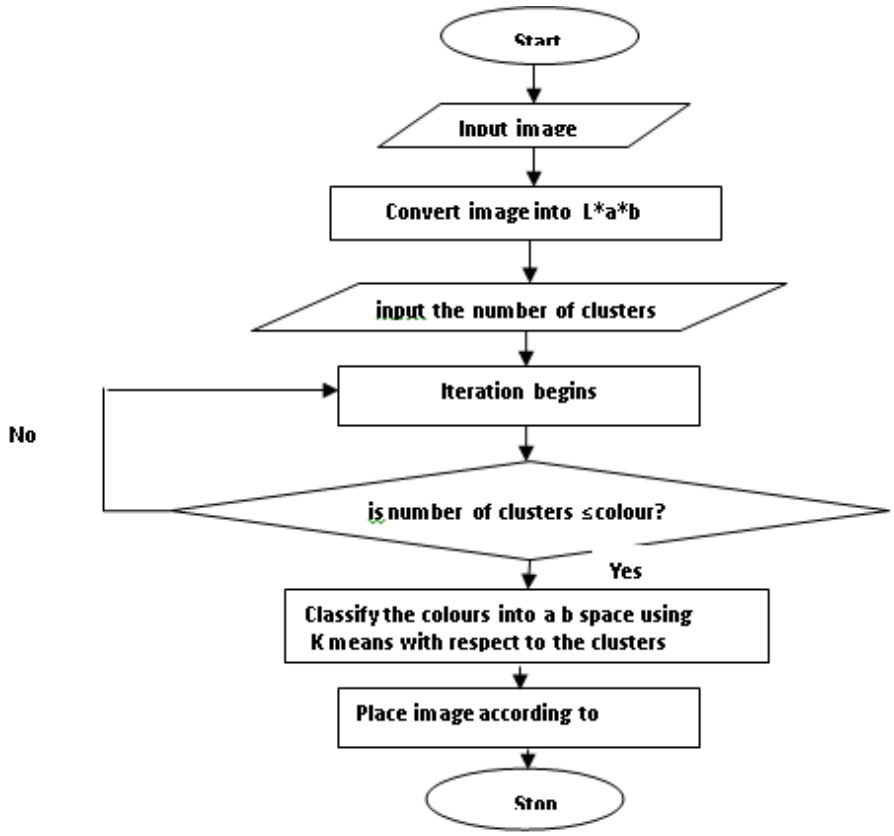


Figure 2: K-Means Flowchart

Fuzzy K-C-Mean Algorithm and Flowchart

In Fuzzy K-C-Means the interest is on making the number of iterations equal to that of the fuzzy c means, and still get an optimum result. This implies that irrespective of the lower number of iteration, we will still get an accurate result.

The algorithm has the following steps:

1. Read the image into the Matlab environment
2. Try to identify the number of iteration it might possibly do within a given period of time
3. Reduce number of iteration with distance check
4. Get the size of the image
5. Calculate the distance possible size using repeating structure
6. Concatenate the given dimension for the image size
7. Repeat the matrix to generate large data items in carrying out possibly distance calculation
8. Reduce repeating when possible distance has been attained
9. Iterations begin by identifying large component of data vis a vis the value of the pixel
10. Iteration stops when possible identification elapses
11. Time is generated.

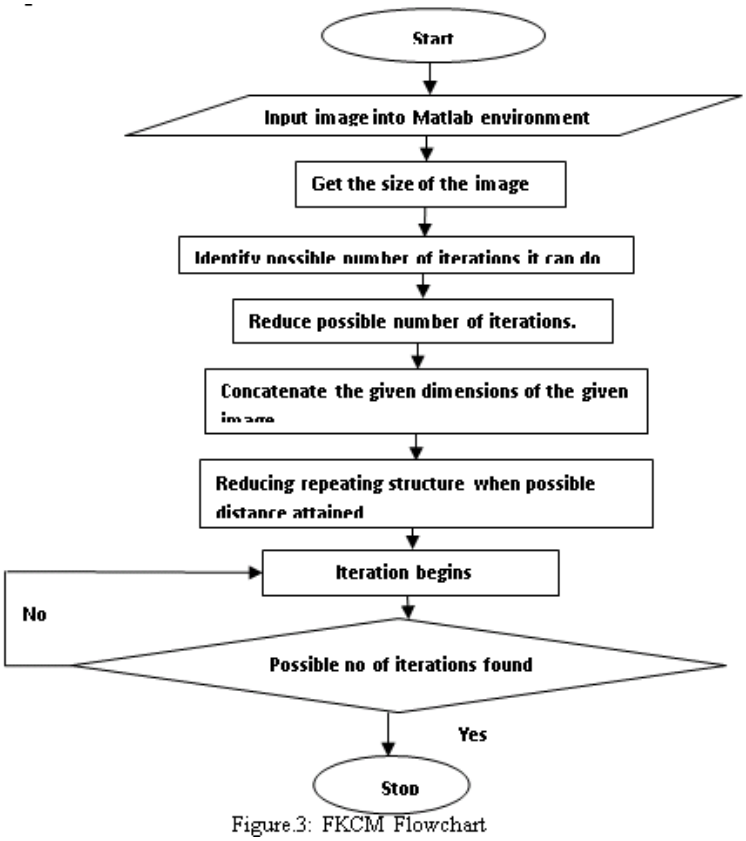


Figure.3: FKCM Flowchart

4. Results and Discussion

The implemented clustering methods have been done in MATLAB. Three images acquired through Magnetic Resonance Imaging (MRI) were used for comparing the performances of the three methods. The machine on which they are tested is made up of the following: Pentium (R) M, Processor speed of 1400 MHz, 512 MB of RAM. The following are the benchmarks used to compare:

- The mode of operation
- The time taken
- The accuracy

Mode of Operation

K-means demands that the user specifies the number of clusters before the segmentation commences. As a result, the number of clusters is predetermined. The k-means method considered here is operating based on colours contained by the image. The number of clusters specified by the user must correspond to the number of colour. It is not necessary to have the pre-knowledge of the number of colours contained by the image because there is provision made for re-inputting the number of clusters. Maximum number of possible colours provided for is 9 since most images may have as much as 5-6 colours. It is possible to have an image whose colours are more than this range, hence the provision for more colours. As soon as k-means gets to the end of the clusters specified it stops.

Fuzzy C-Means converts a coloured image into grey scale before commencing the segmentation. That is it segments using grey scale. If the image inputted is a non-coloured it will still segment it unlike the k-means which only segments a coloured image. Usually, Fuzzy C-means iterates based on the number of clusters it comes across on the image being considered. Unlike K-means, the fuzzy c-means will return the number of clusters after the segmentation has been done. Therefore the number clusters is approximately the number of iterations.

Fuzzy K-C-Means is a method generated from both fuzzy c-means and k-means but it carries more of fuzzy c-means properties than that of k-means. Fuzzy k-c-means works on grey scale images like fuzzy c-means, generates the same number of iterations as in fuzzy c-means.

Time Taken to segment

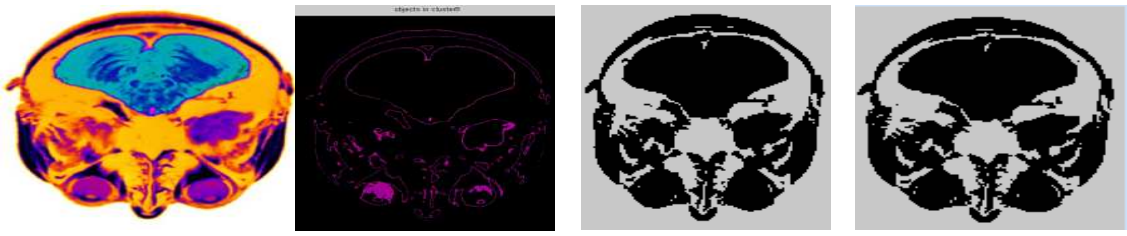
Based on the tested images k-means appears to be faster than fuzzy c-means while in some cases fuzzy c-means also appears to be faster than k-means. Whereas both fuzzy c-means and k-means are competing in terms of time, fuzzy k-c-means has been programmed to generate the same number of iteration with fuzzy c-means with a faster operation time. That is fuzzy k-c-means is faster than both fuzzy c-means and k-means. The conflict in time between fuzzy c-means and k-means is assumed to account from the properties of the image under consideration, the efficiency of the machine on which the methods are tested.

Accuracy

In terms of accuracy, the number iteration is put into consideration. The more the iterations the more the accuracy. The iteration that k-means can perform depend largely on the number of colours contained by an image which make its iterative ability limited unlike that of fuzzy c-means and fuzzy k-c-means which segment based on the number of iterations or clusters contained in an image. Consequent to this, k-means is less accurate than the other two methods.

Segmentation results on MRI brain using the Methods

K-means, Fuzzy c-means and Fuzzy k-c-means have been used in segmenting three MRI images in order to compare the results in each case.



(a) (b) (c)

(d)

Figure 4 (a) Image I, segmentation results; (b) K-Means (c) Fuzzy C-Means (d) Fuzzy K-C Means

Table 1: Comparison of segmentation results on image I

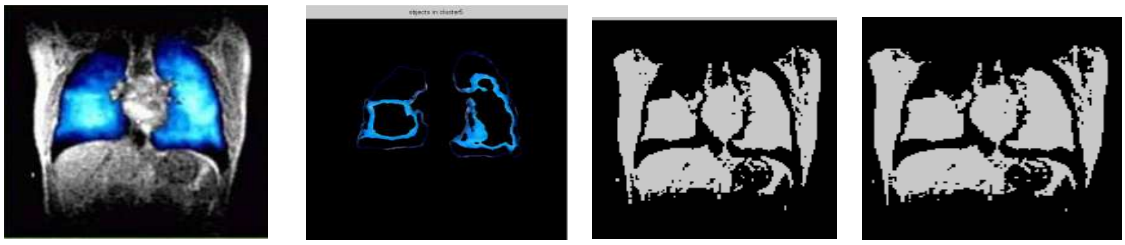
METHODS	TIME TAKEN (s)	NUMBER OF CLUSTERS	NUMBER OF ITERATION
K-MEANS	31.67	9	9
FUZZY C-MEANS	32.45	13	13
FUZZY K-C-MEANS	30.09	13	13

(a) (b) (c) (d)

Figure 5 (a) Image II, segmentation results; (b) K-Means (c) Fuzzy C-Means (d) Fuzzy K-C Means

Table 2: Comparison of segmentation results on image II

METHODS	TIME TAKEN (s)	NUMBER OF CLUSTERS	NUMBER OF ITERATIONS
K-MEANS	31.23	9	9
FUZZY C-MEANS	12.06	11	11
FUZZY K-C-MEANS	10.58	11	11



(a) (b) (c) (d)

Figure 6 (a) Image III, segmentation results; (b) K-Means (c) Fuzzy C-Means (d) Fuzzy K-C Means

Table 3: Comparison of segmentation results on image II

METHODS	TIME TAKEN (IN SECONDS)	NUMBER OF CLUSTERS	NUMBER OF ITERATIONS
K-MEANS	38.44	5	5
FUZZY C-MEANS	37.56	13	13
FUZZY K-C-MEANS	35.68	13	13

It is pertinent to note in this work that;

i. for k-means the user is expected to input the number of clusters before segmentation and this is equal to the number of iterations.

ii. for fuzzy c-means number of clusters is generated iteratively during segmentation and this is equal to the number of clusters.

iii. for fuzzy k-c-means number of iterations is equal to number of clusters.

The method with the highest iteration value and segments within the shortest period of time takes the more accuracy. In this case fuzzy k-c-means and fuzzy c-means should have been considered but with clear observation fuzzy c-means is slower than fuzzy k-c-means therefore fuzzy k-c-means takes the highest accuracy.

5. Conclusion

Medical image segmentation is a case study that is fascinating and very important as well. Fuzzy C-Means, K-Means and Fuzzy K-C-Means clustering algorithms have been considered so far they have been seen effective in the image segmentation. They are easy to use unlike some other methods in existence. Time, accuracy, and iterations have been the major focus here. But there are still limitations that like k-means segmenting with predetermined number of clusters Fuzzy C-means generating an overlapping results and not being able to segment coloured images until they are converted into grey scale. Fuzzy K-C-Means also operates almost like Fuzzy C-Means.

6. References

Ng, H.P. Ong, S.H, Foong, K.W.C, Nowinski, W.L. (2005): "An improved watershed algorithm for medical image segmentation", Proceedings 12th International Conference on Biomedical Engineering.

MacQueen (1967): "Some methods for classification and analysis of multivariate observations", Proceedings 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281-297.

Dzung L. Pham, Chenyang Xu, Jerry L. Prince (2010) " A survey of current methods in medical image segmentation, Journal of image processing

Heyer, Kruglyak, Yooseph (1999) "Quality Threshold Clustering"

This academic article was published by The International Institute for Science, Technology and Education (IISTE). The IISTE is a pioneer in the Open Access Publishing service based in the U.S. and Europe. The aim of the institute is Accelerating Global Knowledge Sharing.

More information about the publisher can be found in the IISTE's homepage:
http://www.iiste.org

The IISTE is currently hosting more than 30 peer-reviewed academic journals and collaborating with academic institutions around the world. **Prospective authors of IISTE journals can find the submission instruction on the following page:** http://www.iiste.org/Journals/

The IISTE editorial team promises to the review and publish all the qualified submissions in a fast manner. All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Printed version of the journals is also available upon request of readers and authors.

IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar

ISSN:2442-5567 | E-ISSN:2443-289X

PENENTUAN PENERIMA BEASISWA DENGAN ALGORITMA *FUZZY C-MEANS* DI UNIVERSITAS MEGOW PAK TULANG BAWANG

Muhardi[1] Nisar[2]

[1]Universitas Megow Pak [2]MTI Institut Informatika & Bisnis Darmajaya
Email: ardyardy68@yahoo.co.id

ABSTRAK

Beasiswa adalah pemberian berupa bantuan keuangan yang diberikan kepada perorangan yang bertujuan untuk digunakan demi keberlangsungan pendidikan yang ditempuh. Penentuan beasiswa dapat dikelompokan berdasarkan kriteria Indeks Prestasi Kumulatif (IPK), tingkat kemiskinan, tanggungan orang tua dan prestasi mahasiswa dalam proses rekruitmen beasiswa. Algoritma *Fuzzy C-Means* merupakan satu algoritma yang mudah dan sering digunakan dalam pengelompokan data karena membuat suatu perkiraan yang efisien dan tidak memerlukan banyak parameter. Beberapa penelitian telah menghasilkan kesimpulan bahwa metode *Fuzzy C-Means* dapat digunakan untuk mengelompokkan data berdasarkan atribut-atribut tertentu. Penerapan algoritma *Fuzzy C-Means* dalam penentuan beasiswa di kelompokan menjadi tiga *cluster* yaitu menerima, dipertimbangkan dan tidak berhak menerima beasiswa , sempel data sebenyak 75 data mahasiswa diperoleh tiga *cluster* berdasarkan nilai rata-rata penentuan beasiswa kemudian setiap *cluster* diklasifikasikan berdasarkan kriteria mana yang lebih diprioritaskan dengan nilai terbesar pada jarak akhir merupakan *cluster* yang menerima beasiswa, sedangkan *cluster* dengan nilai terkecil merupakan *cluster* yang tidak berhak menerima beasiswa.

Kata Kunci: *Beasiswa, clustering, Fuzzy C- Means.*

ABSTRACT

The scholarship is the grant in the form of financial assistance given to an individual who intends to use for the sake of the continuity of education traveled. Determination of the scholarship can be grouped based on the criteria of the cumulative Achievement Index (IPK), the level of poverty, dependent parent and student achievement in the process of recruitment of scholarship. Fuzzy C-Means algorithm is an algorithm that is easily and often used in grouping data because making an estimate of efficient and doesn't require a lot of parameters. Some research has resulted in the conclusion that the Fuzzy C-Means method can be used to classify data based on certain attributes. Application of Fuzzy C-Means algorithm in determination of scholarship in kelompokan into three clusters, namely receiving, considering and not eligible to receive the scholarship, sempel 75 student data sebenyak data retrieved three clusters based on the average value of the determination of a scholarship then each cluster are classified according to criteria which is more prioritized with the greatest value at the end of the cluster is receiving a scholarship, whereas the smallest value cluster is a cluster that is not eligible to receive the scholarship.

Keywords*: scholarship, clustering, Fuzzy C- Means*

ISSN:2442-5567 | E-ISSN:2443-289X

1. PENDAHULUAN

Beasiswa adalah pemberian berupa bantuan keuangan yang diberikan kepada perorangan yang diberikan kapada perorangan yang bertujuan untuk digunakan demi keberlansungan pendidikan yang ditempuh. Pemberian beasiswa merupakan program kerja yang ada di setiap universitas atau perguruan tinggi. Program beasiswa diadakan untuk meringankan beban mahasiswa dalam menempuh masa studi kuliah khususnya dalam masalah biaya. Pemberian beasiswa kepada mahasiswa dilakukan secara selektif sesuai dengan jenis beasiswa yang diadakan. Univeritas Megow Pak Tulang Bawang menyediakan beberpa program beasiswa, sebagai contoh yaitu beasiswa Peningkatan Akademik (PPA), Beasiswa Bantuan Belajar Mahasiswa (BBM) dan lain sebagainya. Indeks prestasi kumulatif, Tingkat Kemiskinan, tanggungan orang tua dan prestasi menjadi kriteria dalam proses rekruitmen beasiswa.

Proses seleksi penerimaan beasiswa secara manual yaitu dengan menginputkan satu persatu data mahasiswa ke dalam file *excel* kemudian melakukan *sorting* data mahasiswa seringkali melakukan beberapa permasalahan, antara lain membutuhkan waktu yang lama dan ketelitian yang tinggi. Selain itu, transparansi serta ketidak jelasan metodologi yang digunakan dalam proses komputasi penerimaan beasiswa juga menjadi salah satu permasalahan, sehingga dibutuhkan suatu sistem yang dapat membantu dalam proses pengambilan keputusan siapa saja mahasiswa yang menerima beasiswa berdasarkan kriteria-kriteria yang telah ditentukan secara capat dan tepat sasaran.

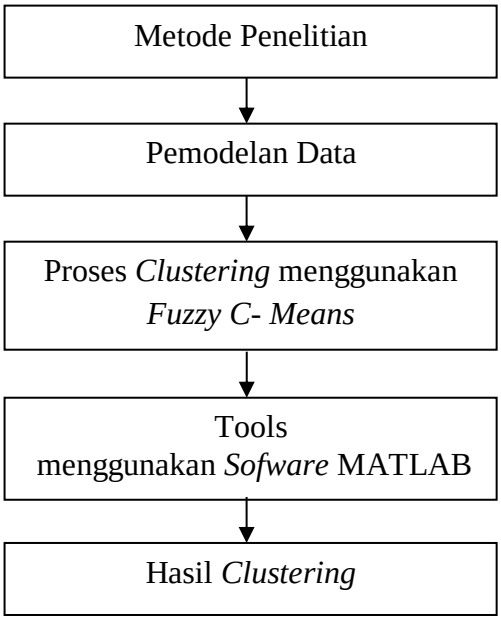
Salah satu metode yang dapat digunakan untuk menyelesaikan permasalahan tersebut adalah metode *Fuzzy C-Means.* Algoritma *Fuzzy C-Means* merupakan satu algoritma yang mudah dan sering digunakan dalam pengelompokan data kerana membuat suatu perkiraan yang efsien dan tidak memerlukan banyak parameter. Beberapa penelitian telah menghasilkan kesimpulan bahwa metode *Fuzzy CMeans* dapat digunakan untuk mengelompokkan data berdasarkan atribut-atribut tertentu. Pada kasus penelitian ini akan menganalisis penerapan metode *Fuzzy C-Means* untuk mengelompokkan data mahasiswa berdasarkan kemampuan mahasiswa dibidang akademik untuk proses penentuan beasiswa.
Pada penelitian sebelumnya, Bahari (2011) melakukan Penerapan metode *Fuzzy C-Means* untuk mengelompokan data siswa dalam penentuan jurusan di Sekolah Menengah Atas pada 81 sampel data siswa yang diuji dalam penelitian menunjukan bahwa metode *Fuzzy C-Means* memiliki tingkat akurasi yang lebih tinggi (rata-rata 78,39%), jika dibandingkan dengan metode penentuan jurusan secara manual yang di lakukan (hanya memiliki tingkat akurasi rata-rata 56,17 %).

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitan

Penelitian ini akan dilaksanakan berdasarkan rancangan penelitian seperti yang di tunjukan pada Gambar 1.

ISSN:2442-5567 | E-ISSN:2443-289X



Gambar 1. Alur Kerangka Penelitian

Penelitian yang dilaksanakan adalah jenis penelitian eksprimen, yaitu melakukan pengujian tingkat akurasi algoritma *Fuzzy C-Means clustering* sebagai model untuk menentukan penerima beasiswa. Data eksperimen diambil dari tempat penelitian yaitu Universitas Megow Pak Tulang Bawang.

2.2 Pengumpulan Data

Pengumpulan data digunakan untuk mengumpulkan data-data dan informasi-informasi yang diperlukan dalam pembuatan sistem pendukung keputusan. Pengumpulan data pada penelitian ini menggunakan metode pengumpulan data *study literature* dan telaah dokumen.

a. *Study Literature*

Study literature dilakukan dengan cara mencari bahan materi yang berhubungan dengan permasalahan, algoritma *Fuzzy C-Means*, sistem pendukung keputusan dan beasiswa, guna mempermudah proses implementasi sistem. Pencarian materi dilakukan melalui pencarian di buku panduan dan internet.

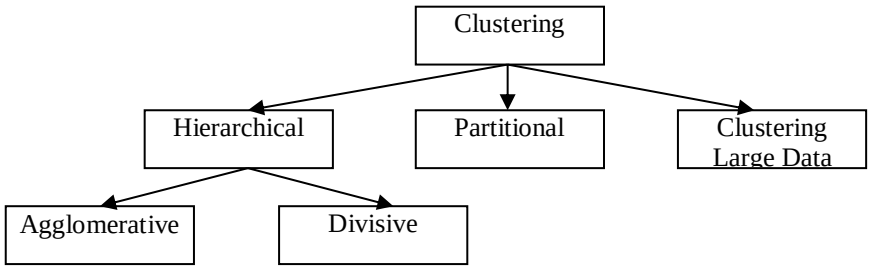
b. Telaah Dokumen

Telaah dokumen adalah pengumpulan data dengan cara mengumpulkan dan mempelajari dokumen-dokumen yang didapatkan dari pihak Jurusan Informaitka FMIPA Universitas Megow Pak Tulang Bawang. Dari metode pengumpulan data ini diperoleh 75 data mahasiswa.

2.3 Algoritma *Clustering*

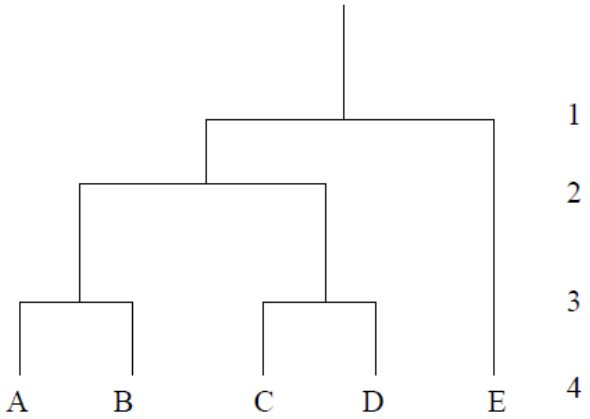
Secara umum pembagian algoritma *clustering* dapat digambarkan sebagai berikut:

ISSN:2442-5567 | E-ISSN:2443-289X



Gambar 2. Kategori Algoritma *Clustering*

Hierarchical clustering menentukan sendiri jumlah *cluster* yang dihasilkan. Hasil dari metode ini adalah suatu struktur data berbentuk pohon yang disebut dendogram dimana data dikelompokkan secara bertingkat dari yang paling bawah dimana tiap *instance* data merupakan satu *cluster* sendiri, hingga tingkat paling atas dimana keseluruhan data membentuk satu *cluster* besar berisi *cluster-cluster* seperti gambar 2.



Gambar 3. *Dendogram*

Divisive hierarchical clustering mengelompokkan data dari kelompok yang terbesar hingga ke kelompok yang terkecil, yaitu masing-masing *instance* dari kelompok data tersebut. Sebaliknya, *agglomerative hierarchical clustering* mulai mengelompokkan data dari kelompok yang terkecil hingga kelompok yang terbesar. Beberapa algoritma yang menggunakan metode ini adalah: *RObust Clustering Using LinKs* (ROCK), *Chameleon, Cobweb, Shared Nearest Neighbor* (SNN).

Partitional clustering yang mengelompokkan data ke dalam k *cluster* dimana k adalah banyaknya *cluster* dari input user. Kategori ini biasanya memerlukan pengetahuan yang cukup mendalam tentang data dan proses bisnis yang memanfaatkannya untuk mendapatkan kisaran nilai input yang sesuai. Beberapa algoritma yang masuk dalam kategori ini antara lain: *K-Means, Fuzzy C-Means, Clustering Large Aplications* (CLARA), *Expectation Maximation* (EM), *Bond Energy Algorithm* (BEA), *algoritma Genetika*, Jaringan Saraf Tiruan.

ISSN:2442-5567 | E-ISSN:2443-289X

a. Algoritma *Fuzzy Clustering C-Means* (FCM)

Pada proses pengklasteran (*clustering*) secara klasik (misalnya pada algoritma *Clustering K-Means*), pembentukan partisi dilakukan sedemikian rupa sehingga setiap obyek berada tepat pada satu partisi. Namun, adakalanya tidak dapat menempatkan suatu obyek tepat pada suatu partisi, karena sebenarnya obyek tersebut terletak di antara 2 atau lebih partisi yang lain. Pada logika *fuzzy*, metode yang dapat digunakan untuk melakukan pengelompokan sejumlah data dikenal dengan nama *fuzzy clustering*. *Fuzzy Clustering* lebih alami jika dibandingkan dengan pengklasteran secara klasik. Suatu algoritma *clustering* dikatakan sebagai *fuzzy clustering* jika algoritma tersebut menggunakan parameter strategi adaptasi secara *soft competitive*. Sebagian besar algoritma *fuzzy clustering* didasarkan atas optimasi fungsi obyektif atau modifikasi dari fungsi obyektif tersebut.

Salah satu teknik *fuzzy clustering* adalah *Fuzzy C-Means* (FCM). FCM adalah suatu teknik pengklasteran data yang keberadaan tiap-tiap data dalam suatu *cluster* ditentukan oleh nilai/derajat keanggotaan tertentu. Teknik ini pertama kali diperkenalkan oleh Jim Bezdek pada tahun 1981. Berbeda dengan teknik pengklasteran secara klasik (dimana suatu obyek hanya akan menjadi anggota suatu klaster tertentu), dalam FCM setiap data bisa menjadi anggota dari beberapa *cluster*.Batas-batas *cluster* dalam FCM adalah lunak (*soft*). Konsep dasar FCM, pertama kali adalah menentukan pusat *cluster* yang akan menandai lokasi rata-rata untuk tiaptiap *cluster*. Pada kondisi awal, pusat *cluster* ini masih belum akurat. Tiap-tiap data memiliki derajat keanggotaan untuk tiap-tiap *cluster*. Dengan cara memperbaiki pusat *cluster* dan nilai keanggotaan tiap-tiap data secara berulang, maa akan terlihat bahwa pusat *cluster* akan bergerak menuju lokasi yang tepat. Perulangan ini didasarkan pada minimasi fungsi obyektif. Fungsi Obyektif yang digunakan pada FCM adalah:

$$J_w(U,V;X) = \sum_{k-1}^{n} \sum_{i-=1}^{c} (\mu_{ik})^w (d_{ik})^2 \quad (1)$$

dengan + ∈ ,1,∞),

$$d_{ik} = d(x_k - v_i) = [\sum_{j=1}^{m} (x_{kj} - x_{ij})]^{\frac{1}{2}} \quad (2)$$

x adalah data yang akan diklaster:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix} \quad (3)$$

dan v adalah matriks pusat *cluster* :

$$V = \begin{bmatrix} v_{11} & \cdots & v_{1m} \\ \vdots & & \vdots \\ v_{m1} & \cdots & v_{mm} \end{bmatrix} \quad (4)$$

nilai Jw terkecil adalah yang terbaik, sehingga:

$$J_w^* (U^*, U^*; x) = \min J(U, U; x) \quad (5)$$

Jika $d_{ik} > 0, \forall i, k$; $w > 1$ dan X setidaknya memiliki m elemen, maka (u, v) ∈ $M_{fm} > 0, \forall i, k$; w dapat meminimasi Jw hanya jika:

$$\mu_{ik}=\frac{[\sum_{j=1}^{m}(x_{ij}-v_{kj})^{2}]^{\frac{-1}{w-1}}}{\sum_{i=1}^{n}[\sum_{j=1}^{m}(x_{ij}-v_{kj})^{2}]^{\frac{-1}{w-1}}}; \quad (6)$$

dan

$$v_{kj}=\frac{\sum_{i=1}^{n}((\mu_{ik})^{w}*X_{ij})}{\sum_{i=1}^{n}(\mu_{ik})^{w}};1 \leq i \leq m;1 \leq J \leq m \quad (7)$$

Algoritma *Fuzzy C-Means* (FCM) diberikan sebagai berikut :

1. Menentukan data yang akan di cluster X, berupa matriks berukuran n x m (n=jumlah sampel data, m = atribut setiap data). Xij=data sampel ke-I (i=1,2,…,n), atribut ke-j (j=1,2,…,m).
2. Menentukan :
   - Jumlah cluster = c
   - Pangkat = w
   - Maksimum interasi = MaxIter
   - Error terkecil yang diharapkan = ξ
   - Fungsi objektif awal = $P_0 = 0$
   - Interasi awal = t = 1

3. Membangkitkan bilangan random μik, i=1,2,3 …,n; k=1,2,3…c; sebagai elemen-elemen matriks partisi awal U.
   Menghitung jumlah setiap kolom :

$$Q_i=\sum_{k=1}^{c}\mu_{ik} \quad (8)$$

dengan j=1,2,…n.

Menghitung :

$$\mu_{ik}=\frac{\mu_{ik}}{Q_i} \quad (9)$$

4. Menghitung pusat cluster ke-k: V_{kj}, dengan k=1,2,…c; dan j=1,2,…m

$$V_{kj}=\frac{\sum_{i=1}^{n}((\mu_{ik})^{w}*X_{ij}}{\sum_{i=1}^{n}((\mu_{ik})^{w}} \quad (10)$$

5. Menghitung fungsi objektif pada interasi ke-t :

$$P_t=\sum_{i=1}^{n}\sum_{k}^{c}([\sum_{j=1}^{m}(X_{ij}-V_{kj})^{2}](\mu_{ik})^{w}) \quad (11)$$

6. Menghitung perubahan matriks partisi :

$$\mu_{ik}=\frac{[\sum_{j=1}^{m}(X_{ij}-V_{kj})^{2}]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}[\sum_{j=1}^{m}(X_{ij}-V_{kj})^{2}]^{\frac{-1}{w-1}}} \quad (12)$$

b. Proses *Clustering* menggunakan *Fuzzy C-Means*

Tahap ini akan diterapkan metode *Fuzzy C-Means* untuk mengelompokkan data. Hasil pengelompokkan ini kemudian akan digunakan untuk pertimbangan menentukan

ISSN:2442-5567 | E-ISSN:2443-289X

mahasiswa yang berhak menerima beasiswa. Adapun algoritma *C-Means Clustering* pada penerima beasiswa adalah sebagai berikut :

1. Menetapkan matriks partisi awal U berupa matriks berukuran n x m (n adalah jumlah sampel data, yaitu=75, dan m adalah parameter/atribut setiap data, yaitu=2). X_{ij}=data sampel ke-i (i=1,2,…,n), atribut ke-j (j=1,2,..,m).
2. Menentukan Nilai Parameter Awal :
   - Jumlah *cluster* (c) = 3
   - Pangkat/bobot (w) = 2
   - Maksimum interasi (MaxIter) = 100
   - Error terkecil yang diharapkan (ξ)= 10^{-5}
   - Fungsi Objektif awal (Po) = 0
   - Interasi awal (t) = 1
3. Membangkitkan bilangan random μik, i=1,2,…c; sebagai elemen-elemen matriks partisi awal (U).

3. HASIL DAN PEMBAHASAN

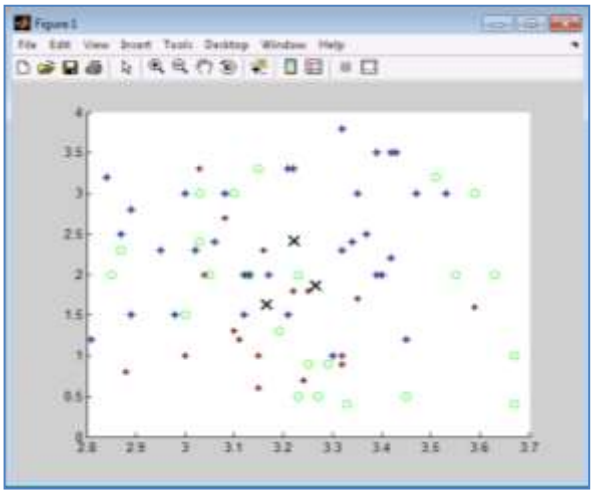
3.1 Tools menggunakan *Sofware* MATLAB

Data yang digunakan dalam penelitian ini adalah data penentuan beasiswa sebanyak 75 mahasiswa. Dalam proses *clustering*, IPK dijadikan sebagai Xi1, penghasilan orang tua dijadikan Xi2, tanggungan orang tua dijadikan Xi3 dan prestasi dijadikan Xi4. Pengolahan data pada penelitian ini menggunakan bantuan program yaitu Matlab. Dari *clustering* yang dilakukan diperoleh hasil yaitu nilai fungsi obyektif selama iterasi, pusat *cluster* atau *center* serta derajat keanggotaan lulusan untuk setiap *cluster* pada iterasi terakhir. Dalam penelitian ini, proses iterasinya berhenti pada iterasi ke-37 karena nilai |Pt - Pt-1|< ξ. Nilai fungsi obyektif pada iterasi terakhir yang diperoleh adalah 111.949781. Untuk penelitian ini *clustering* menggunakan *sofware* Matlab denga menjalankan di *Command Window* yaitu: >> fcm_k3.m

```
Iteration count = 1, obj. fcn = 185.869420
Iteration count = 2, obj. fcn = 146.530736
Iteration count = 3, obj. fcn = 141.726376
Iteration count = 4, obj. fcn = 134.833704
Iteration count = 5, obj. fcn = 130.961108
Iteration count = 6, obj. fcn = 129.145894
Iteration count = 7, obj. fcn = 127.219322
Iteration count = 8, obj. fcn = 124.903487
Iteration count = 9, obj. fcn = 122.880711
Iteration count = 10, obj. fcn = 121.359551
Iteration count = 11, obj. fcn = 119.898753
Iteration count = 12, obj. fcn = 118.236843
Iteration count = 13, obj. fcn = 116.491898
Iteration count = 14, obj. fcn = 114.963865
Iteration count = 15, obj. fcn = 113.865257
Iteration count = 16, obj. fcn = 113.180881
Iteration count = 17, obj. fcn = 112.770379
Iteration count = 18, obj. fcn = 112.513345
Iteration count = 19, obj. fcn = 112.342297
Iteration count = 20, obj. fcn = 112.223512
Iteration count = 21, obj. fcn = 112.139405
```

```
Iteration count = 22, obj. fcn = 112.079712
Iteration count = 23, obj. fcn = 112.037682
Iteration count = 24, obj. fcn = 112.008476
Iteration count = 25, obj. fcn = 111.988487
Iteration count = 26, obj. fcn = 111.975011
Iteration count = 27, obj. fcn = 111.966053
Iteration count = 28, obj. fcn = 111.960170
Iteration count = 29, obj. fcn = 111.956348
Iteration count = 30, obj. fcn = 111.953887
Iteration count = 31, obj. fcn = 111.952314
Iteration count = 32, obj. fcn = 111.951315
Iteration count = 33, obj. fcn = 111.950683
Iteration count = 34, obj. fcn = 111.950285
Iteration count = 35, obj. fcn = 111.950035
Iteration count = 36, obj. fcn = 111.949879
Iteration count = 37, obj. fcn = 111.949781
C =
   3.2649   1.8625   2.2484   4.2701
   3.1653   1.6271   5.1550   1.1770
   3.2225   2.4133   3.0759   0.7967
```

Penyebaran masing-masing anggota *cluster* pada iterasi terakhir dapat dilihat pada *cluster interface* gambar dibawah ini :



Gambar 4. Posisi *Claster* Iterasi terakhir

Keterangan :
Tanda * = *cluster* 1
Tanda + = *cluster* 2
Tanda ○ = *cluster* 3
Tanda X = pusat ke 3 *cluster*

Detail Matriks partisi U yang dihasilkan pada iterasi terakhir (iterasi ke-37) menggunakan fungsi Matlab adalah :

ISSN:2442-5567 | E-ISSN:2443-289X

Columns 1 through 10

0.1634	0.5724	0.0523	0.0467	0.9198	0.0643	0.0525	0.1063	0.0438	0.0605
0.1810	0.1653	0.0954	0.3021	0.0345	0.7542	0.2756	0.1726	0.3300	0.1892
0.6557	0.2623	0.8522	0.6512	0.0457	0.1815	0.6719	0.7211	0.6262	0.7503

Columns 11 through 20

0.0507	0.0576	0.6670	0.0923	0.6497	0.1132	0.6893	0.0108	0.7094	0.0427
0.7378	0.2660	0.1091	0.0971	0.1801	0.6406	0.1642	0.9393	0.0919	0.3048
0.2115	0.6763	0.2239	0.8107	0.1701	0.2463	0.1465	0.0500	0.1987	0.6525

Columns 21 through 30

0.0463	0.0299	0.8531	0.3603	0.0327	0.0592	0.0435	0.0470	0.8602	0.0845
0.7249	0.0572	0.0618	0.3797	0.8855	0.1855	0.3987	0.3511	0.0572	0.7507
0.2288	0.9129	0.0852	0.2600	0.0818	0.7553	0.5578	0.6018	0.0826	0.1648

Columns 31 through 40

0.8104	0.0560	0.1747	0.0183	0.3397	0.6931	0.0611	0.0556	0.2106	0.6219
0.0792	0.8198	0.6184	0.0430	0.3074	0.1693	0.1484	0.1314	0.5727	0.1510
0.1105	0.1242	0.2069	0.9386	0.3530	0.1376	0.7904	0.8130	0.2167	0.2271

Columns 41 through 50

0.0513	0.1739	0.5558	0.8307	0.8829	0.1085	0.0982	0.0307	0.0732	0.0897
0.8324	0.6221	0.2029	0.0788	0.0535	0.2561	0.1566	0.0896	0.4973	0.0934
0.1163	0.2040	0.2413	0.0905	0.0636	0.6354	0.7452	0.8797	0.4295	0.8170

Columns 51 through 60

0.0050	0.1026	0.0297	0.0466	0.0378	0.0673	0.0446	0.0707	0.5855	0.5157
0.9734	0.1039	0.0565	0.2871	0.1144	0.1508	0.3947	0.5757	0.1587	0.2346
0.0216	0.7934	0.9139	0.6663	0.8478	0.7819	0.5607	0.3536	0.2558	0.2497

Columns 61 through 70

0.0825	0.0826	0.0433	0.1208	0.6158	0.0877	0.1063	0.7764	0.0346	0.8865
0.2752	0.2753	0.3423	0.1954	0.1421	0.1292	0.1073	0.0887	0.0653	0.0543
0.6423	0.6421	0.6144	0.6838	0.2421	0.7831	0.7864	0.1349	0.9000	0.0592

Columns 71 through 75

0.8488	0.0548	0.7033	0.6666	0.7703
0.0635	0.4938	0.0943	0.1236	0.1006
0.0877	0.4514	0.2024	0.2097	0.1291

Pada iterasi terakhir (iterasi ke-37), pusat *cluster Vkj* yang dihasilkan dengan k=1,2,3,4 dan j=1,2 adalah:

$$V_{kj} = \begin{pmatrix} 3.2649 & 1.8625 & 2.2484 & 4.2701 \\ 3.1653 & 1.6271 & 5.1550 & 1.1770 \\ 3.2225 & 2.4133 & 3.0759 & 0.7967 \end{pmatrix}$$

ISSN:2442-5567 | E-ISSN:2443-289X

Nilai ini merupakan nilai dari koordinat ketiga titik pusat *cluster* dan memberikan garis besar tiap *cluster* yaitu:

1. Kelompok pertama (*cluster* ke-1), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.2649; memiliki rata-rata penghasilan orang tua sekitar 1.8625; tanggungan orang tua 2.2484 dan memiliki prestasi 4.2701.
2. Kelompok kedua (*cluster* ke-2), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.1653; memiliki rata-rata penghasilan orang tua sekitar 1.6271; tanggungan orang tua 5.1550 dan memiliki prestasi 1.1770.
3. Kelompok ketiga (*cluster* ke-3), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.2225; memiliki rata-rata penghasilan orang tua sekitar 2.4133; tanggungan orang tua 3.0759 dan memiliki prestasi 0.7967.

Derajat keanggotaan penentuan beasiswa untuk sertiap *cluster* pada iterasi terakhir (iterasi ke-37) dapat dilihat pada Tabel 1. di bawah ini:

Tabel 1. Derajat Keanggotaan Penentuan Beasiswa Untuk Setiap *Cluster* Pada Iterasi Terakhir

No	Derajat Keanggotaan (μ) pada Iterasi Terakhir			Kecenderungan data masuk pada *cluster*		
	(μ_1)	(μ_2)	(μ_3)	C1	C2	C3
1	0.1634	0.1810	0.6557			*
2	0.5724	0.1653	0.2623	*		
3	0.0523	0.0954	0.8522			*
4	0.0467	0.3021	0.6512			*
5	0.9198	0.0345	0.0457	*		
6	0.0643	0.7542	0.1815		*	
7	0.0525	0.2756	0.6719			*
8	0.1063	0.1726	0.7211			*
9	0.0438	0.3300	0.6262			*
10	0.0605	0.1892	0.7503			*
11	0.0507	0.7378	0.2115		*	
12	0.0576	0.2660	0.6763			*
13	0.6670	0.1091	0.2239	*		
14	0.0923	0.0971	0.8107			*
15	0.6497	0.1801	0.1701	*		
16	0.1132	0.6406	0.2463		*	
17	0.6893	0.1642	0.1465	*		
18	0.0108	0.9393	0.0500		*	
19	0.7094	0.0919	0.1987	*		
20	0.0427	0.3048	0.6525			*
21	0.0463	0.7249	0.2288		*	
22	0.0299	0.0572	0.9129			*
23	0.8531	0.0618	0.0852	*		
24	0.3603	0.3797	0.2600		*	
25	0.0327	0.8855	0.0818		*	
26	0.0592	0.1855	0.7553			*

ISSN:2442-5567 | E-ISSN:2443-289X

Nilai ini merupakan nilai dari koordinat ketiga titik pusat *cluster* dan memberikan garis besar tiap *cluster* yaitu:

1. Kelompok pertama (*cluster* ke-1), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.2649; memiliki rata-rata penghasilan orang tua sekitar 1.8625; tanggungan orang tua 2.2484 dan memiliki prestasi 4.2701.
2. Kelompok kedua (*cluster* ke-2), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.1653; memiliki rata-rata penghasilan orang tua sekitar 1.6271; tanggungan orang tua 5.1550 dan memiliki prestasi 1.1770.
3. Kelompok ketiga (*cluster* ke-3), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.2225; memiliki rata-rata penghasilan orang tua sekitar 2.4133; tanggungan orang tua 3.0759 dan memiliki prestasi 0.7967.

Derajat keanggotaan penentuan beasiswa untuk sertiap *cluster* pada iterasi terakhir (iterasi ke-37) dapat dilihat pada Tabel 2. di bawah ini:

Tabel 2. Derajat Keanggotaan Penentuan Beasiswa Untuk Setiap *Cluster* Pada Iterasi Terakhir

No	Derajat Keanggotaan (μ) pada Iterasi Terakhir			Kecenderungan data masuk pada *cluster*		
	(μ_1)	(μ_2)	(μ_3)	C1	C2	C3
1	0.1634	0.1810	0.6557			*
2	0.5724	0.1653	0.2623	*		
3	0.0523	0.0954	0.8522			*
4	0.0467	0.3021	0.6512			*
5	0.9198	0.0345	0.0457	*		
6	0.0643	0.7542	0.1815		*	
7	0.0525	0.2756	0.6719			*
8	0.1063	0.1726	0.7211			*
9	0.0438	0.3300	0.6262			*
10	0.0605	0.1892	0.7503			*
11	0.0507	0.7378	0.2115		*	
12	0.0576	0.2660	0.6763			*
13	0.6670	0.1091	0.2239	*		
14	0.0923	0.0971	0.8107			*
15	0.6497	0.1801	0.1701	*		
16	0.1132	0.6406	0.2463		*	
17	0.6893	0.1642	0.1465	*		
18	0.0108	0.9393	0.0500		*	
19	0.7094	0.0919	0.1987	*		
20	0.0427	0.3048	0.6525			*
21	0.0463	0.7249	0.2288		*	
22	0.0299	0.0572	0.9129			*
23	0.8531	0.0618	0.0852	*		
24	0.3603	0.3797	0.2600		*	
25	0.0327	0.8855	0.0818		*	
26	0.0592	0.1855	0.7553			*

Nilai ini merupakan nilai dari koordinat ketiga titik pusat *cluster* dan memberikan garis besar tiap *cluster* yaitu:

1. Kelompok pertama (*cluster* ke-1), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.2649; memiliki rata-rata penghasilan orang tua sekitar 1.8625; tanggungan orang tua 2.2484 dan memiliki prestasi 4.2701.
2. Kelompok kedua (*cluster* ke-2), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.1653; memiliki rata-rata penghasilan orang tua sekitar 1.6271; tanggungan orang tua 5.1550 dan memiliki prestasi 1.1770.
3. Kelompok ketiga (*cluster* ke-3), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.2225; memiliki rata-rata penghasilan orang tua sekitar 2.4133; tanggungan orang tua 3.0759 dan memiliki prestasi 0.7967.

Derajat keanggotaan penentuan beasiswa untuk sertiap *cluster* pada iterasi terakhir (iterasi ke-37) dapat dilihat pada Tabel 3. di bawah ini:

Tabel 3. Derajat Keanggotaan Penentuan Beasiswa Untuk Setiap *Cluster* Pada Iterasi Terakhir

No	Derajat Keanggotaan (μ) pada Iterasi Terakhir			Kecenderungan data masuk pada *cluster*		
	(μ_1)	(μ_2)	(μ_3)	C1	C2	C3
1	0.1634	0.1810	0.6557			*
2	0.5724	0.1653	0.2623	*		
3	0.0523	0.0954	0.8522			*
4	0.0467	0.3021	0.6512			*
5	0.9198	0.0345	0.0457	*		
6	0.0643	0.7542	0.1815		*	
7	0.0525	0.2756	0.6719			*
8	0.1063	0.1726	0.7211			*
9	0.0438	0.3300	0.6262			*
10	0.0605	0.1892	0.7503			*
11	0.0507	0.7378	0.2115		*	
12	0.0576	0.2660	0.6763			*
13	0.6670	0.1091	0.2239	*		
14	0.0923	0.0971	0.8107			*
15	0.6497	0.1801	0.1701	*		
16	0.1132	0.6406	0.2463		*	
17	0.6893	0.1642	0.1465	*		
18	0.0108	0.9393	0.0500		*	
19	0.7094	0.0919	0.1987	*		
20	0.0427	0.3048	0.6525			*
21	0.0463	0.7249	0.2288		*	
22	0.0299	0.0572	0.9129			*
23	0.8531	0.0618	0.0852	*		
24	0.3603	0.3797	0.2600		*	
25	0.0327	0.8855	0.0818		*	
26	0.0592	0.1855	0.7553			*

ISSN:2442-5567 | E-ISSN:2443-289X

1. Kelompok pertama (*cluster* ke-1), akan berisi data mahasiswa ke : 2, 5, 13, 15, 17, 19, 23, 29, 31, 36, 40, 43, 44, 45, 59, 60, 65, 68, 70, 71, 73, 74 dan 75.
2. Kelompok kedua (*cluster* ke-2), akan berisi data mahasiswa ke : 6, 11, 16, 18, 21, 24, 25, 30, 32, 33, 39, 41, 42, 49, 51, 58 dan 72.
3. Kelompok ketiga (*cluster* ke-3), akan berisi data mahasiswa ke : 1, 3, 4, 7, 8, 9, 10, 12, 14, 20, 22, 26, 27, 28, 34, 35, 37, 38, 46, 47, 48, 50, 52, 53, 54, 55, 56, 57, 61, 62, 63, 64, 66, 67 dan 69.

3.2 PEMBAHAAN

Proses *Clustering* memerlukan klasifikasi untuk menentukan kelompok (*cluster*) mana yang berhak untuk menerima beasiswa. Dalam penelitian akan mengelompokkan mahasiswa mejadi tiga (3) *cluster* yaitu :

1. *Cluster* yang menerima beasiswa
2. *Cluster* yang dipertimbangkan menerima beasiswa.
3. *Cluster* yang tidak menerima beasiswa.

Kemudian setiap *cluster* dibagi berdasarkan kriteria mana yang lebih diprioritaskan (berdasarkan IPK atau PO (Penghasilan Orang tua), Tanggungan Orang Tua (TO), Prestasi mahasiswa selama menjadi mahasiswa).

Iterasi pada percobaan ini berhenti pada iterasi ke-37. Hasil akhir *clutering* yang di perolah adalah :

1. *Cluster* pertama memiliki pusat *cluster* (3.2649; 1.8625; 2.2484; 4.2701)
2. *Cluster* kedua memiliki pusat *cluster* (3.1653; 1.6271; 5.1550; 1.1770)
3. *Cluster* ketiga memiliki pusat *cluster* (3.2225; 2.4133; 3.0759; 0.7967)

Hasil klasifikasi *clustering* dapat dilihat pada Tabel 4.

Tabel 4. Hasil Kalasifikasi

Proritas IPK		
Menerima	Dipertimbangkan	Tidak Menerima
Cluster 1	*Cluster* 3	*Cluster* 2
3.2649	3.2225	3.1653
Prioritas TK		
Menerima	Dipertimbangkan	Tidak Menerima
Cluster 3	*Cluster* 1	*Cluster* 2
2.4133	1.8625	1.6271
Prioritas TO		
Menerima	Dipertimbangkan	Tidak Menerima
Cluster 2	*Cluster* 3	*Cluster* 1
5.1550	3.0759	2.2484
Prioritas Prestasi		
Menerima	Dipertimbangkan	Tidak Menerima
Cluster 1	*Cluster* 2	*Cluster* 3
4.2701	1.1770	0.7967

ISSN:2442-5567 | E-ISSN:2443-289X

Hasil klasifikasi mahasiswa berdasarkan Tabel 4 untuk 75 data dapat dilihat pada Tabel 5 di bawah ini.

Tabel 5. Hasil klasifikasi prioritas mahasiswa

No	IPK	PO	TO	Prestasi	*C*	IPK	PO	TO	Prestasi
1	2.81	1.2	2	1	3	DP	DP	M	TM
2	2.85	2	3	3	1	M	TM	DP	M
3	2.84	3.2	3	1	3	DP	DP	M	TM
4	2.87	2.5	4	1	3	DP	DP	M	TM
5	2.87	2.3	2	5	1	M	TM	DP	M
6	2.88	0.8	6	0	2	TM	M	TM	DP
7	2.89	2.8	4	1	3	DP	DP	M	TM
8	2.89	1.5	2	0	3	DP	DP	M	TM
9	2.95	2.3	4	1	3	DP	DP	M	TM
10	2.98	1.5	3	0	3	DP	DP	M	TM
11	3	1	5	0	2	TM	M	TM	DP
12	3	3	4	1	3	DP	DP	M	TM
13	3	1.5	1	3	1	M	TM	DP	M
14	3.02	2.3	2	1	3	DP	DP	M	TM
15	3.03	3	4	5	3	DP	DP	M	TM
16	3.03	3.3	7	1	2	TM	M	TM	DP
17	3.03	2.4	4	5	1	M	TM	DP	M
18	3.04	2	5	1	2	TM	M	TM	DP
19	3.05	2	2	3	1	M	TM	DP	M
20	3.06	2.4	4	1	3	DP	DP	M	TM
21	3.08	2.7	5	1	2	TM	M	TM	DP
22	3.08	3	3	1	3	DP	DP	M	TM
23	3.1	3	2	5	1	M	TM	DP	M
24	3.1	1.3	4	3	2	TM	M	TM	DP
25	3.11	1.2	6	1	2	TM	M	TM	DP
26	3.12	1.5	3	0	3	DP	DP	M	TM
27	3.12	2	4	1	3	DP	DP	M	TM
28	3.13	2	4	0	1	M	TM	DP	M
29	3.13	2	1	5	2	TM	M	TM	DP
30	3.15	1	7	1	2	TM	M	TM	DP
31	3.15	3.3	2	5	1	M	TM	DP	M
32	3.15	0.6	6	1	2	TM	M	TM	DP
33	3.16	2.3	6	3	2	TM	M	TM	DP
34	3.17	2	3	1	3	DP	DP	M	TM
35	3.22	3.3	4	3	3	DP	DP	M	TM
36	3.19	1.3	4	5	1	M	TM	DP	M
37	3.21	1.5	3	1	3	DP	DP	M	TM

ISSN:2442-5567 | E-ISSN:2443-289X

38	3.21	3.3	3	0	3	DP	DP	M	TM
39	3.22	1.8	5	3	2	TM	M	TM	DP
40	3.23	0.5	2	3	1	M	TM	DP	M
41	3.24	0.7	6	1	2	TM	M	TM	DP
42	3.25	1.8	7	3	2	TM	M	TM	DP
43	3.25	0.9	3	3	1	M	TM	DP	M
44	3.27	0.5	2	5	1	M	TM	DP	M
45	3.29	0.9	2	5	1	M	TM	DP	M
46	3.3	1	3	1	3	DP	DP	M	TM
47	3.32	3.8	3	1	3	DP	DP	M	TM
48	3.32	2.3	3	0	3	DP	DP	M	TM
49	3.32	0.9	4	0	2	TM	M	TM	DP
50	3.34	2.4	2	1	3	DP	DP	M	TM
51	3.35	1.7	5	1	2	TM	M	TM	DP
52	3.35	3	2	1	3	DP	DP	M	TM
53	3.35	3	3	1	3	DP	DP	M	TM
54	3.37	2.5	4	0	3	DP	DP	M	TM
55	3.39	2	3	0	3	DP	DP	M	TM
56	3.39	3.5	3	0	3	DP	DP	M	TM
57	3.4	2	4	1	3	DP	DP	M	TM
58	3.32	1	4	1	2	TM	M	TM	DP
59	3.23	2	3	3	2	TM	M	TM	DP
60	3.33	0.4	3	3	1	M	TM	DP	M
61	3.42	3.5	4	1	3	DP	DP	M	TM
62	3.43	3.5	4	1	3	DP	DP	M	TM
63	3.42	2.2	4	1	3	DP	DP	M	TM
64	3.45	1.2	2	0	3	DP	DP	M	TM
65	3.45	0.5	1	3	1	M	TM	DP	M
66	3.47	3	2	0	3	DP	DP	M	TM
67	3.53	3	2	1	3	DP	DP	M	TM
68	3.51	3.2	1	5	1	M	TM	DP	M
69	3.53	3	3	1	3	DP	DP	M	TM
70	3.55	2	3	5	1	M	TM	DP	M
71	3.59	3	2	5	1	M	TM	DP	M
72	3.59	1.6	4	1	2	TM	M	TM	DP
73	3.63	2	2	3	1	M	TM	DP	M
74	3.67	1	2	3	1	M	TM	DP	M
75	3.67	0.4	1	5	1	M	TM	DP	M

Keterangan :
M : menerima.
DP : dipertimbangkan.

TM : tidak menerima.

Berdasarkan tabel prioritas di atas terdapat 3 *cluster* yaitu menerima, dipertimbangkan dan tidak menerima yang di hasilkan dari tabel kalsifikasi.

4. KESIMPULAN DAN SARAN

4.1 KESIMPULAN

Berdasarkan pembahasan maka dapat diberikan beberapa simpulan sebagai berikut:

1. Algoritma *Fuzzy C-means* dapat digunakan untuk mengelompokan data mahasiswa lebih halus dengan menerap- kan besarnya derajat keanggotaan setiap elemen untuk masuk ke dalam kelom- pok-kelompok yang ada.
2. Pengujian data dilakukan sebanyak 37 iterasi, diperoleh tiga kelompok berdasarkan nilai rata-rata penentuan beasiswa, yaitu :
   a. Kelompok pertama (*cluster* ke-1), berisi mahasiswa yang memiliki rata-rata ipk sekitar 3.2649; memiliki rata-rata pendapatan orang tua sekitar 1.8625; tanggungan orang tua 2.2484 dan memiliki prestasi 4.2701.
   b. Kelompok kedua (*cluster* ke-2), berisi mahasiswa yang memiliki rata-rata ipk sekitar 3.1653; memiliki rata-rata pendapatan orang tua sekitar 1.6271; tanggungan orang tua 5.1550 dan memiliki prestasi 1.1770.
   c. Kelompok ketiga (*cluster* ke-3), berisi mahasiswa yang memiliki rata-rata ipk sekitar 3.2225; memiliki rata-rata pendapatan orang tua sekitar 2.4133; tanggungan orang tua 3.0759 dan memiliki prestasi 0.7967.
3. Kemudian setiap *cluster* diklasifikasikan berdasarkan kriteria mana yang lebih diprioritaskan dengan nilai terbesar pada jarak akhir merupakan *cluster* yang menerima beasiswa, sedangkan *cluster* dengan nilai terkecil merupakan *cluster* yang tidak berhak menerima beasiswa.

4.2 SARAN

Saran yang dapat digunakan untuk mendapat tingkat akurasi yang lebih baik dalam penerapan algoritma *Fuzzy C-Means* untuk penentuan beasiswa di Univeristas Megou Pak Tulang Bawang, serta untuk memberikan konstribusi yang lebih besar di dunia riset, disarankan agar hasil penelitian ini dikembangkan dengan cara memodifikasi/*updating* algoritma *Fuzzy C-means* yang digunakan saat ini, atau dengan menggabungkan algoritma *Fuzzy C-means* dengan algoritma lain.

REFERENSI

[1] Ahmad Mauliyadi M, dkk. (2013) , " Perbandingan Metode *Fuzzy C-Means* (Fcm) Dan *Fuzzy Gustafson-Kessel* (Fgk) Menggunakan Data Citra Satelit *Quickbird", Matematika 00 (2013) 01–05,*

[2] Ali Keshavarzi, Fereydoon Sarmadian (2012)," Fuzzy Clustering Analysis For Modeling Of Soil Cation Exchange Capacity", AJAE 3(1):27-33 (2012), ISSN:1836-9448.

[3] Apriansyah Putra, Dinna Yunika Hardiyanti (2011) "Penentuan Penerima Beasiswa Dengan Menggunakan *Fuzzy Multiple Atribute Decission Makin ",Vol. 3, No. 1, April 2011, Halaman 286-293.*

ISSN:2442-5567 | E-ISSN:2443-289X

[4] Bahar, Vincent Suhartono, dkk (2011), "Penentuan Jurusan Sekolah Menengah Atas Dengan Algoritma Fuzzy C-Means", Jurnal Teknologi Informasi, Issn 1414-9999.

[5] Dimas Wahyu Wibowo, dkk (2013) " Perhitungan Jumlah dan Jenis Kendaraan Menggunakan Metode Fuzzy C-means dan Segmentasi Deteksi Tepi Canny" , *Vol. 7, No. 2, Desember 2013, Halaman 103-110.*

[6] Eko Sediyono, dkk (2006), " Penentuan Lokasi Fasilitas Gudang Menggunakan *Fuzzy C – Means* (FCM)", Vol.2, No. 2, Desember 2006: 155 – 166.

[7] Ivan Hardiyanto, dkk (2012), " Implementasi Segmentasi Citra Dengan Menggunakan Metode Generalized Fuzzy Cmeans Clustering Algorithm With Improved Fuzzy Partitions", *Vol. 1, No. 1, (2012) 1-5*

[8] Heru Dibyo Laksono, Abraham Arief, (2012)," Perkiraan Kebutuhan Energi Listrik Jangka Panjang Di Provinsi Sumatera Barat Sampai Tahun 2021 Dengan Logika Fuzzy Clustering ",Vol. 5 No. 2 September 2012, Issn : 2086 – 4981.

[9] Kusumadewi, S., Purnomo, H., (2010), *Aplikasi Fuzzy Untuk Pendukung Keputusan*, Graha Ilmu, Jakarta

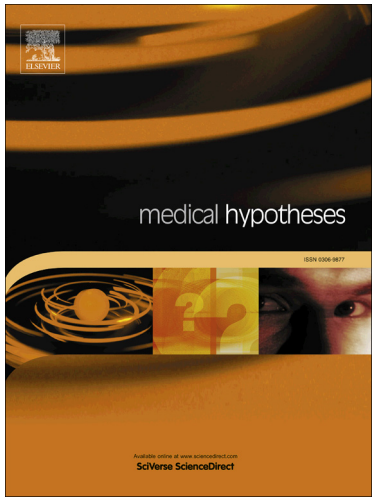
[10] Kusrini, (2006), Algoritma Data Mining, Penerbit ANDI, Yogyakarta.

[11] Noor Fitriana Hastuti, (2010), Pemanfaatan Metode *K-Means Clustering* dalam Penentuan Penerima Beasiswa, Jurusan Informatika. Fakultas MIPA. Universitas Sebelas Maret.

[12] Prabowo Pudjo Widodo, dkk (2013), " Penerapan *Data Mining* dengan *Matlab",* Penerbit Rekayasa Sains, Bandung.

[13] Turban, Efraim, et al . (2005). Decision Support Systems and Intelligent Systems. Yogyakarta : Penerbit Andi

Journal Pre-proofs



Brain Tumor Segmentation Approach Based on the Extreme Learning Machine and Significantly Fast and Robust Fuzzy C-Means Clustering Algorithms Running on Raspberry Pi Hardware

Fatih Ş İ Ş İK, Eser Sert

PII: S0306-9877(19)31095-3
DOI: https://doi.org/10.1016/j.mehy.2019.109507
Reference: YMEHY 109507

To appear in: *Medical Hypotheses*

Received Date: 30 September 2019
Revised Date: 11 November 2019
Accepted Date: 16 November 2019

Please cite this article as: F. Ş İ Ş İK, E. Sert, Brain Tumor Segmentation Approach Based on the Extreme Learning Machine and Significantly Fast and Robust Fuzzy C-Means Clustering Algorithms Running on Raspberry Pi Hardware, *Medical Hypotheses* (2019), doi: https://doi.org/10.1016/j.mehy.2019.109507

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2019 Published by Elsevier Ltd.

Journal Pre-proofs

Brain Tumor Segmentation Approach Based on the Extreme Learning Machine and Significantly Fast and Robust Fuzzy C-Means Clustering Algorithms Running on Raspberry Pi Hardware

Fatih ŞİŞİK[1], Eser SERT[2*]

[1] Göksun Vocational School, Department of Computer Programming, Kahramanmaras Sutcu Imam University, K.Maras, Turkey
[2] Department of Computer Engineering, Engineering and Architecture Faculty, Kahramanmaras Sutcu Imam University, K.Maras, Turkey
∗ Corresponding author e-mail address: esersert@ksu.edu.tr

ABSTRACT

Automatic decision support systems have gained importance in health sector in recent years. In parallel with recent developments in the fields of artificial intelligence and image processing, embedded systems are also used in decision support systems for tumor diagnosis. Extreme learning machine (ELM), is a recently developed, quick and efficient algorithm which can quickly and flawlessly diagnose tumors using machine learning techniques. Similarly, significantly fast and robust fuzzy C-means clustering algorithm (FRFCM) is a novel and fast algorithm which can display a high performance. In the present study, a brain tumor segmentation approach is proposed based on extreme learning machine and significantly fast and robust fuzzy C-means clustering algorithms (BTS-ELM-FRFCM) running on Raspberry Pi (PRI) hardware. The present study mainly aims to introduce a new segmentation system hardware containing new algorithms and offering a high level of accuracy the health sector. PRI's are useful mobile devices due to their cost-effectiveness and satisfying hardware. 3200 pieces of T1 gadolium (T1-GD) sequence MRI images were used to train ELM in the present study. 20 pieces of MRI images were used for testing process. Figure of merid (FOM), jaccard similarity coefficient (JSC) and Dice indexes were used in order to evaluate the performance of the proposed approach. In addition, the proposed method was compared with brain tumor segmentation based on support vector machine (BTS-SVM), brain tumor segmentation based on fuzzy C-means (BTS-FCM) and brain tumor segmentation based on self-organizing maps and k-means (BTS-SOM). The statistical analysis on FOM, JSC and Dice

Journal Pre-proofs

results obtained using four different approaches indicated that BTS-ELM-FRFCM displayed the highest performance. Thus, it can be concluded that the embedded system designed in the present study can perform brain tumor segmentation with a high accuracy rate.

Key words: Extreme learning machine, Brain tumor segmentation, Segmentation, Magnetic resonance imaging, Significantly fast and robust fuzzy C-means clustering, Raspberry pi

1. Introduction

Various new developments have recently occurred in the field of embedded systems, image processing and machine learning. Thanks to the combination of these systems, new support systems can be designed as a contribution to image recognition, segmentation [1], edge detection [2] and data extraction. One of the leading sectors where these systems are widely used is health as they provide doctors with solutions to various problems such as the diagnosis of various diseases and detection of small tumors.

Various studies have so far been conducted on brain tumor segmentation in the literature. Zotin et al. [3] performed edge detection in MRI brain tumor images using fuzzy C-means (FCM). Yilmaz et al. [4] performed skull stripping in brain MRI using multi-stable cellular neural networks. Chandra et al. [5] performed brain tumor detection on MRI images using Genetic Algorithm. Mittal et al. [6] proposed a deep learning-based approach which detected brain tumors on MRI images. Havaei et al. [7] performed brain tumor segmentation using deep neural networks. Additionally, it can be observed in the current literature that brain tumor segmentation was performed using superpixel based classification [8], adaptive neuro-fuzzy inference system (ANFIS) [9], support vector machines (SVM) [10, 11], extremely randomized trees [12] and neutrosophic expert maximum fuzzy-sure Entropy (NS-EMFSE) [13]. Several studies [14-16] have also been conducted in the literature on convolutional neural network (CNN) based brain tumor classification.

ELM has so far been used in various studies. An example is Zhu et al. [17] who benefited from ELM for retinal vessel segmentation. Similarly, Zeng et al. [18] used ELM for liver vessel segmentation. Huang et al. [19] conducted a study on neovascularization using ELM. In addition,

Journal Pre-proofs

ELM algorithm was used for different purposes such as breast mass classification [20], facial age range estimation [21] and action recognition [22].

ELM approach does not possess an iterative structure, thus being trained faster compared to conventional back propagation approaches [23]. Additionally, this approach requires a fairly short amount of time for calculation [24]. Therefore, it can be argued that ELM is a more successful and faster approach compared to numerous neural network based approaches and machine learning techniques. Approaches such as neural network based segmentation, edge detection and object recognition usually perform well. However, the performance of these systems is heavily influenced by the variety of training data. Although FCM approach is widely used for segmentation and edge detection applications in the field of medicine, its greatest drawback is its sensitivity to noise. Similarly, the greatest disadvantage of fuzzy based segmentation approaches is that it encounters difficulties in creating fuzzy rules.

Several approaches were proposed for Raspberry Pi (RPI) hardware in the literature. Anandhalli et al. [25] developed a system for real time vehicle detection and tracking using RPI hardware. Sajjad et al. [26] developed a real-time facial expression recognition system using RPI device as a contribution to law enforcement. In addition to these studies, RPI hardware was also used for a tour guide robot [27], autonomous unmanned ground vehicle [28], vein imaging system [29], image segmentation [30], and automatic behavior analysis system for honey bees [31].

This study proposes a brain tumor segmentation approach based on the extreme learning machine and significantly fast and robust fuzzy C-means clustering algorithms (BTS-ELM-FRFCM) running on RPI hardware. As shown in Figure 3, the proposed BTS-ELM-FRFCM approach first trains ELM. Later, as shown in Figure 4, the trained ELM is used for MRI segmentation. The approach proposed in the present study benefits from FRFCM approach by Lei et al. [32] in order to apply pre-segmentation to MRI images in the training and testing process.

Main contributions of the present study can be summarized as follows:

Journal Pre-proofs

Journal Pre-proofs

1. The proposed BTS-ELM-FRFCM approach can be used on RPI embedded system. It can be noted in the literature that the number of segmentation approaches developed for an embedded system is fairly low, which can be considered as one of the most important contributions of the present study.
2. RPI hardware offers a mobile, user-friendly and low-cost device which can be used on a Linux operating system. Therefore, the segmentation approach proposed in the present study was uploaded to a RPI device.
3. ELM algorithm is the fastest artificial neural network (ANN) model. In the present study, ELM and FRFCM algorithms were used on RPI hardware in order to propose a brain tumor segmentation with a higher accuracy rate. No studies have been so far conducted in the existing literature as far as the segmentation using algorithms such as ELM and FRFCM on an embedded system platform are concerned.
4. Our previous study on NS-EMFSE [13] focused on brain tumor segmentation in an unsupervised machine learning technique. In the present study, on the other hand, a brain tumor segmentation approach using ELM was proposed in a supervised structure.

The proposed BTS-ELM-FRFCM approach was experimentally compared with brain tumor segmentation based on support vector machine (BTS-SVM), brain tumor segmentation based on fuzzy C-means (BTS-FCM) and brain tumor segmentation based on self-organizing maps and k-means (BTS-SOM) approaches in terms of segmentation performance.

Support vector machine (SVM) [33] is a statistical learning method which is widely used in various applications such as edge detection, segmentation and classification of various objects. Similar to SVM, Fuzzy C-means (FCM) [34] an algorithm used for performance comparison in the present study, is an unsupervised technique widely used for edge detection and segmentation approaches. Introduced by Kohonen [35], self-organizing maps (SOM) is an artificial neural network learning method and widely used in segmentation applications.

Journal Pre-proofs

Journal Pre-proofs

The present study is divided into the following sections: The database and proposed method are described in Section 2. BTS-SVM, BTS-FCM and BTS-SOM approaches are described in Section 3, 4 and 5, respectively. The metrics used for the measurement of segmentation performances are described in Section 6. Experimental studies are discussed in Section 7, and the findings are presented in Section 8.

2. Theoretical background

2.1 Database

In the present study, The Cancer Genome Atlas Glioblastoma Multiforme (TCGA-GBM) [36,37] and REMBRANDT [37,38] data sets in Cancer Imaging Archive were used for ELM training and testing processes. These are open access data sets, and thus do not require ethical committee approval. Brain tumors can be found on MRI images in these datasets, and many researchers resort to these data sets around the world. T1-gadolinium (T1-Gd) sequence MRI images enable a highly realistic brain tumor segmentation. Therefore, the present study benefits from T1-Gd sequence MRI images in these datasets.

2.2. RPI hardware

PRI hardware contains small-size equipment which can also be found in a computer. It has its own operating system, Raspbian, and can also be used on different operating systems. In the present study, RPI 3 was used on a Linux operating system, and Octave was preferred as a software in order to operate the proposed approach. RPI 3 hardware with LCD screen used in the present study is shown in **Fig. 1**.

Journal Pre-proofs

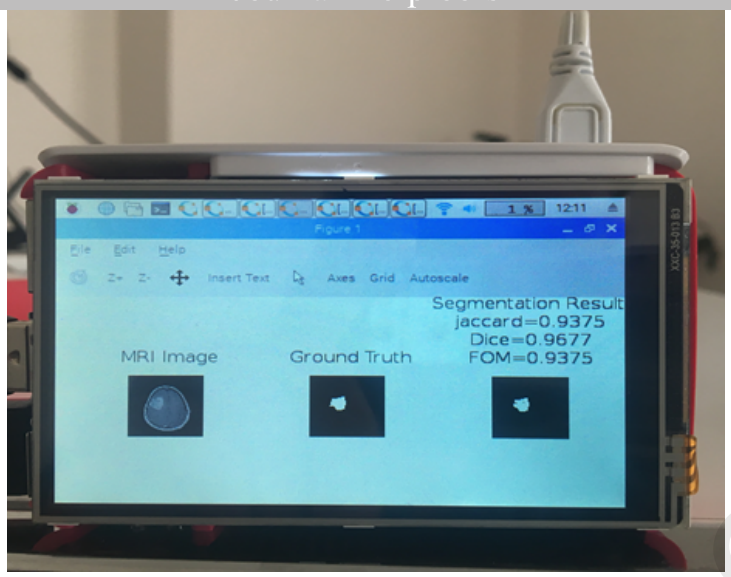


Fig. 1. RPI and LCD

2.3. ELM

ANN is a machine learning method based on human neural system. ELM is a feedforward neural network developed by Huang et al. The structure of single-hidden layer feed forward network (SLFN) [39] is shown in **Fig. 2**. ELM is widely used for regression [24], classification [40], segmentation [17], and object detection [41].

For S optional samples (x_i, t_i) for $\in R^n$ x R^m, the mathematical equation of ELM with SLFN structure is expressed as follows [42]:

$$\sum_{i=1}^{N} \beta_i g(w_i.x_j + b_i) = O_j. \quad j = 1,....S \tag{1}$$

here:

- $x_i = [x_{i1}, x_{i2}, x_{i3}, \ldots, x_{in}] \in R^n$, and $t_i = [t_{i1}, t_{i2}, t_{i3}, \ldots, t_{im}] \in R^m$,
- N is the number of hidden nodes,
- $w_i = [w_{i1}, w_{i2}, w_{i3}, \ldots, w_{in}]^T$ is ith weight vector connected to the hidden node,
- $\beta_i = [\beta_{i1}, \beta_{i2}, \beta_{i3}, \ldots, \beta_{im}]^T$ is the ith weight vector connection between the hidden node and output nodes,
- $b \in R$ is the bias of ith hidden node,

Journal Pre-proofs

- $O_j = [O_{j1}, O_{j2}, \ldots, O_{jm}]^T$ is the output of SLFN.

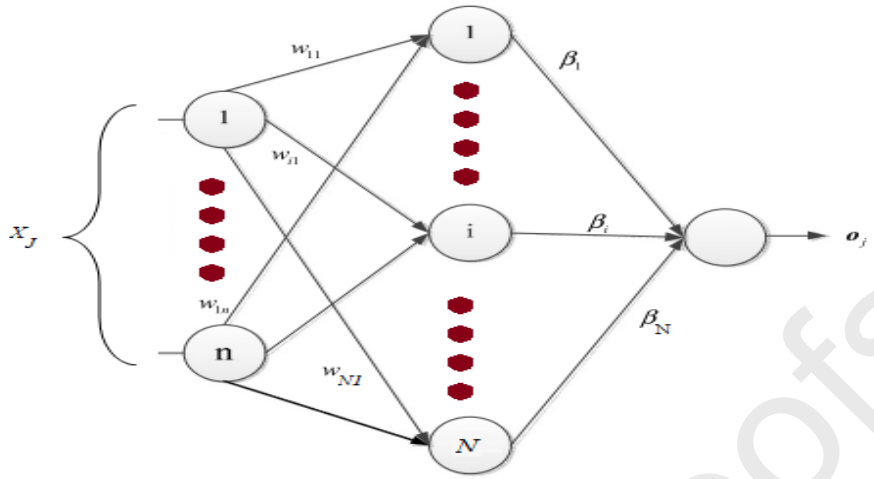


Fig. 2. The structure of SLFN.

Depending on the following formula, ELM converges to 0 mean error for N number of samples:

$$\sum_{i=1}^{N} \| O_j - t_j \| = 0 \tag{2}$$

where w_i, b_i and β_i is equalized to t_j as follows:

$$\sum_{i=1}^{L} \beta_i g_i(w_i.x_j + b_i) = t_j. \qquad j = 1, \ldots. N \tag{3}$$

The above-mentioned equations are expressed in matrix form as follows:

$$H\beta = T \tag{4}$$

H is the hidden layer matrix and T is the output matrix. The hidden layer matrix and output matrix are given in Equation 4 and 5.

$$H = \begin{bmatrix} G(w_1.x_1 + b_1) & \ldots & G(w_N.x_1 + b_N) \\ \vdots & \ddots & \vdots \\ G(w_1.x_S + b_1) & \ldots & G(w_N.x_S + b_N) \end{bmatrix}_{NxL} \tag{5}$$

$$T = [t_1 \quad \ldots \quad t_N]_{NXm} \qquad \text{and} \qquad \beta = [\beta_1 \quad \ldots \quad \beta_N]_{SXm} \tag{6}$$

Journal Pre-proofs

H is SLFN hidden layer output matrix. After hidden node parameters (a_i, b_i) are randomly selected, ELM is trained using least squares depending on the following formula:

$$\beta\hat{} = H^{+}T \tag{7}$$

Here H^{+} is the generalized inverse Moore-Penrose of H. More information on ELM can be found in [39].

2.4. FRFCM approach

Conventional FCM algorithm is sensitive to noise and thus encounters some problems in the segmentation process. In addition, FCM is sometimes a time-consuming approach for segmentation. Therefore, in order to minimize these problems, Lei et al. [32] proposed FRFCM algorithm which is a highly better and useful approach compared to FCM as it increases segmentation performance using morphological reconstruction and membership filtering. This algorithm displays an effective segmentation performance even in noisy images, which eliminates the need for selecting different filters for different types of noise [32]. Because FRFCM requires a low level of computation compared to the conventional FCM approach, it can be considered as a more fast approach. Further information on FRFCM can be found in [32].

2.5. The proposed approach

BTS-ELM-FRFCM approach proposed in the present study operates in two steps: training and testing process. As shown in Figure 3, 200 MRI images are used at the beginning and, thanks to the algorithm process of the proposed method, 3200 training images are obtained. These images and train label data are used to train ELM. As summarized in Figure 3, 20 MRI images in DICOM format are used to obtain test data thanks to the algorithm process of the proposed method. These obtained data and test labels are used to perform ELM based segmentation. The whole process takes place in RPI hardware.

Journal Pre-proofs

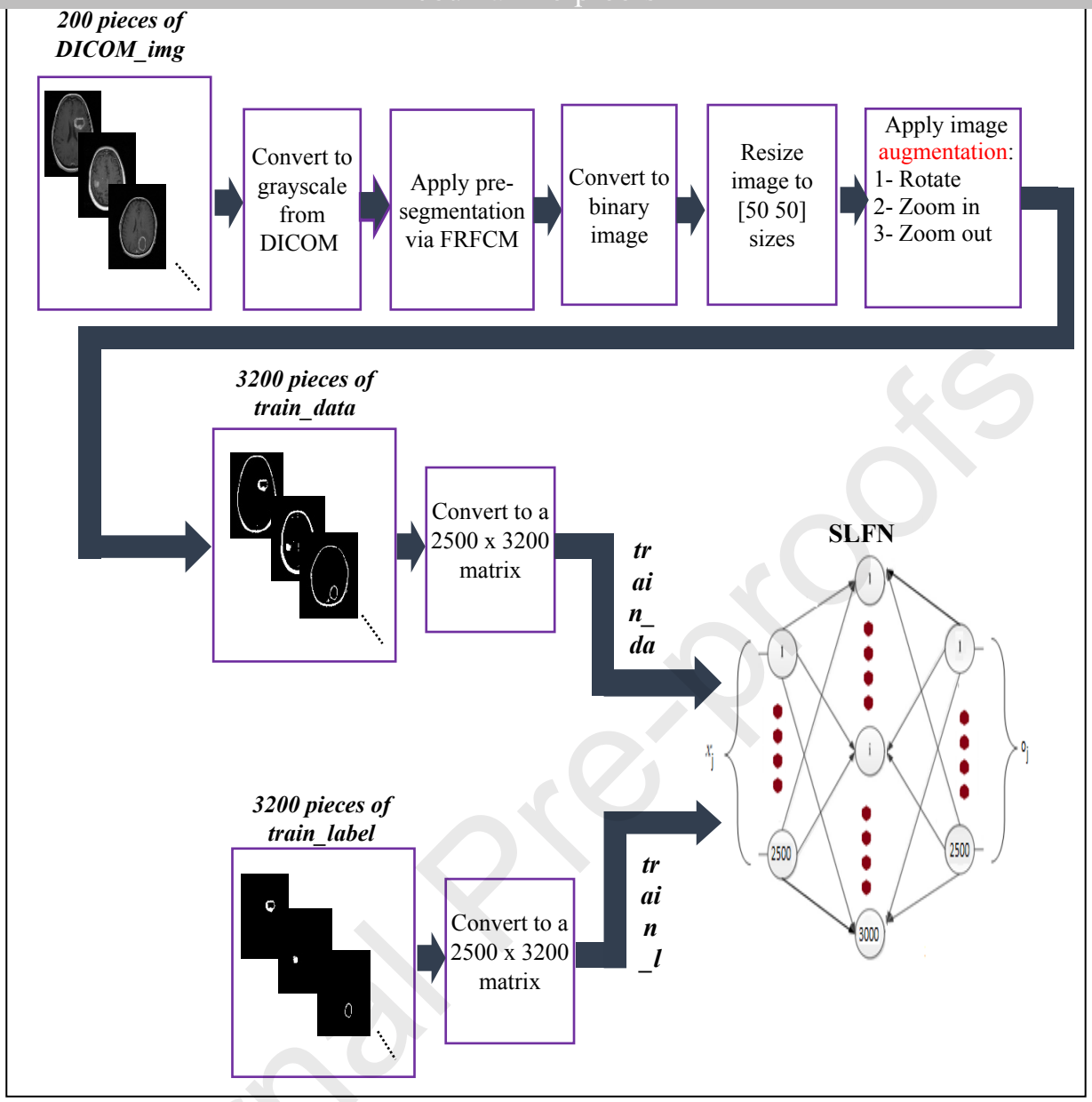


Fig. 3. The summary of BTS-ELM-FRFCM training process

Journal Pre-proofs

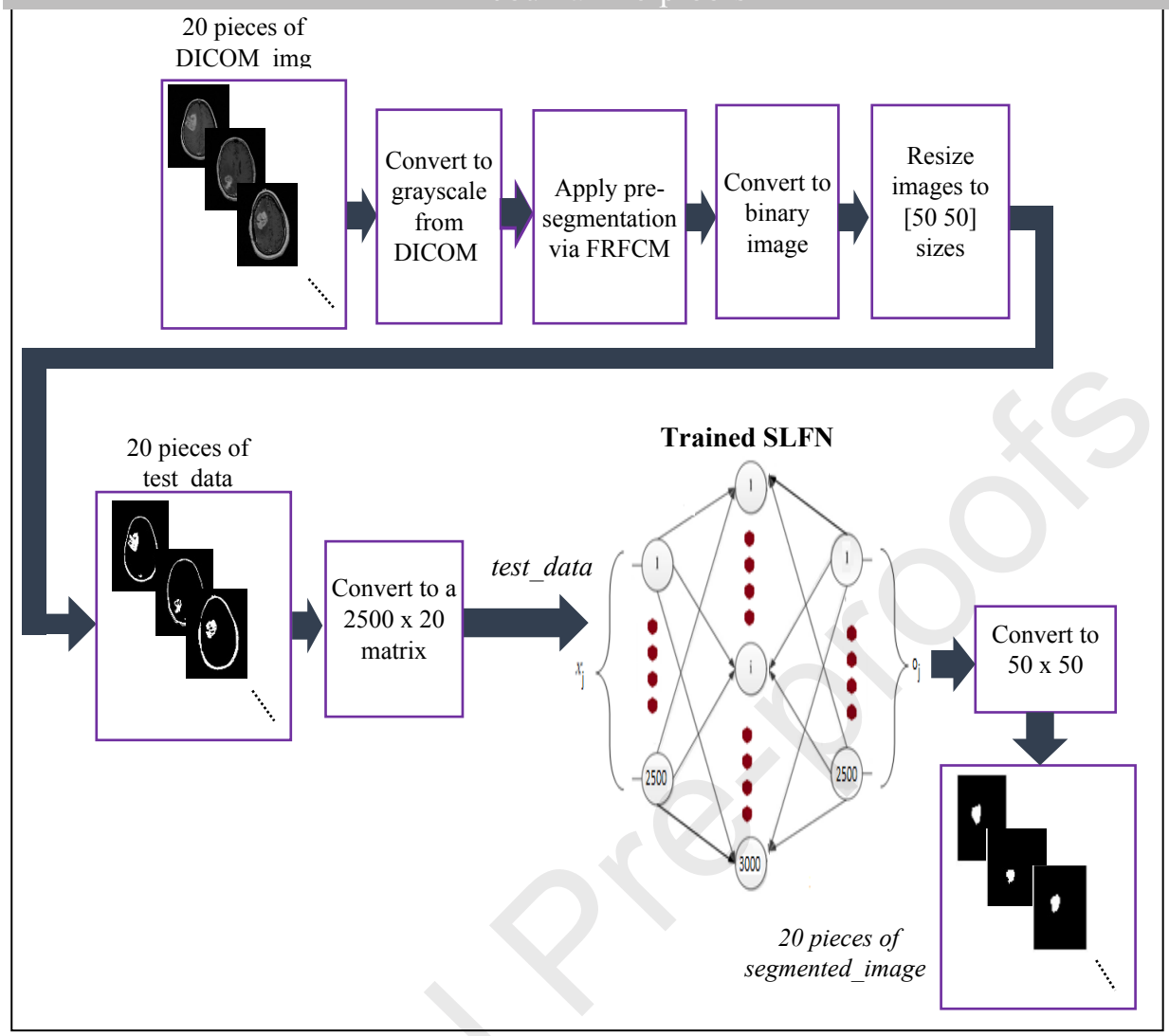


Fig. 4. BTS-ELM-FRFCM testing process

2.5.1. Training process

The details of BTS-ELM-FRFCM approach training process are given in **Algorithm 1**. The number of hidden layers in ELM training and testing process were adjusted as 3000. The experimental studies indicated that this value was sufficient to optimally operate RPI and ELM. The activation function of ELM was selected as sigmoid, and the numbers of input and outputs were adjusted as 2500 and 2500, respectively.

200 MRI images were used at the beginning of the training process. These images are described in detail in the Section "2.1. Database". MRI image in DICOM format is uploaded in **Algorithm 1** Step 1. Because the pixel values of images in DICOM format may vary in a broad range, they were converted to grayscale to make this range vary between 0 and 255. Thus, image processing became

Journal Pre-proofs

much quicker, and no apparent problems were observed in terms of BTS-ELM-FRFCM training quality. In Algorithm 1 Step 6, a pre-segmentation process is applied to images using FRFCM approach developed by Lei et al. [32]. For a quicker ELM training process, images was resized at a dimension of 50 x 50 in Step 5. In **Algorithm 1** Step 6, the image augmentation process was performed after training images had been rotated, zoomed in and zoomed out. In the first step, MRI images are rotated at 30, 60, 90, 120 and 150 degrees. In the second step, the MRI image is zoomed in at 5%, 10%, 15%, 20% and 25%. In the third step, the MRI image is zoomed out at 5%, 10%, 15%, 20% and 25%. In Fig.5, 15 images obtained by applying image augmentation to a sample MRI image are given. At the end of the image augmentation process, 3200 training images are obtained, train_data variable at a dimension of 3200 x 50 x 50 train_data is created.

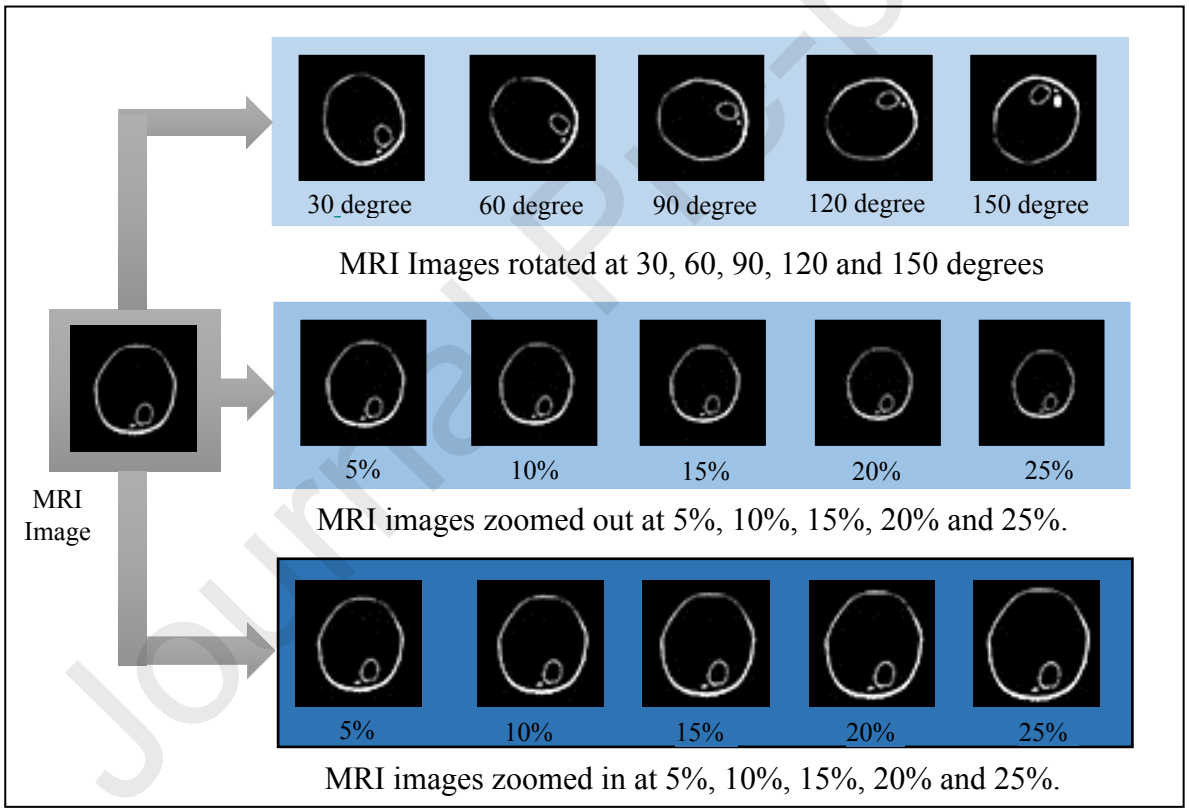


Fig. 5. The image augmentation process

A matrix at a dimension of 50 x 50 x 3200 in train_data variable was obtained, which was later converted to a matrix at a dimension of 2500 x 3200 in **Algorithm 1** Step 7 in order for ELM to process this matrix. ELM training is performed in Step 9. **Algorithm 1** is summarized in **Fig. 3**.

Algorithm 1. The preparation of training data and BTS-ELM-FRFCM training

Step 1: Use 200 MRI images in DICOM format

Step 2: Convert images to Grayscale format

Step 3: Apply pre-segmentation to images using **FRFCM** algorithm

Step 4: Convert images to binary format

Step 5: resized image [50 50], create 200 x 50 x 50 training images

Step 6: Increase training images to 3200 by applyling rotate, zoom in and zoom out processes and store them in **train_data**, create 3200 x 50 x 50 train images

Step 7: Resize **train_data** at 2500 x 3200

Step 8: Prepare **test_label** for 3200 test data and resize them at 2500 x 3200

Step 9: Train ELM using (**train_data, train_label**)

2.5.2. Testing process

Testing process of the proposed BTS-ELM-FRFCM approach is performed as shown in **Algorithm 2.** MRI images in DICOM format is uploaded in **Algorithm 1** Step 1. In Step 2, they were converted to grayscale. In Step 3, a pre-segmentation process is applied to images using FRFCM [32]. The images were converted to a matrix at a dimension of 2500 x 20 in Step 6, respectively in order for ELM to process MRI images. In Step 7, test label images containing ground truth information of 20 MRI images to be used in testing process were prepared. In Step 8, the trained ELM uses test_image and test_label to perform segmentation process. In the last step, segmented_image is resized at 20 x 50 x 50 in order to display it as an image. The testing process of BTS-ELM-FRFCM is summarized in **Fig. 4**. Images obtained by applying **Algorithm 2** to a sample MRI image are shown in **Table 1**.

Algorithm 2. The testing process

Step 1: Use 20 MRI images in DICOM format

Step 2: Convert images to Grayscale format

Step 3: Apply pre-segmentation to images using FRFCM algorithm

Journal Pre-proofs

Step 4: Convert images to binary format

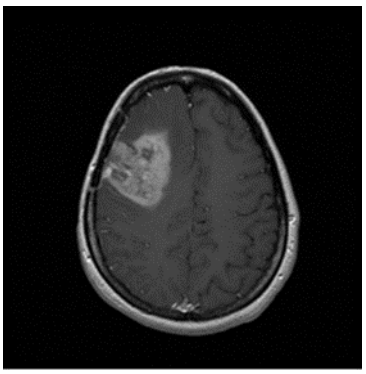
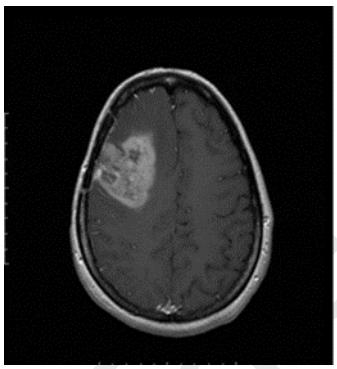
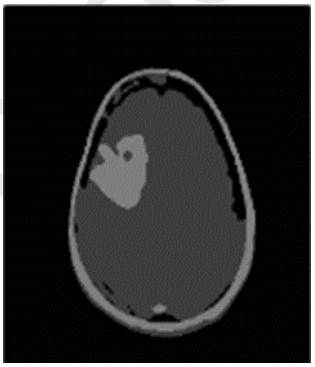
Step 5: resized images [50 50], create 20 x 50 x 50 testing images, and store them in **test_data** variable.

Step 6: Resize **test_data** at 2500 x 20

Step 7: Prepare **test_label** for **test_data** and resize them at 2500 x 20

Step 8: segmented_image ← Perform brain tumor segmentation using trained ELM (**test_image, test_label**)

Table 1. Test Results

DICOM Image (**Step 1**)	Grayscale Image (**Step 2**)	Pre-segmented image via FRFCM Image (**Step 3**)
		
Binary image (**Step 4**)	[50 50] resized image (**Step 5**)	segmented_image (**Step 8**)

3. BTS-SVM approach

Journal Pre-proofs

Introduced by Vladimir N. Vapnik in 1995, SVM is a successful and widely used statistical learning method in many fields [33]. It is divided into two categories as linear and non-linear SVM. Further information on SVM can be found in [33].

In the present study, BTS-SVM was used as an algorithm for comparison among different approaches. One Class SVM was preferred as a type of SVM, and gradient descent was used to minimize energy. The number of maximum iteration was set at 1000.

4. BTS-FCM approach

Fuzzy C-means algorithm is a popular data clustering algorithm developed by [43]. FCM is an uncontrolled learning approach which can successfully perform segmentation on medical images. An objective function is used in this algorithm, and it aims to group iteratively similar preliminary data which are randomly assigned. Group process is completed when it reaches a certain error value. In the image segmentation process, FCM algorithm groups similar pixels in an image, and behaves in accordance with fuzzy logic approach. Further information on the algorithm can be found in [43].

In the present study, BTS-FCM is the second approach used for brain tumor segmentation. The approach is based on a 3-class fuzzy c-means clustering structure. FCM input is set as “cut between the middle and large class”.

5. BTS-SOM approach

Introduced by Kohonen [35], SOM is an artificial neural network learning method. It contains a competitive artificial neural network and operates in two steps [44]. The first step involves creating an artificial neural network through a training set and a map through the competitive learning network. In the second step, the algorithm rapidly predicts the clusters of new data on the map [45]. The structure of SOM network is shown in **Fig. 6**.

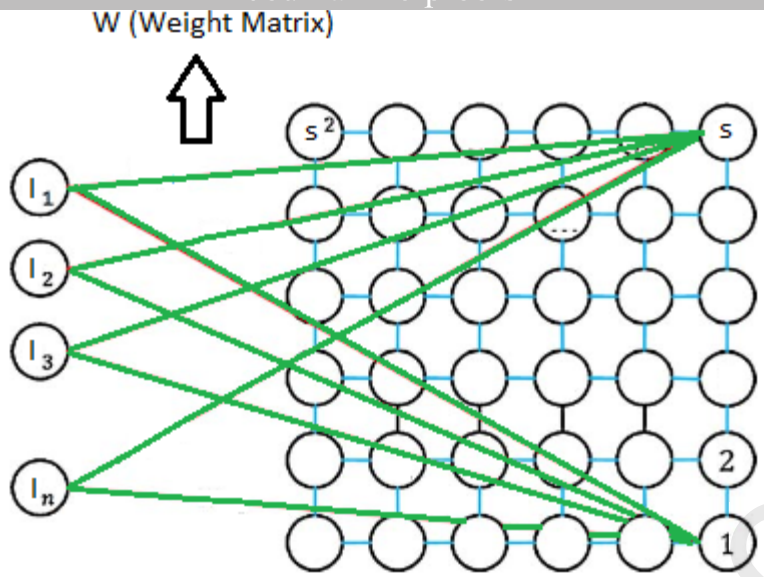


Fig. 6. Architecture of the SOM network

where (I1, I2, ... , Ii, ... , In) are a set of continuous-valued input patterns. I has a fixed dimension n. The SOM network has an array of S=s × s neurons. W is weight matrix. K-means is one of the most widely used unsupervised learning approaches. The approach relies on data for clustering process which aims to put all data in a single cluster. In K-means clustering approach, m number of data is grouped in N number of clusters. This process mainly aims to reach a high level of similarity within the clusters, while it aims to find a low level of similarity among clusters [46,47].

BTS-SOM, which is the third algorithm used for performance comparison in the present study, performs brain tumor segmentation by combining SOM and K-means. Cluster parameter in this approach is used to set the number of segmentation clusters. As a result, the number of segmentation images generated by BTS-SOM is equal to clusters at the end of segmentation process. The number of classes and SOM iterations were set at 4 and 200, respectively.

6. Metric Criteria for Segmentation Performance Comparison

In the present study, Jaccard Similarity Coefficient (JSC), Dice and Figure of Merid (FOM) were used to compare segmentation performances of BTS-ELM-FRFCM, BTS-SVM, BTS-FCM and BTS-SOM approaches. These metrics are described in detail in the following sections.

6.2. JSC

Journal Pre-proofs

Journal Pre-proofs

JSC is often used to evaluate the segmentation performance of an approach. JSC compares two different images in order to measure segmentation performance depending on the following formula [13]:

$$\mathrm{JSC} = \frac{|\mathrm{SI} \cap \mathrm{GT}|}{|\mathrm{SI} \cup \mathrm{GT}|} \tag{8}$$

where SI is segmented_image, and GT represents ground truth. SI corresponds to the segmented image obtained by the segmentation approach. GT corresponds to the complete and accurate segmentation result obtained manually. JSC value varies between 0 and 1, and the segmentation performance increases as the value converges to 1.

6.2. DC

DC is another popular metric used for the measurement of segmentation performance. It measures the segmentation performance using the mathematical formula given in Eq. 9 [48]. DC values varies between 0 and 1, the similarity increases as the value converges to 1.

$$\mathrm{DC} = \frac{2 * |\mathrm{SI} \cap \mathrm{GT}|}{|\mathrm{SI}| + |\mathrm{GT}|} \tag{9}$$

where SI is segmented_image, and GT represents ground truth.

6.3. FOM

FOM, which was developed by Pratt [49] is another metric used for the measurement of image processing performance. It benefits from the following mathematical formula:

$$\mathrm{FOM} = \frac{1}{\max(ESI, EGT)} \sum_{i=1}^{EGT} \frac{1}{1 + d_m{}^2(i)} \tag{10}$$

where ESI (Edge segmented_image) is the number of edge pixels in the segmentation result obtained using edge detection approach, while EGT (Edge Ground Truth) represents the number of edge pixels in ground truth images. m (i) denotes the distance between ith boundary pixel and the

Journal Pre-proofs

nearest ground-truth pixel. d is the scaling constant. FOM values vary between 0 and 1, and the segmentation performance increases as this value converges to 1.

7. Experimental results

As described in Section "2.5.1. Training process" in detail, 3200 training data were used for BTS-ELM-FRFCM training. Following this process, as described in Section "2.5.2. Testing process" in detail, the segmentation of 20 different MRI images was performed using BTS-ELM-FRFCM approach. In addition, BTS-SVM, BTS-FCM, BTS-SOM approaches were also applied to 20 MRI images used in the testing process of the proposed approach in order to compare the segmentation performance of BTS-ELM-FRFCM with different approaches. The findings obtained on RPI device with LCD screen following the segmentation process of the first five MRI images using BTS-ELM-FRFCM are shown in **Fig. 6**. The first five test results obtained from MRI images, ground truth images and three approaches in the first five tests are given in **Table 2**.

Journal Pre-proofs

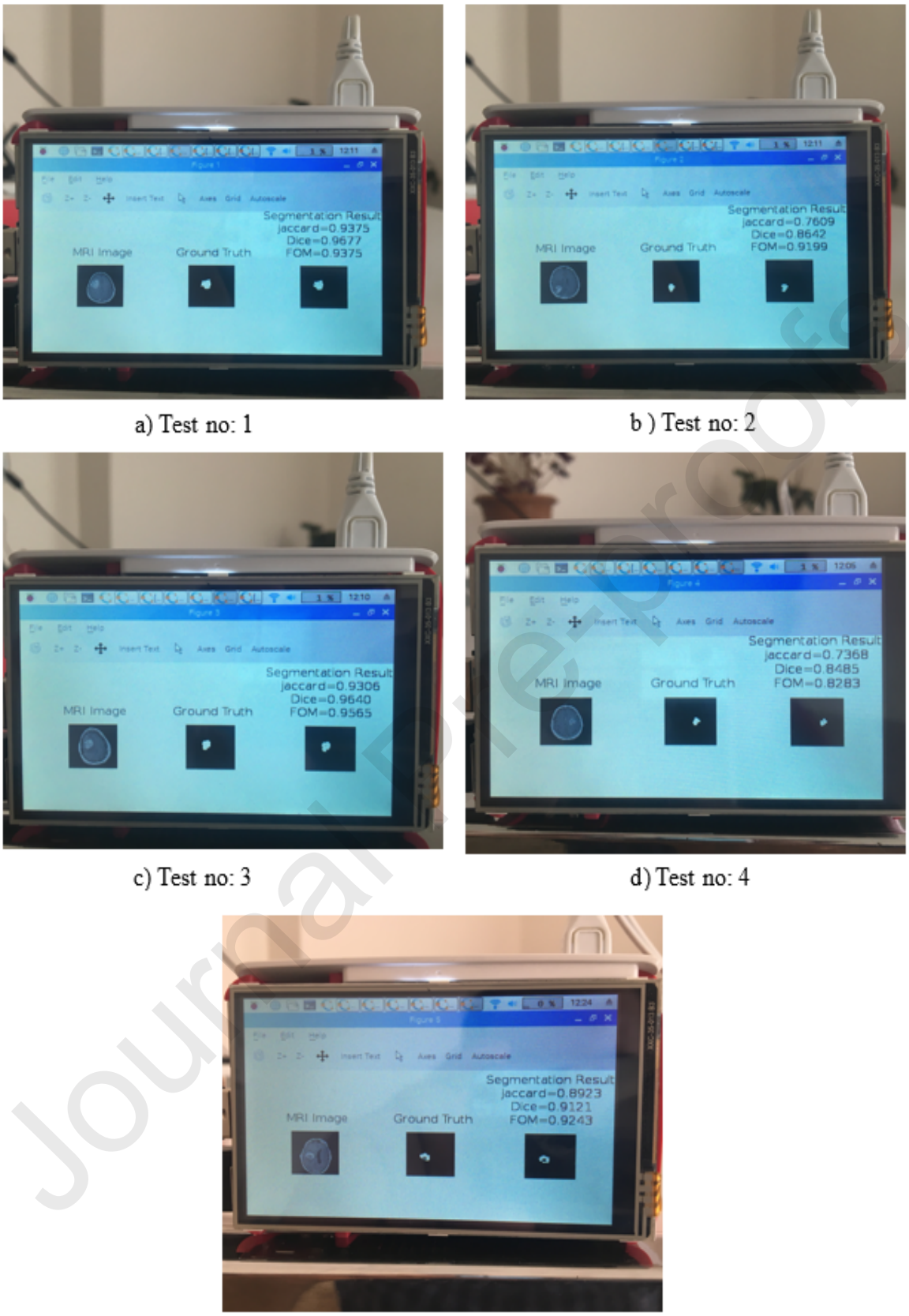
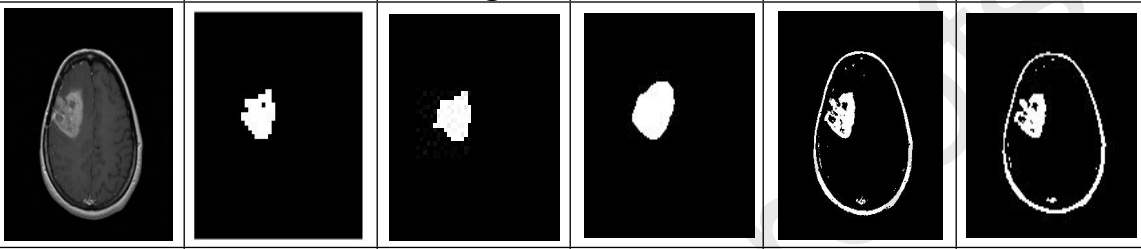
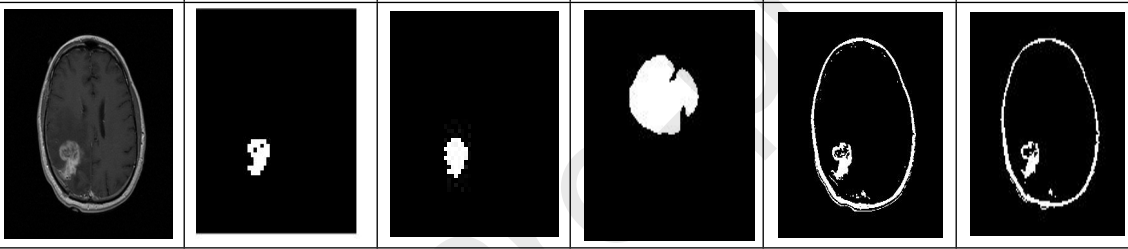
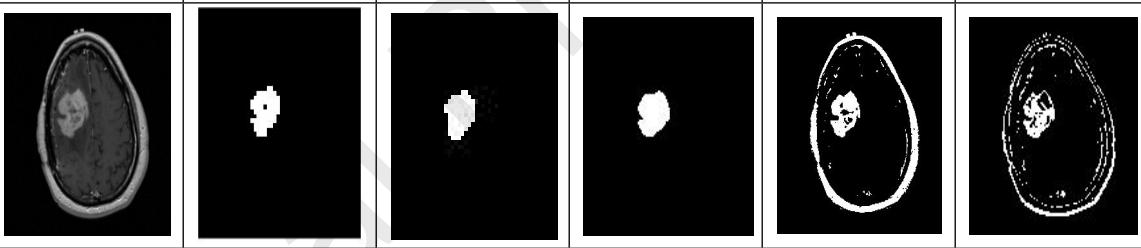
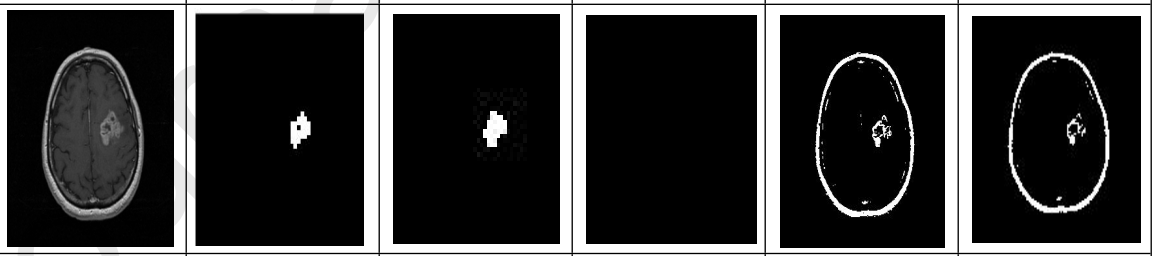
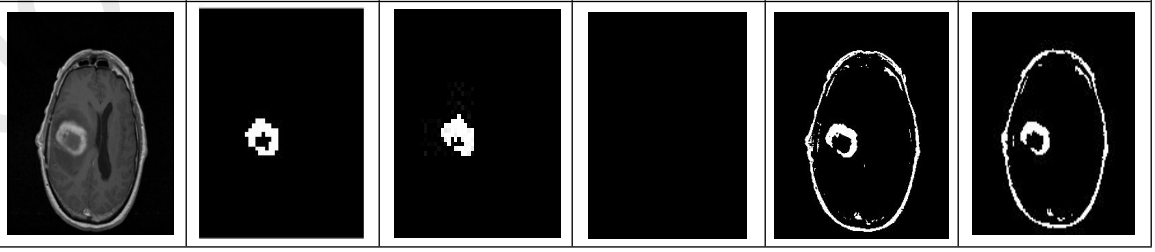


Fig. 7. Test results shown on RPI hardware

Journal Pre-proofs

The results of JSC, FOM and Dice test results obtained from ELM-FRFCM, BTS-SVM, BTS-FCM and BTS-SOM on 20 MRI images are given in **Table 3**.

Table 2. The segmentation images following ELM training

Test	T1-GD MRI Image	Ground truth	BTS-ELM-FRFCM segmented image	BTS-SVM segmented image	BTS-FCM segmented image	BTS-SOM segmented image
1						
2						
3						
4						
5						

Journal Pre-proofs

Table 3. The comparison of the findings with machine learning results

	BTS-ELM-FRFCM			BTS-SVM			BTS-FCM			BTS-SOM		
Test	**FOM**	**Dice**	**JSC**	**FOM**	**Dice**	**JSC**	**FOM**	**Dice**	**JSC**	**FOM**	**Dice**	**JSC**
1	0.9375	0.9677	0.9375	0.5255	0.6890	0.5255	0.4118	0.529	0.3596	0.4016	0.5366	0.3667
2	0.9199	0.8642	0.7609	0	0	0	0.2845	0.3745	0.2845	0.229	0.3438	0.2075
3	0.9565	0.9640	0.9306	0.7311	0.7878	0.6708	0.3101	0.3625	0.2465	0.5446	0.5157	0.3475
4	0.8283	0.8485	0.7368	0	0	0	0.1765	0.2058	0.1421	0.1392	0.1159	0.0615
5	0.9243	0.9121	0.8923	0	0	0	0.289	0.3615	0.2496	0.2632	0.3694	0.2266
6	0.8000	0.8889	0.8000	0	0	0	0.0582	0.0954	0.0501	0.0677	0.1101	0.0583
7	0.9878	0.9891	0.9743	0.6592	0.7946	0.6592	0.1931	0.3096	0.1831	0.1881	0.3069	0.1813
8	0.9399	0.8696	0.8121	0.8203	0.9013	0.8203	0.2566	0.4033	0.2526	0.2604	0.4132	0.2604
9	0.9899	0.8889	0.8000	0.0183	0.036	0.0183	0.0275	0.0535	0.0275	0.0179	0.0351	0.0179
10	0.7959	0.8864	0.7959	0	0	0	0.1335	0.1996	0.1109	0.1508	0.2245	0.1264
11	0.8194	0.9008	0.8194	0.0326	0.0338	0.0172	0.0928	0.0768	0.0399	0.0726	0.0286	0.0145
12	0.7917	0.8837	0.7917	0.0816	0.0962	0.0506	0.1947	0.2177	0.1222	0.1743	0.1911	0.1056
13	0.9811	0.8261	0.7037	0.0123	0	0	0.2734	0.3908	0.2428	0.2921	0.3607	0.22
14	0.9161	0.8800	0.7857	0	0	0	0.0904	0.1288	0.0688	0.1581	0.08	0.0417
15	0.9712	0.9616	0.9511	0	0	0	0.2403	0.3673	0.225	0.2668	0.3974	0.248
16	0.9524	0.9756	0.9524	0.2665	0.4208	0.2665	0.3117	0.4459	0.2869	0.3248	0.4368	0.2794
17	0.9849	0.9444	0.8947	0	0	0	0.1834	0.2685	0.155	0.1932	0.3023	0.1781
18	0.9241	0.8932	0.8612	0	0	0	0.2096	0.3081	0.1821	0.2184	0.34	0.2048
19	0.9686	0.9232	0.8935	0.0167	0	0	0.2417	0.2781	0.1615	0.252	0.2586	0.1485
20	0.8328	0.8421	0.7273	0.0102	0	0	0.1581	0.2432	0.1384	0.1681	0.2581	0.1481

.

1) Various statistical analyses were performed in order to evaluate segmentation results given in Table 3. In this respect, ANOVA analysis was performed on JSC, Dice and FOM results obtained using 4 different segmentation approaches, and the findings are presented in Table 4. It can be observed in Table 4 that mean, minimum and maximum and 95% confidence interval for mean values of JSC, Dice and FOM results obtained using BTS-ELM-FRFCM approach were higher compared to other approaches, and that its standard deviation values were lower compared to other approaches. Therefore, it can be stated that the proposed method displayed a better segmentation performance among all methods.

Table 4. Statistics with ANOVA

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		

Journal Pre-proofs

FOM	BTS-ELM-FRFCM	20	0.9111	0.0712	0.0159	0.8778	0.9444	0.7917	0.9899
	BTS-SVM	20	0.1587	0.2804	0.0627	0.0275	0.2899	0.0000	0.8203
	BTS-FCM	20	0.2068	0.0963	0.0215	0.1618	0.2519	0.0275	0.4118
	BTS-SOM	20	0.2191	0.1189	0.0266	0.1635	0.2748	0.0179	0.5446
	Total	80	0.3740	0.3517	0.0393	0.2957	0.4522	0.0000	0.9899
Dice	BTS-ELM-FRFCM	20	0.9055	0.0476	0.0106	0.8832	0.9278	0.8261	0.9891
	BTS-SVM	20	0.1880	0.3262	0.0729	0.0353	0.3406	0.0000	0.9013
	BTS-FCM	20	0.2810	0.1289	0.0288	0.2207	0.3413	0.0535	0.5290
	BTS-SOM	20	0.2812	0.1507	0.0337	0.2107	0.3518	0.0286	0.5366
	Total	80	0.4139	0.3444	0.0385	0.3373	0.4906	0.0000	0.9891
JSC	BTS-ELM-FRFCM	20	0.8411	0.0826	0.0185	0.8024	0.8797	0.7037	0.9743
	BTS-SVM	20	0.1514	0.2762	0.0618	0.0222	0.2807	0.0000	0.8203
	BTS-FCM	20	0.1765	0.0915	0.0205	0.1336	0.2193	0.0275	0.3596
	BTS-SOM	20	0.1721	0.1026	0.0229	0.1241	0.2201	0.0145	0.3667
	Total	80	0.3353	0.3331	0.0372	0.2611	0.4094	0.0000	0.9743

2) In the second step, ANOVA significance analysis was performed in order to reveal any differences among FOM, Dice and JSC results obtained using 4 different segmentation approaches given in Table 3, and the findings are given in Table 5, which demonstrates that FOM, Dice and JSC were calculated as ($F(3, 76)= 96.291$; $p=0.000<0.05$), ($F(3, 76)= 59.1070$; $p=0.000<0.05$) and ($F(3, 76)= 89.2740$; $p=0.000<0.05$), respectively. It can be thus stated that FOM, Dice and JSC results obtained using 4 different segmentation approaches differed significantly. In order to prove this difference numerically, Scheffe analysis was performed.

Journal Pre-proofs

Journal Pre-proofs

Table 5. Significance report of three metrics with ANOVA.

ANOVA						
		Sum of Squares	**df**	**Mean Square**	**F**	**Sig.**
FOM	Between Groups	7.7350	3	2.578	96.291	0.000
	Within Groups	2.0350	76	0.027		
	Total	9.7700	79			
Dice	Between Groups	6.5600	3	2.1870	59.1070	0.000
	Within Groups	2.8110	76	0.0370		
	Total	9.3710	79			
JSC	Between Groups	6.8290	3	2.2760	89.2740	0.000
	Within Groups	1.9380	76	0.0250		
	Total	8.7670	79			

3) In the third step, Scheffe analysis was performed on FOM, Dice and JSC results as given in Table 6-8. It can be understood from Table 6 that the highest Mean Difference (I-J) results were calculated in the first three lines of Table 6 where Scheffe analysis on FOM, Dice and JSC results are given. These findings indicate that arithmetic means of FOM results obtained using BTS-ELM-FRFCM approach were higher compared to other approaches. The lowest Mean Difference (I-J) results were obtained between Line 4-6 in Table 6. Therefore, mean values of FOM results obtained using BTS-SVM approach were lower compared to those of other approaches.

Journal Pre-proofs

Table 6. Scheffe comparison of FOM.

Multiple Comparisons							
Scheffe							
Dependent Variable			**Mean Difference (I-J)**	**Std. Error**	**Sig.**	**95% Confidence Interval**	
						Lower Bound	**Upper Bound**
FOM	BTS-ELM-FRFCM	BTS-SVM	0.7524*	0.0517	0.0000	0.6044	0.9004
		BTS-FCM	0.7042*	0.0517	0.0000	0.5563	0.8522
		BTS-SOM	0.6919*	0.0517	0.0000	0.5440	0.8399
	BTS-SVM	BTS-ELM-FRFCM	-0.7524*	0.0517	0.0000	-0.9004	-0.6044
		BTS-FCM	-0.0481	0.0517	0.8337	-0.1961	0.0998
		BTS-SOM	-0.0604	0.0517	0.7148	-0.2084	0.0875
	BTS-FCM	BTS-ELM-FRFCM	-0.7042*	0.0517	0.0000	-0.8522	-0.5563
		BTS-SVM	0.0481	0.0517	0.8337	-0.0998	0.1961
		BTS-SOM	-0.0123	0.0517	0.9965	-0.1603	0.1357
	BTS-SOM	BTS-ELM-FRFCM	-0.6919*	0.0517	0.0000	-0.8399	-0.5440
		BTS-SVM	0.0604	0.0517	0.7148	-0.0875	0.2084
		BTS-FCM	0.0123	0.0517	0.9965	-0.1357	0.1603

*The mean difference is significant at the 0.05 level.

4) Scheffe analysis was performed on Dice results as given in Table 7, and the highest Mean Difference (I-J) results were obtained in the first three lines of Table 7. These findings demonstrate that arithmetic means of Dice results obtained using BTS-ELM-FRFCM approach were higher compared to those of other approaches. It is also observed that the lowest Mean Difference (I-J) results were obtained between Line 4-6 in Table 7. Therefore, it is evident that arithmetic means of Dice results obtained using BTS-SVM approach were lower compared to those of other approaches.

Journal Pre-proofs

Table 7. Scheffe comparison of Dice.

Multiple Comparisons							
Scheffe							
Dependent Variable			**Mean Difference (I-J)**	**Std. Error**	**Sig.**	**95% Confidence Interval**	
						Lower Bound	**Upper Bound**
Dice	BTS-ELM-FRFCM	BTS-SVM	0.7175*	0.0608	0.0000	0.5436	0.8914
		BTS-FCM	0.6245*	0.0608	0.0000	0.4506	0.7984
		BTS-SOM	0.6242*	0.0608	0.0000	0.4504	0.7982
	BTS-SVM	BTS-ELM-FRFCM	-0.7175*	0.0608	0.0000	-0.8914	-0.5436
		BTS-FCM	-0.0930	0.0608	0.5089	-0.2669	0.0809
		BTS-SOM	-0.0933	0.0608	0.5066	-0.2672	0.0806
	BTS-FCM	BTS-ELM-FRFCM	-0.6245*	0.0608	0.0000	-0.7984	-0.4506
		BTS-SVM	0.0930	0.0608	0.5089	-0.0809	0.2669
		BTS-SOM	- 0.0002	0.0608	1.0000	-0.1741	0.1737
	BTS-SOM	BTS-ELM-FRFCM	- 0.6242*	0.0608	0.0000	-0.7982	-0.4504
		BTS-SVM	0.0933	0.0608	0.5066	-0.0806	0.2672
		BTS-FCM	0.0002	0.0608	1.0000	-0.1737	0.1741

* The mean difference is significant at the 0.05 level.

5) Scheffe analysis was performed on JSC results as given in Table 8, and the highest Mean Difference (I-J) results were obtained in the first three lines of Table 8. These findings demonstrate that arithmetic means of JSC results obtained using BTS-ELM-FRFCM approach were higher compared to those of other approaches. It is also observed that the lowest Mean Difference (I-J) results were obtained between Line 4-6 in Table 8. Therefore, it can be concluded that arithmetic means of JSC results obtained using BTS-SVM approach were lower compared to those of other approaches.

Journal Pre-proofs

Journal Pre-proofs

In conclusion, it was proven in Table 4-8 that FOM, Dice and JSC results obtained using the proposed BTS-ELM-FRFCM approach were statistically higher compared to other edge detection approaches. Thus, it was clearly indicated that the proposed method display a better segmentation performance compared to other approaches. In addition, it was also statistically indicated that BTS-SVM approach displayed a worse performance compared to other approaches. In this respect, it is evident that the proposed ELM-FRFCM approach functioned properly and successfully on RPI hardware.

Table 8. Scheffe comparison of JSC.

Multiple Comparisons							
Scheffe							
Dependent Variable			**Mean Difference (I-J)**	**Std. Error**	**Sig.**	**95% Confidence Interval**	
						Lower Bound	**Upper Bound**
JSC	BTS-ELM-FRFCM	BTS-SVM	0.6896*	0.0505	0.0000	0.5453	0.8340
		BTS-FCM	0.6646*	0.0505	0.0000	0.5202	0.8090
		BTS-SOM	0.6689*	0.0505	0.0000	0.5245	0.8133
	BTS-SVM	BTS-ELM-FRFCM	-0.6896*	0.0505	0.0000	-0.8340	-0.5453
		BTS-FCM	-0.0250	0.0505	0.9697	-0.1694	0.1193
		BTS-SOM	-0.0207	0.0505	0.9824	-0.1651	0.1237
	BTS-FCM	BTS-ELM-FRFCM	-0.6646*	0.0505	0.0000	-0.8090	-0.5202
		BTS-SVM	0.0250	0.0505	0.9697	-0.1193	0.1694
		BTS-SOM	0.0043	0.0505	0.9998	-0.1401	0.1487
	BTS-SOM	BTS-ELM-FRFCM	-0.6689*	0.0505	0.0000	-0.8133	-0.5245
		BTS-SVM	0.0207	0.0505	0.9824	-0.1237	0.1651
		BTS-FCM	-0.0043	0.0505	0.9998	-0.1487	0.1401

* The mean difference is significant at the 0.05 level.

The comparative graphs of JSC, FOM and Dice tests obtained using BTS-ELM-FRFCM, BTS-SVM, BTS-FCM and BTS-SOM in Table 2 are shown in **Fig. 7-9**. It can be understood from these

figures that the proposed BTS-ELM-FRFCM approach displayed a significantly higher segmentation performance compared to other methods.

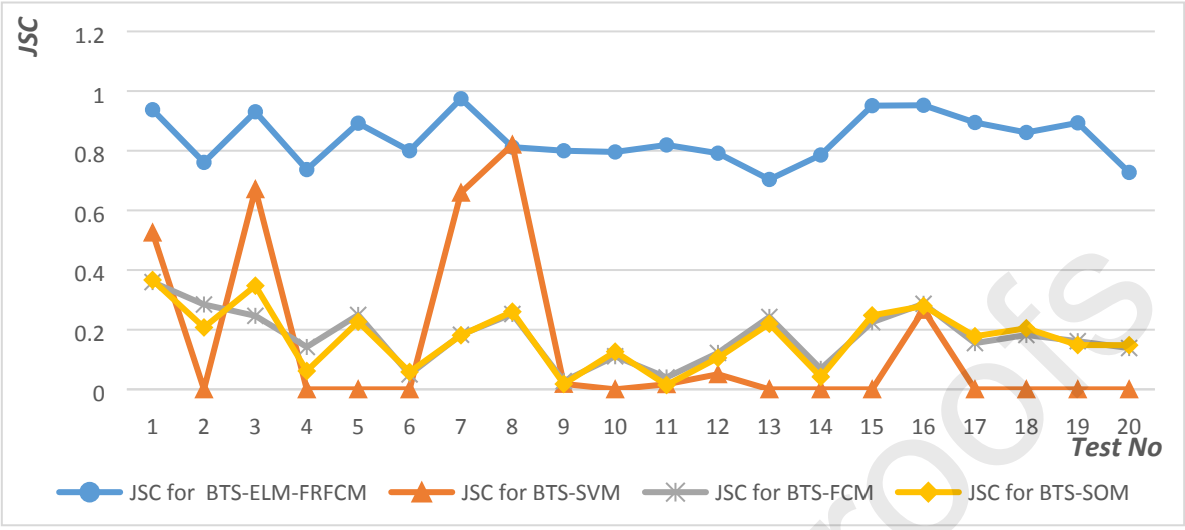


Fig. 8. JSC graph for three methods

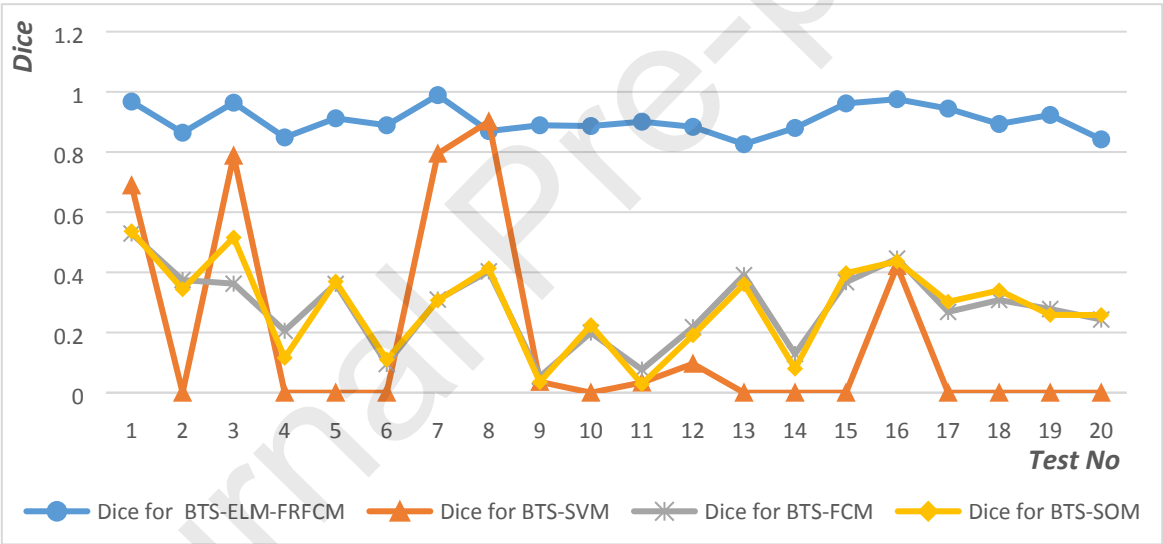


Fig. 9. Dice graph for three methods

Journal Pre-proofs

Journal Pre-proofs

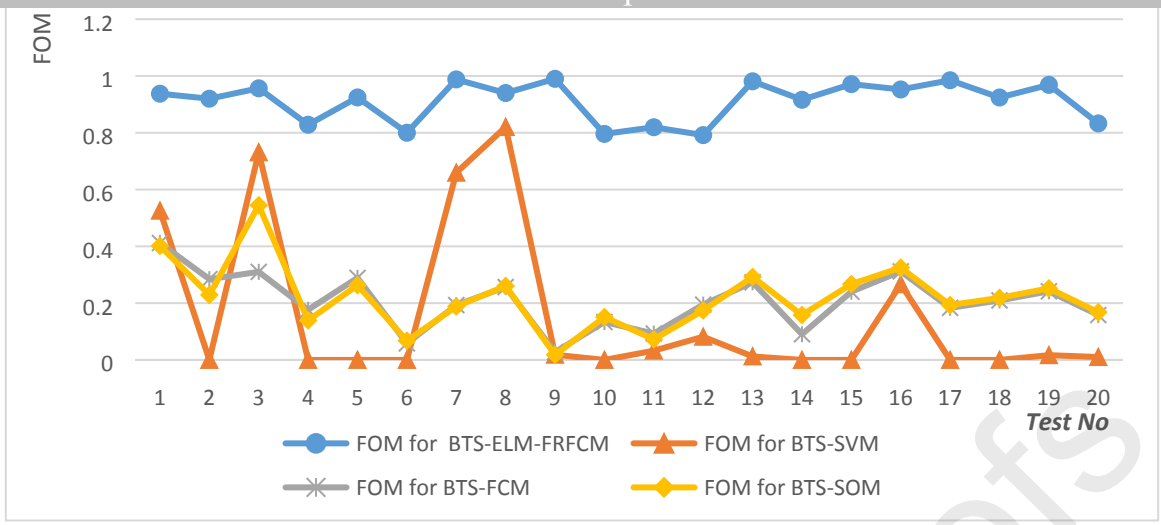


Fig. 10. FOM graph for three methods

8. Conclusion

In the present study, a brain tumor segmentation system was designed using ELM and FRFCM algorithms on RPI hardware. The greatest advantages of RPI devices are their effectiveness and low cost, while they also display high hardware features. Therefore, ELM and FRFCM, which are successful and quick algorithms used in segmentation processes, were used in the present study. It was observed in the experimental studies that BTS-ELM-FRFCM approach proposed for RPI hardware performed a quick segmentation process. The proposed BTS-ELM-FRFCM approach was compared with BTS-SVM, BTS-FCM and BTS-SOM using 20 different MRI images in order to measure the segmentation performances of these four methods. As a result, the mean values of JSC, Dice and FOM analysis on the segmentation results of the proposed method were calculated as 0.8411, 0.9055 and 0.9111, respectively. In the light of statistical analyses such as ANOVA and Scheffe performed on the results of JSC, Dice and FOM results, it can be concluded that the proposed method statistically displayed a better segmentation performance compared to BTS-SVM, BTS-FCM and BTS-SOM. Therefore, a segmentation system which operates on two different algorithms was offered for the health sector.

Journal Pre-proofs

Journal Pre-proofs

REFERENCES

[1] Sert E. A new modified neutrosophic set segmentation approach. Computers & Electrical Engineering 2018;65:576–92. doi:10.1016/j.compeleceng.2017.01.017.

[2] Sert E, Alkan A. Image edge detection based on neutrosophic set approach combined with Chan–Vese algorithm. International Journal of Pattern Recognition and Artificial Intelligence 2019;33(03):1954008. doi:10.1142/S0218001419540089.

[3] Zotin A, Simonov K, Kurako M, Hamad Y, Kirillova S. Edge detection in MRI brain tumor images based on fuzzy c-means clustering. Procedia Computer Science 2018;126:1261–70. doi:10.1016/j.procs.2018.08.069.

[4] Yilmaz B, Durdu A, Emlik GD. A new method for skull stripping in brain MRI using multistable cellular neural networks. Neural Computing and Applications 2018;29(8):79–95. doi:10.1007/s00521-016-2834-2.

[5] Chandra GR, Rao KRH. Tumor detection in brain using genetic algorithm. Procedia Computer Science 2016;79:449–57. doi:10.1016/j.procs.2016.03.058.

[6] Mittal M, Goyal LM, Kaur S, Kaur I, Verma A, Jude Hemanth D. Deep learning based enhanced tumor segmentation approach for MR brain images. Applied Soft Computing Journal 2019;78:346–54. doi:10.1016/j.asoc.2019.02.036.

[7] Havaei M, Davy A, Warde-Farley D, Biard A, Courville A, Bengio Y, et al. Brain tumor segmentation with deep neural networks. Medical Image Analysis 2017;35:18–31. doi:10.1016/j.media.2016.05.004.

[8] Rehman ZU, Naqvi SS, Khan TM, Khan MA, Bashir T. Fully automated multi-parametric brain tumour segmentation using superpixel based classification. Expert Systems with Applications 2019;118:598–613. doi:10.1016/j.eswa.2018.10.040.

[9] Selvapandian A, Manivannan K. Fusion based glioma brain tumor detection and segmentation using ANFIS classification. Computer Methods and Programs in Biomedicine 2018;166:33–8. doi:10.1016/j.cmpb.2018.09.006.

Journal Pre-proofs

[10] Bergner N, Bocklitz T, Romeike BFM, Reichart R, Kalff R, Krafft C, et al. Identification of primary tumors of brain metastases by raman imaging and support vector machines. Chemometrics and Intelligent Laboratory Systems 2012;117:224–32. doi:10.1016/j.chemolab.2012.02.008.

[11] Şişik F, Sert E. Support vector machine working on FPGA and the segmentation method of brain MR screening. International Journal of Innovative Research in Education 2017;4(3):120. doi:10.18844/ijire.v4i3.2549.

[12] Pinto A, Pereira S, Rasteiro D, Silva CA. Hierarchical brain tumour segmentation using extremely randomized trees. Pattern Recognition 2018;82:105–17. doi:10.1016/j.patcog.2018.05.006.

[13] Sert E, Avci D. Brain tumor segmentation using neutrosophic expert maximum fuzzy-sure entropy and other approaches. Biomedical Signal Processing and Control 2019;47:276–87. doi:10.1016/j.bspc.2018.08.025.

[14] Özyurt F., Sert E., Avci E, Dogantekin, E. Brain tumor detection based on Convolutional Neural Network with neutrosophic expert maximum fuzzy sure entropy. Measurement 2019;147: 106830. doi:10.1016/j.measurement.2019.07.058

[15] Sert E., Özyurt F, Doğantekin A. A new approach for brain tumor diagnosis system: single image super resolution based maximum fuzzy entropy segmentation and convolutional neural network. Medical hypotheses 2019;133:109413. doi:10.1016/j.mehy.2019.109413

[16] Özyurt F, Sert E, Avcı D. An expert system for brain tumor detection: fuzzy c-means with super resolution and convolutional neural network with extreme learning machine. Medical Hypotheses 2020; 134: 109433 doi:10.1016/j.mehy.2019.109433

[17] Zhu C, Zou B, Zhao R, Cui J, Duan X, Chen Z, et al. Retinal vessel segmentation in colour fundus images using extreme learning machine. Computerized Medical Imaging and Graphics 2017;55:68–77. doi:10.1016/j.compmedimag.2016.05.004.

[18] Zeng YZ, Zhao YQ, Liao M, Zou BJ, Wang XF, Wang W. Liver vessel segmentation based on extreme learning machine. Physica Medica 2016;32(5):709–16. doi:10.1016/j.ejmp.2016.04.003.

Journal Pre-proofs

Journal Pre-proofs

[19] Huang H, Ma H, JW van Triest H, Wei Y, Qian W. Automatic detection of neovascularization in retinal images using extreme learning machine. Neurocomputing 2018;277:218–27. doi:10.1016/j.neucom.2017.03.093.

[20] Xie W, Li Y, Ma Y. Breast mass classification in digital mammography based on extreme learning machine. Neurocomputing 2016;173:930–41. doi:10.1016/j.neucom.2015.08.048.

[21] Sai P-K, Wang J-G, Teoh E-K. Facial age range estimation with extreme learning machines. Neurocomputing 2015;149:364–72. doi:10.1016/j.neucom.2014.03.074.

[22] Nguyen TV, Mirza B. Dual-layer kernel extreme learning machine for action recognition. Neurocomputing 2017;260:123–30. doi:10.1016/j.neucom.2017.04.007.

[23] Zhao L, Wang X. Seemingly unrelated extreme learning machine. Neurocomputing 2019;355:134–42. doi:10.1016/j.neucom.2019.04.067.

[24] Shang Z, He J. Confidence-weighted extreme learning machine for regression problems. Neurocomputing 2015;148:544–50. doi:10.1016/j.neucom.2014.07.009.

[25] Anandhalli M, Baligar VP. A novel approach in real-time vehicle detection and tracking using raspberry pi. Alexandria Engineering Journal 2018;57(03):1597–607. doi:10.1016/j.aej.2017.06.008.

[26] Sajjad M, Nasir M, Ullah FUM, Muhammad K, Sangaiah AK, Baik SW. Raspberry pi assisted facial expression recognition framework for smart security in law-enforcement services. Information Sciences 2019;479:416–31. doi:10.1016/j.ins.2018.07.027.

[27] Diallo AD, Gobee S, Durairajah V. Autonomous tour guide robot using embedded system control. Procedia Computer Science 2015;76:126–33. doi:10.1016/j.procs.2015.12.302.

[28] Lam Loong Man CKY, Koonjul Y, Nagowah L. A low cost autonomous unmanned ground vehicle. Future Computing and Informatics Journal 2018;3(2):304–20. doi:10.1016/j.fcij.2018.10.001.

[29] Yildiz MZ, Boyraz ÖF. Development of a low-cost microcomputer based vein imaging system. Infrared Physics & Technology 2019;98:27–35. doi:10.1016/j.infrared.2019.02.010.

Journal Pre-proofs

[30] Wang C, Ko M-C, Chen Y-M, Chen L-Q, Lin C-W. An automatic multi-thread image segmentation embedded system for surface plasmon resonance sensor. Sensors and Actuators A: Physical 2019;285:603–12. doi:10.1016/j.sna.2018.12.007.

[31] Tu GJ, Hansen MK, Kryger P, Ahrendt P. Automatic behaviour analysis system for honey bees using computer vision. Computers and Electronics in Agriculture 2016;122:10–8. doi:10.1016/j.compag.2016.01.011.

[32] Lei T, Jia X, Zhang Y, He L, Meng H, Nandi AK. Significantly fast and robust fuzzy c-means clustering algorithm based on morphological reconstruction and membership filtering. IEEE Trans Fuzzy Syst 2018;26(05):3027–41. doi:10.1109/TFUZZ.2018.2796074.

[33] Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995;20(03):273–97. doi:10.1007/BF00994018.

[34] Bezdek JC, Ehrlich R, Full W. FCM: the fuzzy c-means clustering algorithm. Computers & Geosciences 1984;10(02-03):191–203. doi:10.1016/0098-3004(84)90020-7.

[35] Kohonen T, Huang TS, Schroeder MR. Self-organizing maps. Berlin, Heidelberg: Springer Berlin / Heidelberg; 2012.

[36] Scarpace L, Mikkelsen T, Cha S, Rao S, Tekchandani S, Gutman D, et al. Radiology data from the cancer genome atlas glioblastoma multiforme [TCGA-GBM] collection 2016. doi:10.7937/k9/tcia.2016.rnyfuye9.

[37] Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The cancer ımaging archive (TCIA): maintaining and operating a public ınformation repository. Journal of Digital Imaging 2013;26(06):1045–1057. doi:10.1007/s10278-013-9622-7.

[38] Scarpace L, Flanders AE, Jain R, Mikkelsen T, Andrews DW. Data from rembrandt 2015. doi:10.7937/k9/tcia.2015.588ozuzb.

[39] Huang G-B, Zhu Q-Y, Siew C-K. Extreme learning machine: theory and applications. Neurocomputing 2006;70(01-03):489–501. doi:10.1016/j.neucom.2005.12.126.

Journal Pre-proofs

[40] McDonnell MD, Tissera MD, Vladusich T, van Schaik A, Tapson J. Fast, simple and accurate handwritten digit classification by training shallow neural network classifiers with the 'extreme learning machine' algorithm. Plos One 2015;10(08):e0134254. doi:10.1371/journal.pone.0134254.

[41] Tang J, Deng C, Huang G-B, Zhao B. Compressed-domain ship detection on spaceborne optical image using deep neural network and extreme learning machine. IEEE Trans Geosci Remote Sensing 2015;53(03):1174–85. doi:10.1109/TGRS.2014.2335751.

[42] Golestaneh P, Zekri M, Sheikholeslam F. Fuzzy wavelet extreme learning machine. Fuzzy Sets and Systems 2018;342:90–108. doi:10.1016/j.fss.2017.12.006.

[43] Bezdek JC. Pattern Recognition with fuzzy objective function algorithms. Boston, MA: Springer US; 1981.

[44] Haykin SS. Neural networks: a comprehensive foundation. 2nd ed. Upper Saddle River, N.J: Prentice Hall; 1999.

[45] Chi D. Self-Organizing Map-based color image segmentation with k -means clustering and saliency map. ISRN Signal Processing 2011;2011:1–18. doi:10.5402/2011/393891.

[46] Yao H, Duan Q, Li D, Wang J. An improved k -means clustering algorithm for fish image segmentation. Mathematical and Computer Modelling 2013;58(03-04):790–8. doi:10.1016/j.mcm.2012.12.025.

[47] Sert E, Okumus IT. Segmentation of mushroom and cap width measurement using modified k-means clustering algorithm. AEEE 2014;12(04):354–60. doi:10.15598/aeee.v12i4.1200.

[48] Joskowicz L, Cohen D, Caplan N, Sosna J. Automatic segmentation variability estimation with segmentation priors. Medical Image Analysis 2018;50:54–64. doi:10.1016/j.media.2018.08.006.

[49] Pratt WK. Digital image processing. Hoboken, N.J.: Wiley-Liss; 2002.

Journal Pre-proofs

Journal Pre-proofs

Brain Tumor Segmentation Approach Based on the Extreme Learning Machine and Significantly Fast and Robust Fuzzy C-Means Clustering Algorithms Running on Raspberry Pi Hardware

Fatih ŞİŞİK[1], Eser SERT[2*]

[1] Göksun Vocational School, Department of Computer Programming, Kahramanmaras Sutcu Imam University, K.Maras, Turkey
[2] Department of Computer Engineering, Engineering and Architecture Faculty, Kahramanmaras Sutcu Imam University, K.Maras, Turkey
∗ Corresponding author e-mail address: esersert@ksu.edu.tr

Journal Pre-proofs

Journal Pre-proofs

ABSTRACT

Automatic decision support systems have gained importance in health sector in recent years. In parallel with recent developments in the fields of artificial intelligence and image processing, embedded systems are also used in decision support systems for tumor diagnosis. Extreme learning machine (ELM), is a recently developed, quick and efficient algorithm which can quickly and flawlessly diagnose tumors using machine learning techniques. Similarly, significantly fast and robust fuzzy C-means clustering algorithm (FRFCM) is a novel and fast algorithm which can display a high performance. In the present study, a brain tumor segmentation approach is proposed based on extreme learning machine and significantly fast and robust fuzzy C-means clustering algorithms (BTS-ELM-FRFCM) running on Raspberry Pi (PRI) hardware. The present study mainly aims to introduce a new segmentation system hardware containing new algorithms and offering a high level of accuracy the health sector. PRI's are useful mobile devices due to their cost-effectiveness and satisfying hardware. 3200 pieces of T1 gadolium (T1-GD) sequence MRI images were used to train ELM in the present study. 20 pieces of MRI images were used for testing process.Figure of merid (FOM), jaccard similarity coefficient (JSC) and Dice indexes were used in order to evaluate the performance of the proposed approach. In addition, the proposed method was compared with brain tumor segmentation based on support vector machine (BTS-SVM), brain tumor segmentation based on fuzzy C-means (BTS-FCM) and brain tumor segmentation based on self-organizing maps and k-means (BTS-SOM). The statistical analysis on FOM, JSC and Dice results obtained using four different approaches indicated that BTS-ELM-FRFCM displayed the highest performance. Thus, it can be concluded that the embedded system designed in the present study can perform brain tumor segmentation with a high accuracy rate.

Key words: Extreme learning machine, Brain tumor segmentation, Segmentation, Magnetic resonance imaging, Significantly fast and robust fuzzy C-means clustering, Raspberry pi

1. Introduction

Journal Pre-proofs

Various new developments have recently occurred in the field of embedded systems, image processing and machine learning. Thanks to the combination of these systems, new support systems can be designed as a contribution to image recognition, segmentation [1], edge detection [2] and data extraction. One of the leading sectors where these systems are widely used is health as they provide doctors with solutions to various problems such as the diagnosis of various diseases and detection of small tumors.

Various studies have so far been conducted on brain tumor segmentation in the literature. Zotin et al. [3] performed edge detection in MRI brain tumor images using fuzzy C-means (FCM). Yilmaz et al. [4] performed skull stripping in brain MRI using multi-stable cellular neural networks. Chandra et al. [5] performed brain tumor detection on MRI images using Genetic Algorithm. Mittal et al. [6] proposed a deep learning-based approach which detected brain tumors on MRI images. Havaei et al. [7] performed brain tumor segmentation using deep neural networks. Additionally, it can be observed in the current literature that brain tumor segmentation was performed using superpixel based classification [8], adaptive neuro-fuzzy inference system (ANFIS) [9], support vector machines (SVM) [10, 11], extremely randomized trees [12] and neutrosophic expert maximum fuzzy-sure Entropy (NS-EMFSE) [13]. Several studies [14-16] have also been conducted in the literature on convolutional neural network (CNN) based brain tumor classification.

ELM has so far been used in various studies. An example is Zhu et al. [17] who benefited from ELM for retinal vessel segmentation. Similarly, Zeng et al. [18] used ELM for liver vessel segmentation. Huang et al. [19] conducted a study on neovascularization using ELM. In addition, ELM algorithm was used for different purposes such as breast mass classification [20], facial age range estimation [21] and action recognition [22].

ELM approach does not possess an iterative structure, thus being trained faster compared to conventional back propagation approaches [23]. Additionally, this approach requires a fairly short amount of time for calculation [24]. Therefore, it can be argued that ELM is a more successful and faster approach compared to numerous neural network based approaches and machine learning

Journal Pre-proofs

techniques. Approaches such as neural network based segmentation, edge detection and object recognition usually perform well. However, the performance of these systems is heavily influenced by the variety of training data. Although FCM approach is widely used for segmentation and edge detection applications in the field of medicine, its greatest drawback is its sensitivity to noise. Similarly, the greatest disadvantage of fuzzy based segmentation approaches is that it encounters difficulties in creating fuzzy rules.

Several approaches were proposed for Raspberry Pi (RPI) hardware in the literature. Anandhalli et al. [25] developed a system for real time vehicle detection and tracking using RPI hardware. Sajjad et al. [26] developed a real-time facial expression recognition system using RPI device as a contribution to law enforcement. In addition to these studies, RPI hardware was also used for a tour guide robot [27], autonomous unmanned ground vehicle [28], vein imaging system [29], image segmentation [30], and automatic behavior analysis system for honey bees [31].

This study proposes a brain tumor segmentation approach based on the extreme learning machine and significantly fast and robust fuzzy C-means clustering algorithms (BTS-ELM-FRFCM) running on RPI hardware. As shown in Figure 3, the proposed BTS-ELM-FRFCM approach first trains ELM. Later, as shown in Figure 4, the trained ELM is used for MRI segmentation. The approach proposed in the present study benefits from FRFCM approach by Lei et al. [32] in order to apply pre-segmentation to MRI images in the training and testing process.

Main contributions of the present study can be summarized as follows:

1. The proposed BTS-ELM-FRFCM approach can be used on RPI embedded system. It can be noted in the literature that the number of segmentation approaches developed for an embedded system is fairly low, which can be considered as one of the most important contributions of the present study.
2. RPI hardware offers a mobile, user-friendly and low-cost device which can be used on a Linux operating system. Therefore, the segmentation approach proposed in the present study was uploaded to a RPI device.

Journal Pre-proofs

3. ELM algorithm is the fastest artificial neural network (ANN) model. In the present study, ELM and FRFCM algorithms were used on RPI hardware in order to propose a brain tumor segmentation with a higher accuracy rate. No studies have been so far conducted in the existing literature as far as the segmentation using algorithms such as ELM and FRFCM on an embedded system platform are concerned.
4. Our previous study on NS-EMFSE [13] focused on brain tumor segmentation in an unsupervised machine learning technique. In the present study, on the other hand, a brain tumor segmentation approach using ELM was proposed in a supervised structure.

The proposed BTS-ELM-FRFCM approach was experimentally compared with brain tumor segmentation based on support vector machine (BTS-SVM), brain tumor segmentation based on fuzzy C-means (BTS-FCM) and brain tumor segmentation based on self-organizing maps and k-means (BTS-SOM) approaches in terms of segmentation performance.

Support vector machine (SVM) [33] is a statistical learning method which is widely used in various applications such as edge detection, segmentation and classification of various objects. Similar to SVM, Fuzzy C-means (FCM) [34] an algorithm used for performance comparison in the present study, is an unsupervised technique widely used for edge detection and segmentation approaches. Introduced by Kohonen [35], self-organizing maps (SOM) is an artificial neural network learning method and widely used in segmentation applications.

The present study is divided into the following sections: The database and proposed method are described in Section 2. BTS-SVM, BTS-FCM and BTS-SOM approaches are described in Section 3, 4 and 5, respectively. The metrics used for the measurement of segmentation performances are described in Section 6. Experimental studies are discussed in Section 7, and the findings are presented in Section 8.

2. Theoretical background

2.1. Database

Journal Pre-proofs

Journal Pre-proofs

In the present study, The Cancer Genome Atlas Glioblastoma Multiforme (TCGA-GBM) [36,37] and REMBRANDT [37,38] data sets in Cancer Imaging Archive were used for ELM training and testing processes. These are open access data sets, and thus do not require ethical committee approval. Brain tumors can be found on MRI images in these datasets, and many researchers resort to these data sets around the world. T1-gadolinium (T1-Gd) sequence MRI images enable a highly realistic brain tumor segmentation. Therefore, the present study benefits from T1-Gd sequence MRI images in these datasets.

2.2. RPI hardware

PRI hardware contains small-size equipment which can also be found in a computer. It has its own operating system, Raspbian, and can also be used on different operating systems. In the present study, RPI 3 was used on a Linux operating system, and Octave was preferred as a software in order to operate the proposed approach. RPI 3 hardware with LCD screen used in the present study is shown in **Fig. 1**.

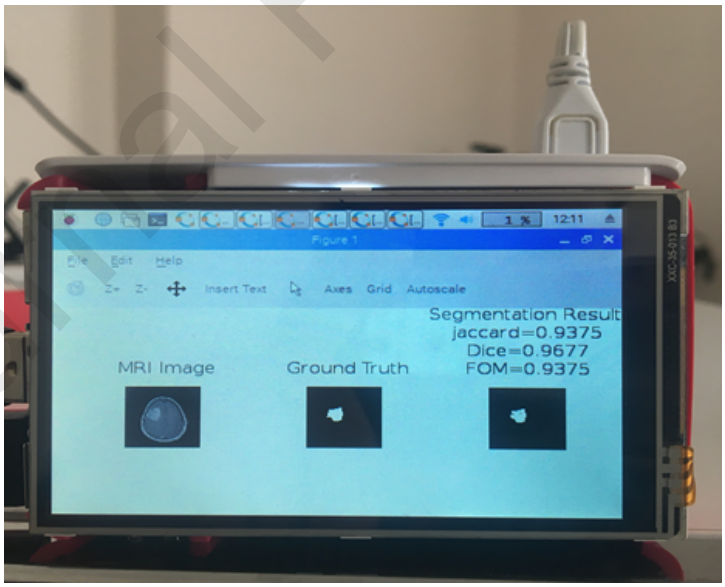


Fig. 1. RPI and LCD

2.3. ELM

ANN is a machine learning method based on human neural system. ELM is a feedforward neural network developed by Huang et al. The structure of single-hidden layer feed forward

Journal Pre-proofs

network (SLFN) [39] is shown in **Fig. 2**. ELM is widely used for regression [24], classification [40], segmentation [17], and object detection [41].

For S optional samples (x_i, t_i) for $\in R^n$ x R^m, the mathematical equation of ELM with SLFN structure is expressed as follows [42]:

$$\sum_{i=1}^{N} \beta_i g(w_i.x_j + b_i) = O_j. \quad j = 1,....S \tag{1}$$

here:

- $x_i = [x_{i1}, x_{i2}, x_{i3}, \ldots, x_{in}] \in R^n$, and $t_i = [t_{i1}, t_{i2}, t_{i3}, \ldots, t_{im}] \in R^m$,
- N is the number of hidden nodes,
- $w_i = [w_{i1}, w_{i2}, w_{i3}, \ldots, w_{in}]^T$ is ith weight vector connected to the hidden node,
- $\beta_i = [\beta_{i1}, \beta_{i2}, \beta_{i3}, \ldots, \beta_{im}]^T$ is the ith weight vector connection between the hidden node and output nodes,
- $b \in R$ is the bias of ith hidden node,
- $Oj = [O_{j1}, O_{j2}, \ldots, O_{jm}]^T$ is the output of SLFN.

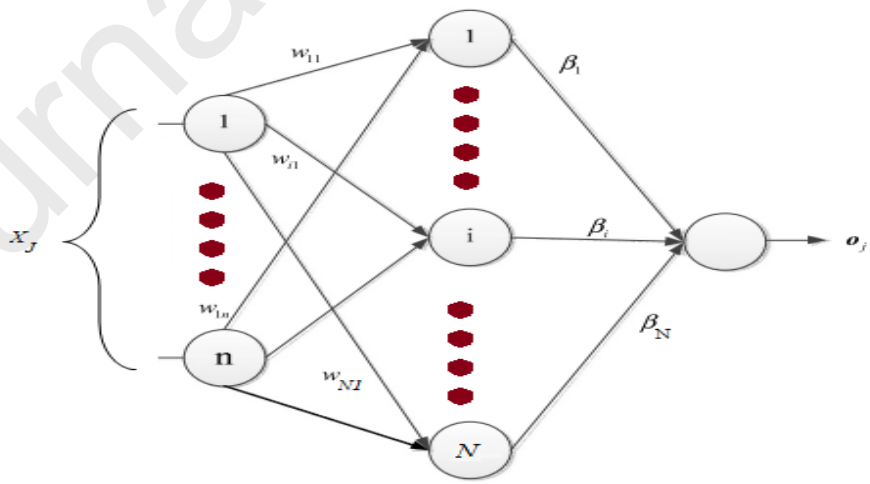


Fig. 2. The structure of SLFN.

Depending on the following formula, ELM converges to 0 mean error for N number of samples:

Journal Pre-proofs

$$\sum_{i=1}^{\infty}\|O_j - t_j\| = 0 \tag{2}$$

where w_i, b_i and β_i is equalized to t_j as follows:

$$\sum_{i=1}^{L}\beta_i g_i(w_i.x_j + b_i) = t_{j.} \qquad j = 1,....N \tag{3}$$

The above-mentioned equations are expressed in matrix form as follows:

$$H\beta = T \tag{4}$$

H is the hidden layer matrix and T is the output matrix. The hidden layer matrix and output matrix are given in Equation 4 and 5.

$$H = \begin{bmatrix} G(w_1.x_1 + b_1) & \dots & G(w_N.x_1 + b_N) \\ \vdots & \ddots & \vdots \\ G(w_1.x_S + b_1) & \dots & G(w_N.x_S + b_N) \end{bmatrix}_{NxL} \tag{5}$$

$$T = [t_1 \quad \cdots \quad t_N]_{NXm} \qquad \text{and} \qquad \beta = [\beta_1 \quad \cdots \quad \beta_N]_{SXm} \tag{6}$$

H is SLFN hidden layer output matrix. After hidden node parameters (a_i, b_i) are randomly selected, ELM is trained using least squares depending on the following formula:

$$\beta\hat{} = H^{+}T \tag{7}$$

Here H^{+} is the generalized inverse Moore-Penrose of H. More information on ELM can be found in [39].

2.4. FRFCM approach

Conventional FCM algorithm is sensitive to noise and thus encounters some problems in the segmentation process. In addition, FCM is sometimes a time-consuming approach for segmentation. Therefore, in order to minimize these problems, Lei et al. [32] proposed FRFCM algorithm which is a highly better and useful approach compared to FCM as it increases segmentation performance

Journal Pre-proofs

using morphological reconstruction and membership filtering. This algorithm displays an effective segmentation performance even in noisy images, which eliminates the need for selecting different filters for different types of noise [32]. Because FRFCM requires a low level of computation compared to the conventional FCM approach, it can be considered as a more fast approach. Further information on FRFCM can be found in [32].

2.5. The proposed approach

BTS-ELM-FRFCM approach proposed in the present study operates in two steps: training and testing process. As shown in Figure 3, 200 MRI images are used at the beginning and, thanks to the algorithm process of the proposed method, 3200 training images are obtained. These images and train label data are used to train ELM. As summarized in Figure 3, 20 MRI images in DICOM format are used to obtain test data thanks to the algorithm process of the proposed method. These obtained data and test labels are used to perform ELM based segmentation. The whole process takes place in RPI hardware.

Journal Pre-proofs

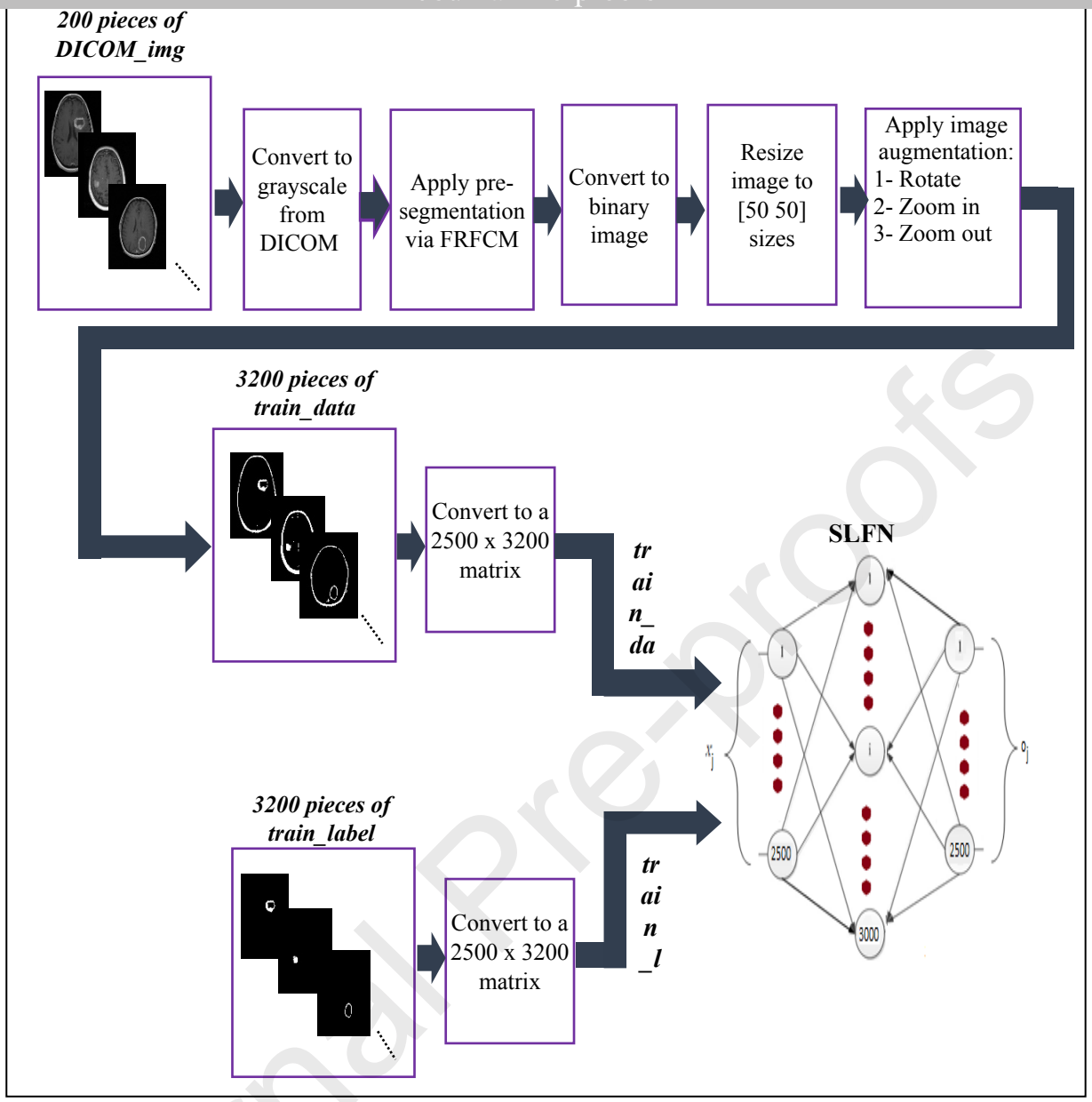


Fig. 3. The summary of BTS-ELM-FRFCM training process

Journal Pre-proofs

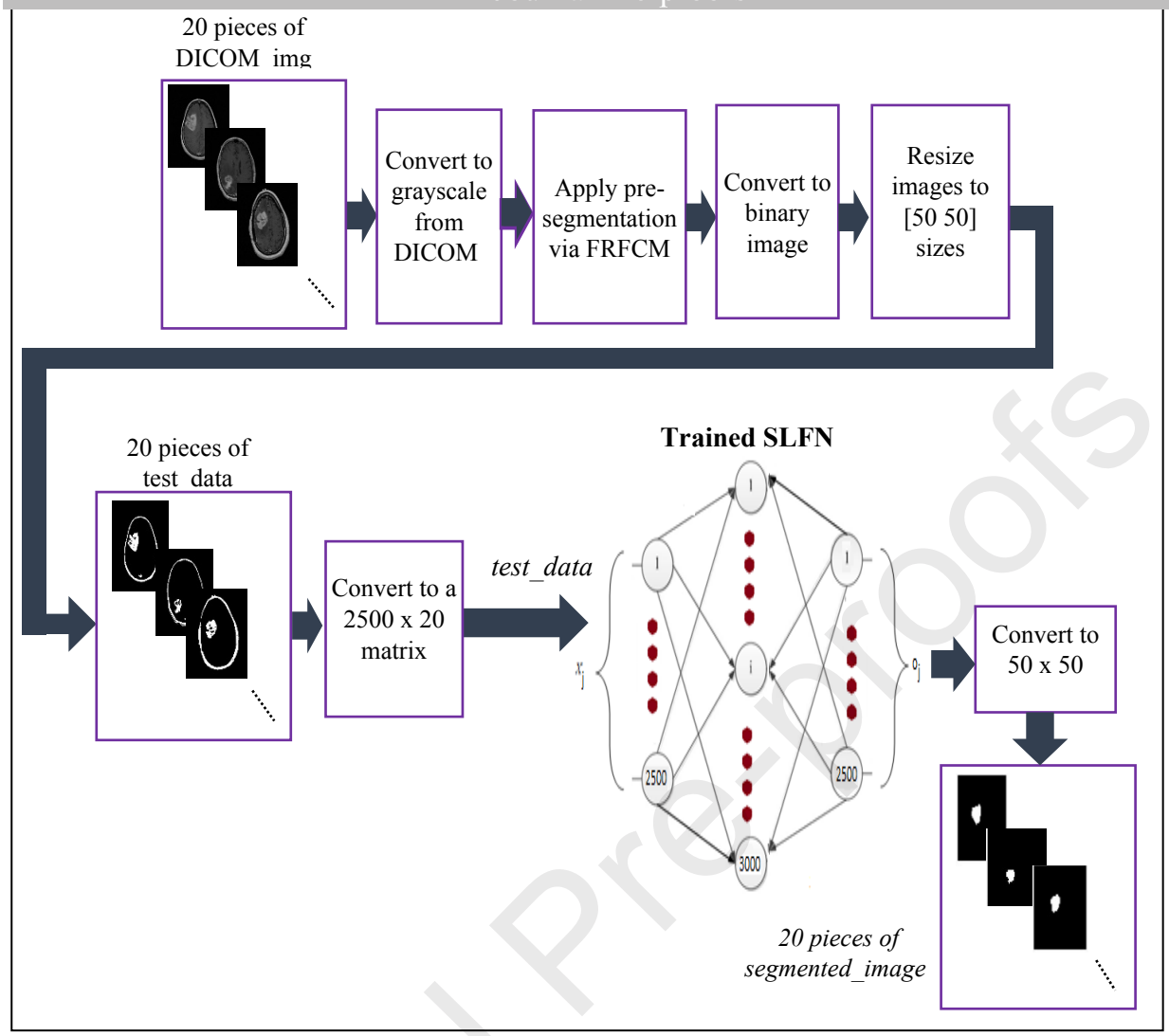


Fig. 4. BTS-ELM-FRFCM testing process

2.5.1. Training process

The details of BTS-ELM-FRFCM approach training process are given in **Algorithm 1**. The number of hidden layers in ELM training and testing process were adjusted as 3000. The experimental studies indicated that this value was sufficient to optimally operate RPI and ELM. The activation function of ELM was selected as sigmoid, and the numbers of input and outputs were adjusted as 2500 and 2500, respectively.

200 MRI images were used at the beginning of the training process. These images are described in detail in the Section "2.1. Database". MRI image in DICOM format is uploaded in **Algorithm 1** Step 1. Because the pixel values of images in DICOM format may vary in a broad range, they were converted to grayscale to make this range vary between 0 and 255. Thus, image processing became

Journal Pre-proofs

much quicker, and no apparent problems were observed in terms of BTS-ELM-FRFCM training quality. In Algorithm 1 Step 6, a pre-segmentation process is applied to images using FRFCM approach developed by Lei et al. [32]. For a quicker ELM training process, images was resized at a dimension of 50 x 50 in Step 5. In **Algorithm 1** Step 6, the image augmentation process was performed after training images had been rotated, zoomed in and zoomed out. In the first step, MRI images are rotated at 30, 60, 90, 120 and 150 degrees. In the second step, the MRI image is zoomed in at 5%, 10%, 15%, 20% and 25%. In the third step, the MRI image is zoomed out at 5%, 10%, 15%, 20% and 25%. In Fig.5, 15 images obtained by applying image augmentation to a sample MRI image are given. At the end of the image augmentation process, 3200 training images are obtained, train_data variable at a dimension of 3200 x 50 x 50 train_data is created.

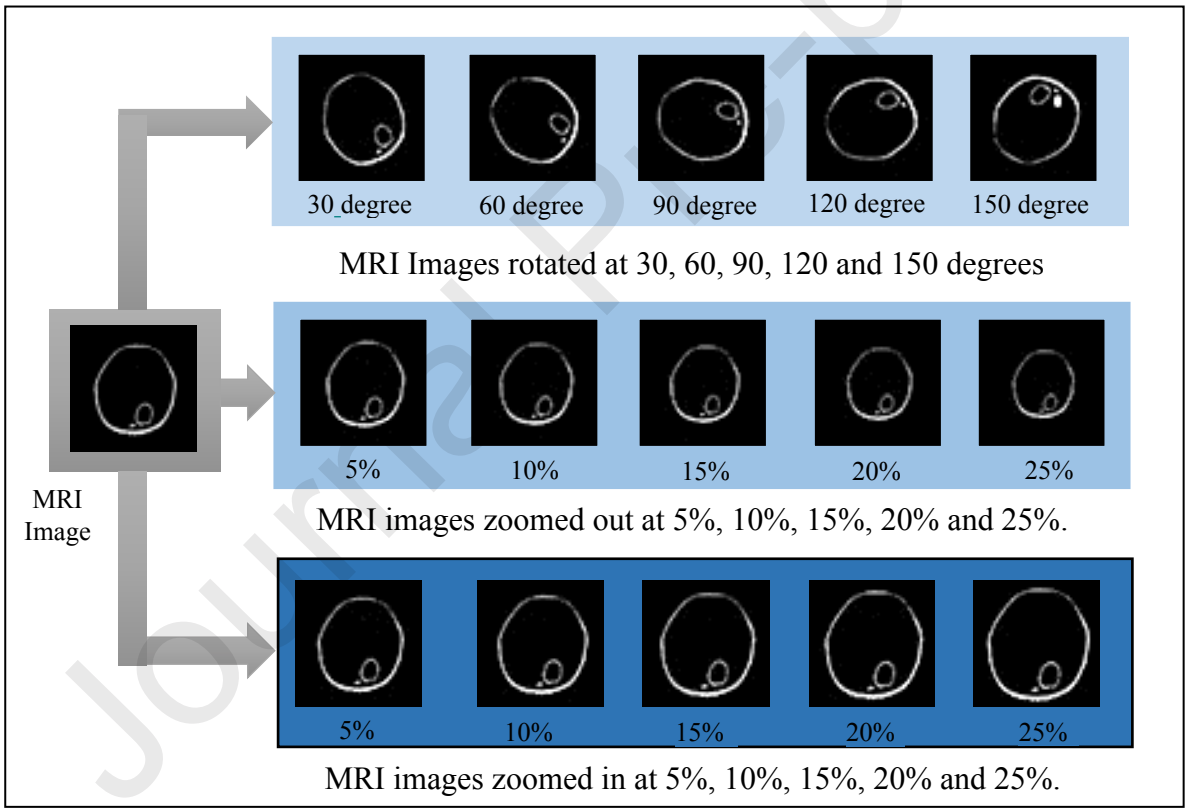


Fig. 5. The image augmentation process

A matrix at a dimension of 50 x 50 x 3200 in train_data variable was obtained, which was later converted to a matrix at a dimension of 2500 x 3200 in **Algorithm 1** Step 7 in order for ELM to process this matrix. ELM training is performed in Step 9. **Algorithm 1** is summarized in **Fig. 3**.

Algorithm 1. The preparation of training data and BTS-ELM-FRFCM training

Step 1: Use 200 MRI images in DICOM format

Step 2: Convert images to Grayscale format

Step 3: Apply pre-segmentation to images using **FRFCM** algorithm

Step 4: Convert images to binary format

Step 5: resized image [50 50], create 200 x 50 x 50 training images

Step 6: Increase training images to 3200 by applyling rotate, zoom in and zoom out processes and store them in **train_data**, create 3200 x 50 x 50 train images

Step 7: Resize **train_data** at 2500 x 3200

Step 8: Prepare **test_label** for 3200 test data and resize them at 2500 x 3200

Step 9: Train ELM using (**train_data, train_label**)

2.5.2. Testing process

Testing process of the proposed BTS-ELM-FRFCM approach is performed as shown in **Algorithm 2.** MRI images in DICOM format is uploaded in **Algorithm 1** Step 1. In Step 2, they were converted to grayscale. In Step 3, a pre-segmentation process is applied to images using FRFCM [32]. The images were converted to a matrix at a dimension of 2500 x 20 in Step 6, respectively in order for ELM to process MRI images. In Step 7, test label images containing ground truth information of 20 MRI images to be used in testing process were prepared. In Step 8, the trained ELM uses test_image and test_label to perform segmentation process. In the last step, segmented_image is resized at 20 x 50 x 50 in order to display it as an image. The testing process of BTS-ELM-FRFCM is summarized in **Fig. 4**. Images obtained by applying **Algorithm 2** to a sample MRI image are shown in **Table 1**.

Journal Pre-proofs

Algorithm 2. The testing process

Step 1: Use 20 MRI images in DICOM format

Step 2: Convert images to Grayscale format

Step 3: Apply pre-segmentation to images using FRFCM algorithm

Step 4: Convert images to binary format

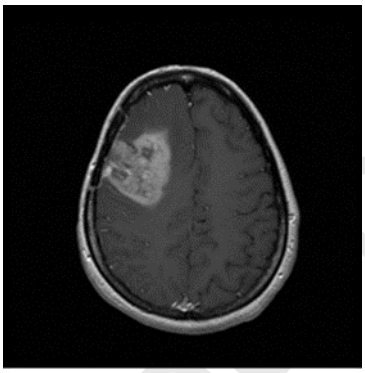
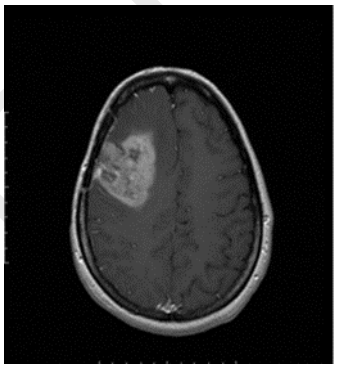
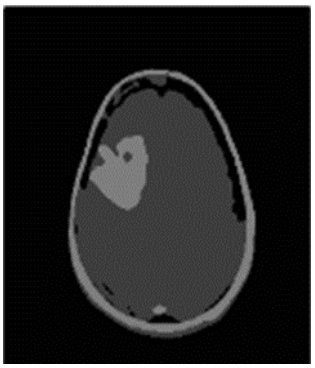
Step 5: resized images [50 50], create 20 x 50 x 50 testing images, and store them in **test_data** variable.

Step 6: Resize **test_data** at 2500 x 20

Step 7: Prepare **test_label** for **test_data** and resize them at 2500 x 20

Step 8: segmented_image ← Perform brain tumor segmentation using trained ELM (**test_image, test_label**)

Table 1. Test Results

DICOM Image (**Step 1**)	Grayscale Image (**Step 2**)	Pre-segmented image via FRFCM Image (**Step 3**)
		
Binary image (**Step 4**)	[50 50] resized image (**Step 5**)	segmented_image (**Step 8**)

Journal Pre-proofs

3. BTS-SVM approach

Introduced by Vladimir N. Vapnik in 1995, SVM is a successful and widely used statistical learning method in many fields [33]. It is divided into two categories as linear and non-linear SVM. Further information on SVM can be found in [33].

In the present study, BTS-SVM was used as an algorithm for comparison among different approaches. One Class SVM was preferred as a type of SVM, and gradient descent was used to minimize energy. The number of maximum iteration was set at 1000.

4. BTS-FCM approach

Fuzzy C-means algorithm is a popular data clustering algorithm developed by [43]. FCM is an uncontrolled learning approach which can successfully perform segmentation on medical images. An objective function is used in this algorithm, and it aims to group iteratively similar preliminary data which are randomly assigned. Group process is completed when it reaches a certain error value. In the image segmentation process, FCM algorithm groups similar pixels in an image, and behaves in accordance with fuzzy logic approach. Further information on the algorithm can be found in [43].

In the present study, BTS-FCM is the second approach used for brain tumor segmentation. The approach is based on a 3-class fuzzy c-means clustering structure. FCM input is set as “cut between the middle and large class”.

5. BTS-SOM approach

Introduced by Kohonen [35], SOM is an artificial neural network learning method. It contains a competitive artificial neural network and operates in two steps [44]. The first step involves creating an artificial neural network through a training set and a map through the competitive learning

Journal Pre-proofs

network. In the second step, the algorithm rapidly predicts the clusters of new data on the map [45]. The structure of SOM network is shown in **Fig. 6**.

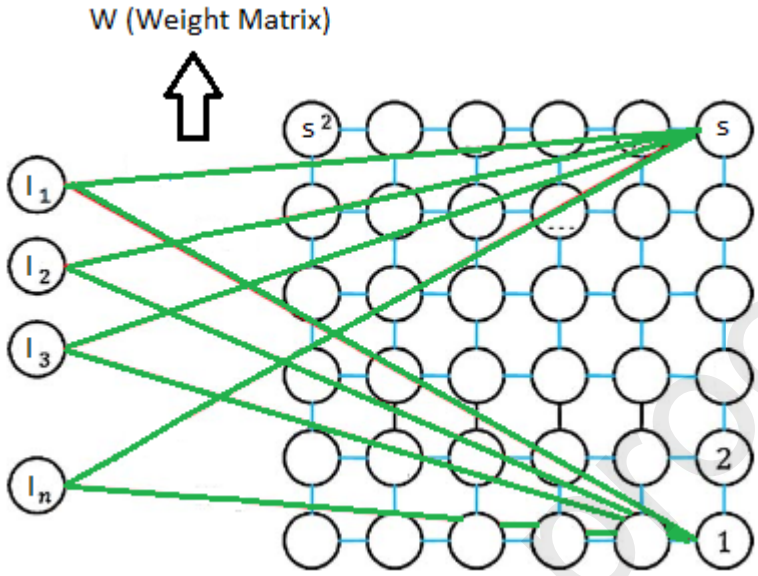


Fig. 6. Architecture of the SOM network

where (I1, I2, ... , Ii, ... , In) are a set of continuous-valued input patterns. I has a fixed dimension n. The SOM network has an array of S=s × s neurons. W is weight matrix. K-means is one of the most widely used unsupervised learning approaches. The approach relies on data for clustering process which aims to put all data in a single cluster. In K-means clustering approach, m number of data is grouped in N number of clusters. This process mainly aims to reach a high level of similarity within the clusters, while it aims to find a low level of similarity among clusters [46,47].

BTS-SOM, which is the third algorithm used for performance comparison in the present study, performs brain tumor segmentation by combining SOM and K-means. Cluster parameter in this approach is used to set the number of segmentation clusters. As a result, the number of segmentation images generated by BTS-SOM is equal to clusters at the end of segmentation process. The number of classes and SOM iterations were set at 4 and 200, respectively.

6. Metric Criteria for Segmentation Performance Comparison

Journal Pre-proofs

Journal Pre-proofs

In the present study, Jaccard Similarity Coefficient (JSC), Dice and Figure of Merid (FOM) were used to compare segmentation performances of BTS-ELM-FRFCM, BTS-SVM, BTS-FCM and BTS-SOM approaches. These metrics are described in detail in the following sections.

6.1. JSC

JSC is often used to evaluate the segmentation performance of an approach. JSC compares two different images in order to measure segmentation performance depending on the following formula [13]:

$$\mathrm{JSC} = \frac{|\mathrm{SI} \cap \mathrm{GT}|}{|\mathrm{SI} \cup \mathrm{GT}|} \tag{8}$$

where SI is segmented_image, and GT represents ground truth. SI corresponds to the segmented image obtained by the segmentation approach. GT corresponds to the complete and accurate segmentation result obtained manually. JSC value varies between 0 and 1, and the segmentation performance increases as the value converges to 1.

6.2. DC

DC is another popular metric used for the measurement of segmentation performance. It measures the segmentation performance using the mathematical formula given in Eq. 9 [48]. DC values varies between 0 and 1, the similarity increases as the value converges to 1.

$$\mathrm{DC} = \frac{2 * |\mathrm{SI} \cap \mathrm{GT}|}{|\mathrm{SI}| + |\mathrm{GT}|} \tag{9}$$

where SI is segmented_image, and GT represents ground truth.

6.3. FOM

FOM, which was developed by Pratt [49] is another metric used for the measurement of image processing performance. It benefits from the following mathematical formula:

Journal Pre-proofs

$$\text{FOM} = \frac{1}{\max(ESI, EGT)} \sum_{i=1}^{ESI} \frac{1}{1 + d_m^2(i)} \quad (10)$$

where ESI (Edge segmented_image) is the number of edge pixels in the segmentation result obtained using edge detection approach, while EGT (Edge Ground Truth) represents the number of edge pixels in ground truth images. m (i) denotes the distance between ith boundary pixel and the nearest ground-truth pixel. d is the scaling constant. FOM values vary between 0 and 1, and the segmentation performance increases as this value converges to 1.

7. Experimental results

As described in Section "2.5.1. Training process" in detail, 3200 training data were used for BTS-ELM-FRFCM training. Following this process, as described in Section "2.5.2. Testing process" in detail, the segmentation of 20 different MRI images was performed using BTS-ELM-FRFCM approach. In addition, BTS-SVM, BTS-FCM, BTS-SOM approaches were also applied to 20 MRI images used in the testing process of the proposed approach in order to compare the segmentation performance of BTS-ELM-FRFCM with different approaches. The findings obtained on RPI device with LCD screen following the segmentation process of the first five MRI images using BTS-ELM-FRFCM are shown in **Fig. 6**. The first five test results obtained from MRI images, ground truth images and three approaches in the first five tests are given in **Table 2**.

Journal Pre-proofs

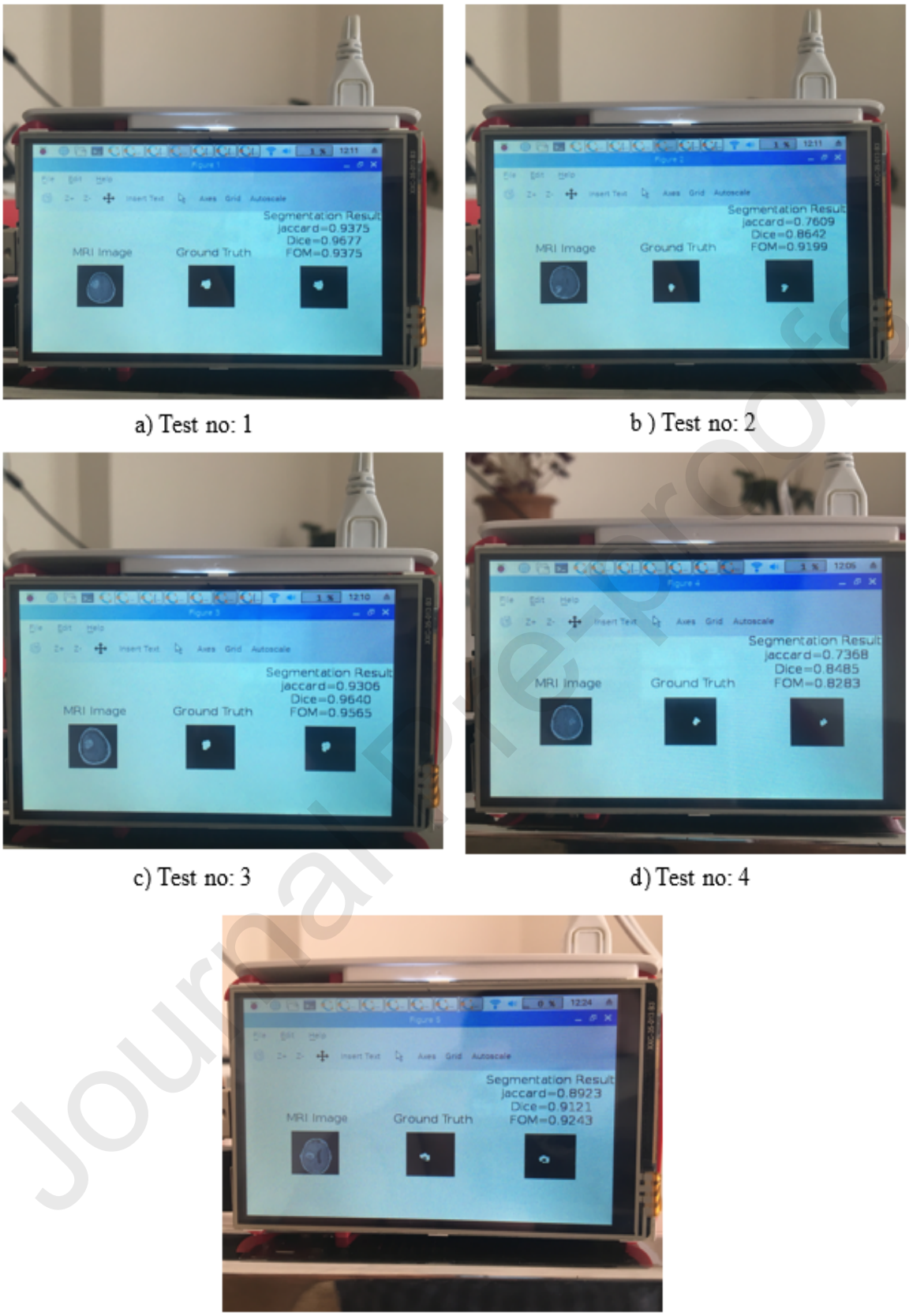
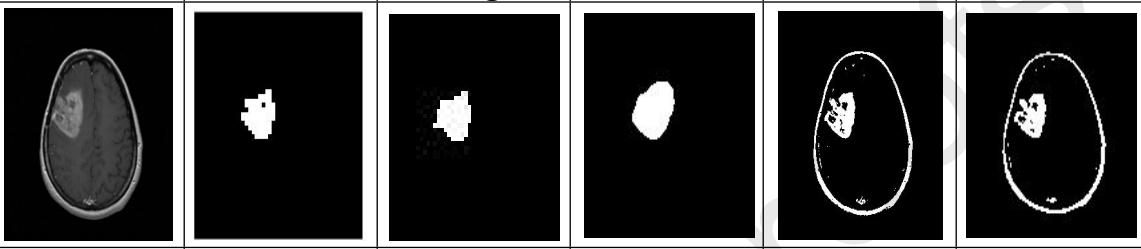
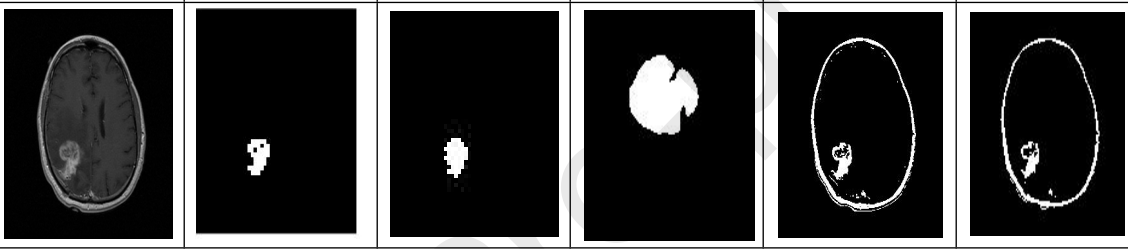
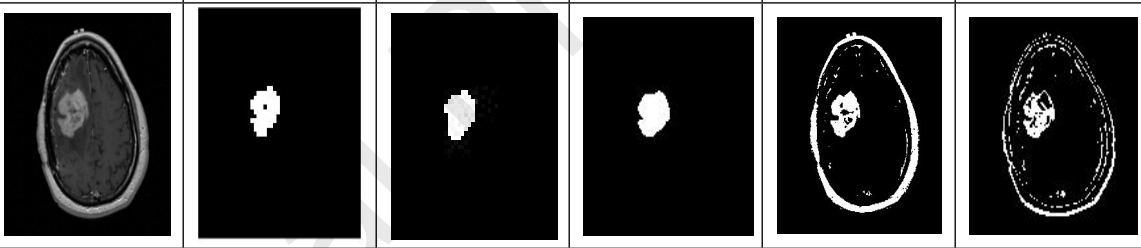
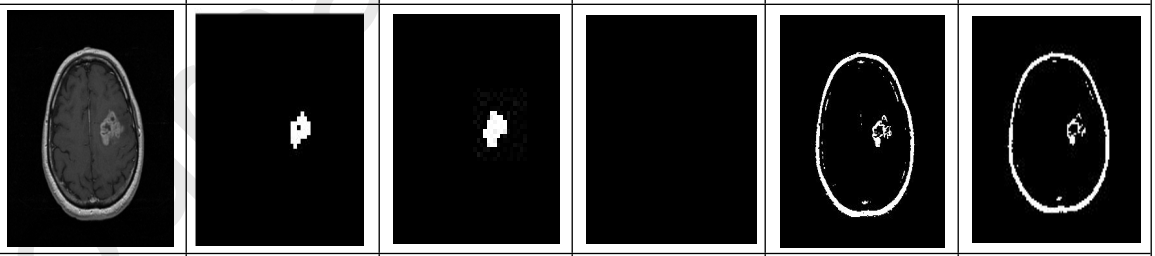
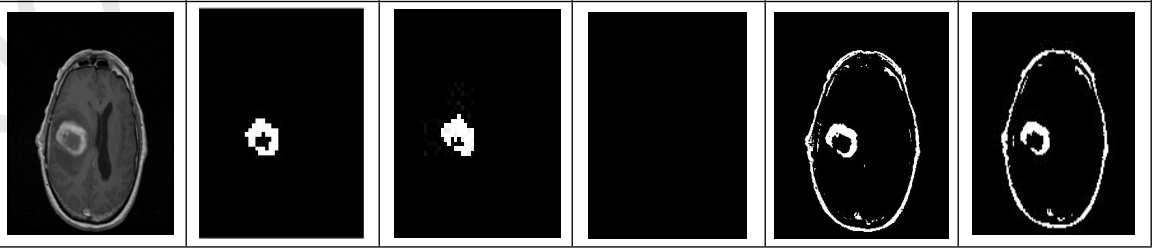


Fig. 7. Test results shown on RPI hardware

Journal Pre-proofs

The results of JSC, FOM and Dice test results obtained from ELM-FRFCM, BTS-SVM, BTS-FCM and BTS-SOM on 20 MRI images are given in **Table 3**.

Table 2. The segmentation images following ELM training

Test	T1-GD MRI Image	Ground truth	BTS-ELM-FRFCM segmented image	BTS-SVM segmented image	BTS-FCM segmented image	BTS-SOM segmented image
1						
2						
3						
4						
5						

Journal Pre-proofs

Table 3. The comparison of the findings with machine learning results

	BTS-ELM-FRFCM			BTS-SVM			BTS-FCM			BTS-SOM		
Test	**FOM**	**Dice**	**JSC**	**FOM**	**Dice**	**JSC**	**FOM**	**Dice**	**JSC**	**FOM**	**Dice**	**JSC**
1	0.9375	0.9677	0.9375	0.5255	0.6890	0.5255	0.4118	0.529	0.3596	0.4016	0.5366	0.3667
2	0.9199	0.8642	0.7609	0	0	0	0.2845	0.3745	0.2845	0.229	0.3438	0.2075
3	0.9565	0.9640	0.9306	0.7311	0.7878	0.6708	0.3101	0.3625	0.2465	0.5446	0.5157	0.3475
4	0.8283	0.8485	0.7368	0	0	0	0.1765	0.2058	0.1421	0.1392	0.1159	0.0615
5	0.9243	0.9121	0.8923	0	0	0	0.289	0.3615	0.2496	0.2632	0.3694	0.2266
6	0.8000	0.8889	0.8000	0	0	0	0.0582	0.0954	0.0501	0.0677	0.1101	0.0583
7	0.9878	0.9891	0.9743	0.6592	0.7946	0.6592	0.1931	0.3096	0.1831	0.1881	0.3069	0.1813
8	0.9399	0.8696	0.8121	0.8203	0.9013	0.8203	0.2566	0.4033	0.2526	0.2604	0.4132	0.2604
9	0.9899	0.8889	0.8000	0.0183	0.036	0.0183	0.0275	0.0535	0.0275	0.0179	0.0351	0.0179
10	0.7959	0.8864	0.7959	0	0	0	0.1335	0.1996	0.1109	0.1508	0.2245	0.1264
11	0.8194	0.9008	0.8194	0.0326	0.0338	0.0172	0.0928	0.0768	0.0399	0.0726	0.0286	0.0145
12	0.7917	0.8837	0.7917	0.0816	0.0962	0.0506	0.1947	0.2177	0.1222	0.1743	0.1911	0.1056
13	0.9811	0.8261	0.7037	0.0123	0	0	0.2734	0.3908	0.2428	0.2921	0.3607	0.22
14	0.9161	0.8800	0.7857	0	0	0	0.0904	0.1288	0.0688	0.1581	0.08	0.0417
15	0.9712	0.9616	0.9511	0	0	0	0.2403	0.3673	0.225	0.2668	0.3974	0.248
16	0.9524	0.9756	0.9524	0.2665	0.4208	0.2665	0.3117	0.4459	0.2869	0.3248	0.4368	0.2794
17	0.9849	0.9444	0.8947	0	0	0	0.1834	0.2685	0.155	0.1932	0.3023	0.1781
18	0.9241	0.8932	0.8612	0	0	0	0.2096	0.3081	0.1821	0.2184	0.34	0.2048
19	0.9686	0.9232	0.8935	0.0167	0	0	0.2417	0.2781	0.1615	0.252	0.2586	0.1485
20	0.8328	0.8421	0.7273	0.0102	0	0	0.1581	0.2432	0.1384	0.1681	0.2581	0.1481

.

1) Various statistical analyses were performed in order to evaluate segmentation results given in Table 3. In this respect, ANOVA analysis was performed on JSC, Dice and FOM results obtained using 4 different segmentation approaches, and the findings are presented in Table 4. It can be observed in Table 4 that mean, minimum and maximum and 95% confidence interval for mean values of JSC, Dice and FOM results obtained using BTS-ELM-FRFCM approach were higher compared to other approaches, and that its standard deviation values were lower compared to other approaches. Therefore, it can be stated that the proposed method displayed a better segmentation performance among all methods.

Journal Pre-proofs

Table 4. Statistics with ANOVA

		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
						Lower Bound	Upper Bound		
FOM	BTS-ELM-FRFCM	20	0.9111	0.0712	0.0159	0.8778	0.9444	0.7917	0.9899
	BTS-SVM	20	0.1587	0.2804	0.0627	0.0275	0.2899	0.0000	0.8203
	BTS-FCM	20	0.2068	0.0963	0.0215	0.1618	0.2519	0.0275	0.4118
	BTS-SOM	20	0.2191	0.1189	0.0266	0.1635	0.2748	0.0179	0.5446
	Total	80	0.3740	0.3517	0.0393	0.2957	0.4522	0.0000	0.9899
Dice	BTS-ELM-FRFCM	20	0.9055	0.0476	0.0106	0.8832	0.9278	0.8261	0.9891
	BTS-SVM	20	0.1880	0.3262	0.0729	0.0353	0.3406	0.0000	0.9013
	BTS-FCM	20	0.2810	0.1289	0.0288	0.2207	0.3413	0.0535	0.5290
	BTS-SOM	20	0.2812	0.1507	0.0337	0.2107	0.3518	0.0286	0.5366
	Total	80	0.4139	0.3444	0.0385	0.3373	0.4906	0.0000	0.9891
JSC	BTS-ELM-FRFCM	20	0.8411	0.0826	0.0185	0.8024	0.8797	0.7037	0.9743
	BTS-SVM	20	0.1514	0.2762	0.0618	0.0222	0.2807	0.0000	0.8203
	BTS-FCM	20	0.1765	0.0915	0.0205	0.1336	0.2193	0.0275	0.3596
	BTS-SOM	20	0.1721	0.1026	0.0229	0.1241	0.2201	0.0145	0.3667
	Total	80	0.3353	0.3331	0.0372	0.2611	0.4094	0.0000	0.9743

2) In the second step, ANOVA significance analysis was performed in order to reveal any differences among FOM, Dice and JSC results obtained using 4 different segmentation approaches given in Table 3, and the findings are given in Table 5, which demonstrates that FOM, Dice and JSC were calculated as ($F(3, 76)= 96.291$; $p=0.000<0.05$), ($F(3, 76)= 59.1070$; $p=0.000<0.05$) and ($F(3, 76)= 89.2740$; $p=0.000<0.05$), respectively. It can be thus stated that FOM, Dice and JSC results obtained using 4 different segmentation approaches differed significantly. In order to prove this difference numerically, Scheffe analysis was performed.

Journal Pre-proofs

Journal Pre-proofs

Table 5. Significance report of three metrics with ANOVA.

ANOVA						
		Sum of Squares	**df**	**Mean Square**	**F**	**Sig.**
FOM	Between Groups	7.7350	3	2.578	96.291	0.000
	Within Groups	2.0350	76	0.027		
	Total	9.7700	79			
Dice	Between Groups	6.5600	3	2.1870	59.1070	0.000
	Within Groups	2.8110	76	0.0370		
	Total	9.3710	79			
JSC	Between Groups	6.8290	3	2.2760	89.2740	0.000
	Within Groups	1.9380	76	0.0250		
	Total	8.7670	79			

3) In the third step, Scheffe analysis was performed on FOM, Dice and JSC results as given in Table 6-8. It can be understood from Table 6 that the highest Mean Difference (I-J) results were calculated in the first three lines of Table 6 where Scheffe analysis on FOM, Dice and JSC results are given. These findings indicate that arithmetic means of FOM results obtained using BTS-ELM-FRFCM approach were higher compared to other approaches. The lowest Mean Difference (I-J) results were obtained between Line 4-6 in Table 6. Therefore, mean values of FOM results obtained using BTS-SVM approach were lower compared to those of other approaches.

Journal Pre-proofs

Table 6. Scheffe comparison of FOM.

Multiple Comparisons							
Scheffe							
Dependent Variable			**Mean Difference (I-J)**	**Std. Error**	**Sig.**	**95% Confidence Interval**	
						Lower Bound	**Upper Bound**
FOM	BTS-ELM-FRFCM	BTS-SVM	0.7524*	0.0517	0.0000	0.6044	0.9004
		BTS-FCM	0.7042*	0.0517	0.0000	0.5563	0.8522
		BTS-SOM	0.6919*	0.0517	0.0000	0.5440	0.8399
	BTS-SVM	BTS-ELM-FRFCM	-0.7524*	0.0517	0.0000	-0.9004	-0.6044
		BTS-FCM	-0.0481	0.0517	0.8337	-0.1961	0.0998
		BTS-SOM	-0.0604	0.0517	0.7148	-0.2084	0.0875
	BTS-FCM	BTS-ELM-FRFCM	-0.7042*	0.0517	0.0000	-0.8522	-0.5563
		BTS-SVM	0.0481	0.0517	0.8337	-0.0998	0.1961
		BTS-SOM	-0.0123	0.0517	0.9965	-0.1603	0.1357
	BTS-SOM	BTS-ELM-FRFCM	-0.6919*	0.0517	0.0000	-0.8399	-0.5440
		BTS-SVM	0.0604	0.0517	0.7148	-0.0875	0.2084
		BTS-FCM	0.0123	0.0517	0.9965	-0.1357	0.1603

*The mean difference is significant at the 0.05 level.

4) Scheffe analysis was performed on Dice results as given in Table 7, and the highest Mean Difference (I-J) results were obtained in the first three lines of Table 7. These findings demonstrate that arithmetic means of Dice results obtained using BTS-ELM-FRFCM approach were higher compared to those of other approaches. It is also observed that the lowest Mean Difference (I-J) results were obtained between Line 4-6 in Table 7. Therefore, it is evident that arithmetic means of Dice results obtained using BTS-SVM approach were lower compared to those of other approaches.

Journal Pre-proofs

Table 7. Scheffe comparison of Dice.

Multiple Comparisons							
Scheffe							
Dependent Variable			**Mean Difference (I-J)**	**Std. Error**	**Sig.**	**95% Confidence Interval**	
						Lower Bound	**Upper Bound**
Dice	BTS-ELM-FRFCM	BTS-SVM	0.7175*	0.0608	0.0000	0.5436	0.8914
		BTS-FCM	0.6245*	0.0608	0.0000	0.4506	0.7984
		BTS-SOM	0.6242*	0.0608	0.0000	0.4504	0.7982
	BTS-SVM	BTS-ELM-FRFCM	-0.7175*	0.0608	0.0000	-0.8914	-0.5436
		BTS-FCM	-0.0930	0.0608	0.5089	-0.2669	0.0809
		BTS-SOM	-0.0933	0.0608	0.5066	-0.2672	0.0806
	BTS-FCM	BTS-ELM-FRFCM	-0.6245*	0.0608	0.0000	-0.7984	-0.4506
		BTS-SVM	0.0930	0.0608	0.5089	-0.0809	0.2669
		BTS-SOM	- 0.0002	0.0608	1.0000	-0.1741	0.1737
	BTS-SOM	BTS-ELM-FRFCM	- 0.6242*	0.0608	0.0000	-0.7982	-0.4504
		BTS-SVM	0.0933	0.0608	0.5066	-0.0806	0.2672
		BTS-FCM	0.0002	0.0608	1.0000	-0.1737	0.1741

* The mean difference is significant at the 0.05 level.

5) Scheffe analysis was performed on JSC results as given in Table 8, and the highest Mean Difference (I-J) results were obtained in the first three lines of Table 8. These findings demonstrate that arithmetic means of JSC results obtained using BTS-ELM-FRFCM approach were higher compared to those of other approaches. It is also observed that the lowest Mean Difference (I-J) results were obtained between Line 4-6 in Table 8. Therefore, it can be concluded that arithmetic means of JSC results obtained using BTS-SVM approach were lower compared to those of other approaches.

Journal Pre-proofs

Journal Pre-proofs

In conclusion, it was proven in Table 4-8 that FOM, Dice and JSC results obtained using the proposed BTS-ELM-FRFCM approach were statistically higher compared to other edge detection approaches. Thus, it was clearly indicated that the proposed method display a better segmentation performance compared to other approaches. In addition, it was also statistically indicated that BTS-SVM approach displayed a worse performance compared to other approaches. In this respect, it is evident that the proposed ELM-FRFCM approach functioned properly and successfully on RPI hardware.

Table 8. Scheffe comparison of JSC.

Multiple Comparisons							
Scheffe							
Dependent Variable			**Mean Difference (I-J)**	**Std. Error**	**Sig.**	**95% Confidence Interval**	
						Lower Bound	**Upper Bound**
JSC	BTS-ELM-FRFCM	BTS-SVM	0.6896*	0.0505	0.0000	0.5453	0.8340
		BTS-FCM	0.6646*	0.0505	0.0000	0.5202	0.8090
		BTS-SOM	0.6689*	0.0505	0.0000	0.5245	0.8133
	BTS-SVM	BTS-ELM-FRFCM	-0.6896*	0.0505	0.0000	-0.8340	-0.5453
		BTS-FCM	-0.0250	0.0505	0.9697	-0.1694	0.1193
		BTS-SOM	-0.0207	0.0505	0.9824	-0.1651	0.1237
	BTS-FCM	BTS-ELM-FRFCM	-0.6646*	0.0505	0.0000	-0.8090	-0.5202
		BTS-SVM	0.0250	0.0505	0.9697	-0.1193	0.1694
		BTS-SOM	0.0043	0.0505	0.9998	-0.1401	0.1487
	BTS-SOM	BTS-ELM-FRFCM	-0.6689*	0.0505	0.0000	-0.8133	-0.5245
		BTS-SVM	0.0207	0.0505	0.9824	-0.1237	0.1651
		BTS-FCM	-0.0043	0.0505	0.9998	-0.1487	0.1401

* The mean difference is significant at the 0.05 level.

The comparative graphs of JSC, FOM and Dice tests obtained using BTS-ELM-FRFCM, BTS-SVM, BTS-FCM and BTS-SOM in Table 2 are shown in **Fig. 7-9**. It can be understood from these

figures that the proposed BTS-ELM-FRFCM approach displayed a significantly higher segmentation performance compared to other methods.

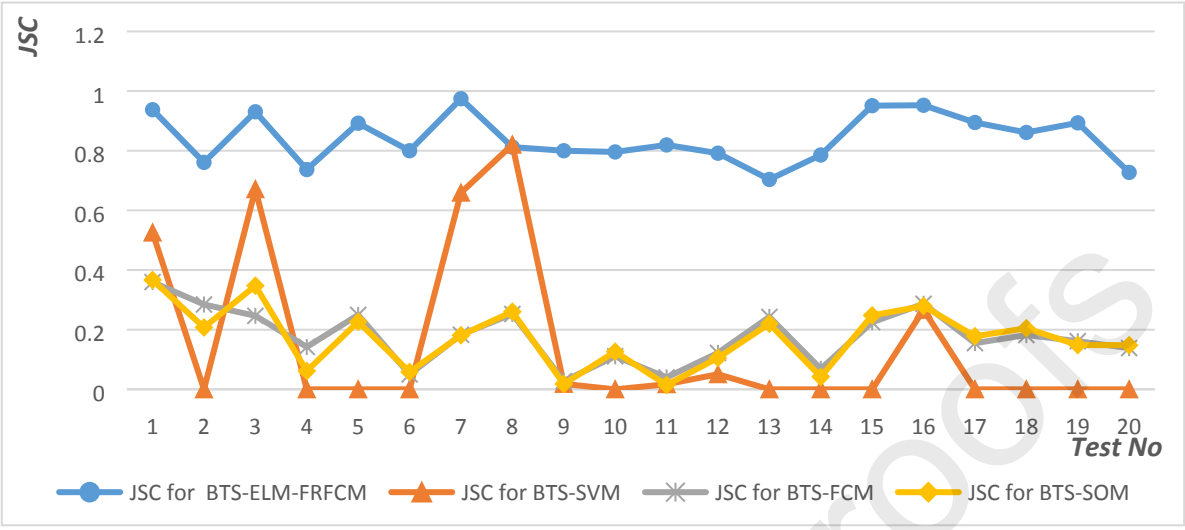


Fig. 8. JSC graph for three methods

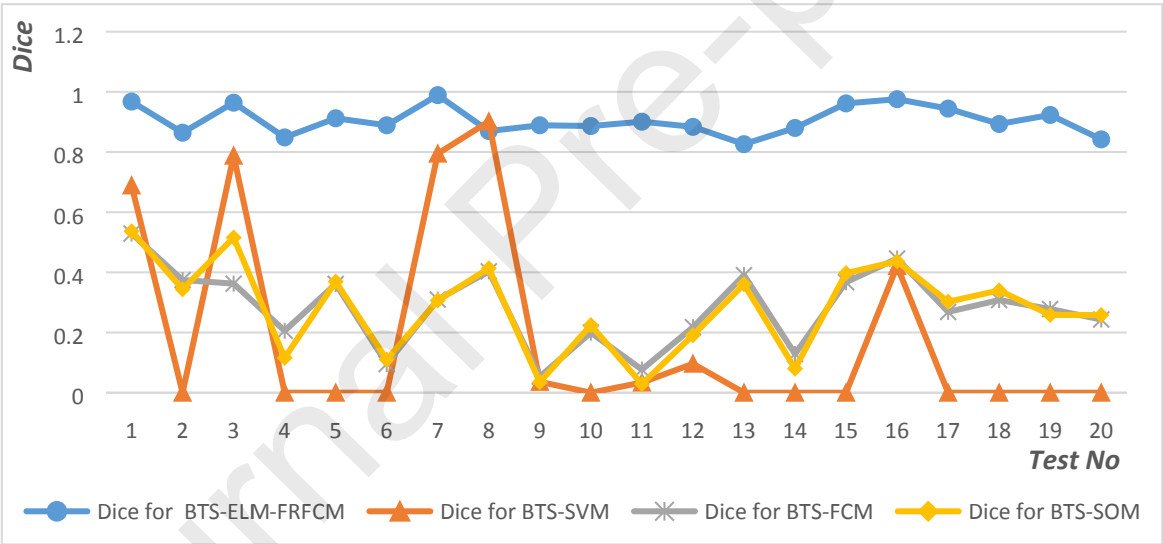


Fig. 9. Dice graph for three methods

Journal Pre-proofs

Journal Pre-proofs

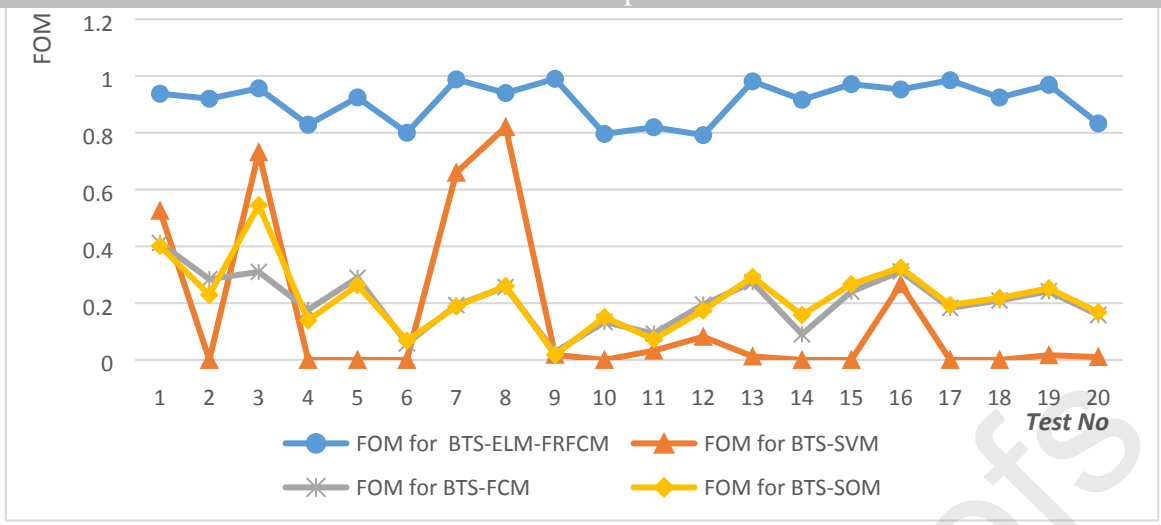


Fig. 10. FOM graph for three methods

8. Conclusion

In the present study, a brain tumor segmentation system was designed using ELM and FRFCM algorithms on RPI hardware. The greatest advantages of RPI devices are their effectiveness and low cost, while they also display high hardware features. Therefore, ELM and FRFCM, which are successful and quick algorithms used in segmentation processes, were used in the present study. It was observed in the experimental studies that BTS-ELM-FRFCM approach proposed for RPI hardware performed a quick segmentation process. The proposed BTS-ELM-FRFCM approach was compared with BTS-SVM, BTS-FCM and BTS-SOM using 20 different MRI images in order to measure the segmentation performances of these four methods. As a result, the mean values of JSC, Dice and FOM analysis on the segmentation results of the proposed method were calculated as 0.8411, 0.9055 and 0.9111, respectively. In the light of statistical analyses such as ANOVA and Scheffe performed on the results of JSC, Dice and FOM results, it can be concluded that the proposed method statistically displayed a better segmentation performance compared to BTS-SVM, BTS-FCM and BTS-SOM. Therefore, a segmentation system which operates on two different algorithms was offered for the health sector.

Journal Pre-proofs

Journal Pre-proofs

REFERENCES

[1] Sert E. A new modified neutrosophic set segmentation approach. Computers & Electrical Engineering 2018;65:576–92. doi:10.1016/j.compeleceng.2017.01.017.

[2] Sert E, Alkan A. Image edge detection based on neutrosophic set approach combined with Chan–Vese algorithm. International Journal of Pattern Recognition and Artificial Intelligence 2019;33(03):1954008. doi:10.1142/S0218001419540089.

[3] Zotin A, Simonov K, Kurako M, Hamad Y, Kirillova S. Edge detection in MRI brain tumor images based on fuzzy c-means clustering. Procedia Computer Science 2018;126:1261–70. doi:10.1016/j.procs.2018.08.069.

[4] Yilmaz B, Durdu A, Emlik GD. A new method for skull stripping in brain MRI using multistable cellular neural networks. Neural Computing and Applications 2018;29(8):79–95. doi:10.1007/s00521-016-2834-2.

[5] Chandra GR, Rao KRH. Tumor detection in brain using genetic algorithm. Procedia Computer Science 2016;79:449–57. doi:10.1016/j.procs.2016.03.058.

[6] Mittal M, Goyal LM, Kaur S, Kaur I, Verma A, Jude Hemanth D. Deep learning based enhanced tumor segmentation approach for MR brain images. Applied Soft Computing Journal 2019;78:346–54. doi:10.1016/j.asoc.2019.02.036.

[7] Havaei M, Davy A, Warde-Farley D, Biard A, Courville A, Bengio Y, et al. Brain tumor segmentation with deep neural networks. Medical Image Analysis 2017;35:18–31. doi:10.1016/j.media.2016.05.004.

[8] Rehman ZU, Naqvi SS, Khan TM, Khan MA, Bashir T. Fully automated multi-parametric brain tumour segmentation using superpixel based classification. Expert Systems with Applications 2019;118:598–613. doi:10.1016/j.eswa.2018.10.040.

[9] Selvapandian A, Manivannan K. Fusion based glioma brain tumor detection and segmentation using ANFIS classification. Computer Methods and Programs in Biomedicine 2018;166:33–8. doi:10.1016/j.cmpb.2018.09.006.

Journal Pre-proofs

[10] Bergner N, Bocklitz T, Romeike BFM, Reichart R, Kalff R, Krafft C, et al. Identification of primary tumors of brain metastases by raman imaging and support vector machines. Chemometrics and Intelligent Laboratory Systems 2012;117:224–32. doi:10.1016/j.chemolab.2012.02.008.

[11] Şişik F, Sert E. Support vector machine working on FPGA and the segmentation method of brain MR screening. International Journal of Innovative Research in Education 2017;4(3):120. doi:10.18844/ijire.v4i3.2549.

[12] Pinto A, Pereira S, Rasteiro D, Silva CA. Hierarchical brain tumour segmentation using extremely randomized trees. Pattern Recognition 2018;82:105–17. doi:10.1016/j.patcog.2018.05.006.

[13] Sert E, Avci D. Brain tumor segmentation using neutrosophic expert maximum fuzzy-sure entropy and other approaches. Biomedical Signal Processing and Control 2019;47:276–87. doi:10.1016/j.bspc.2018.08.025.

[14] Özyurt F., Sert E., Avci E, Dogantekin, E. Brain tumor detection based on Convolutional Neural Network with neutrosophic expert maximum fuzzy sure entropy. Measurement 2019;147: 106830. doi:10.1016/j.measurement.2019.07.058

[15] Sert E., Özyurt F, Doğantekin A. A new approach for brain tumor diagnosis system: single image super resolution based maximum fuzzy entropy segmentation and convolutional neural network. Medical hypotheses 2019;133:109413. doi:10.1016/j.mehy.2019.109413

[16] Özyurt F, Sert E, Avcı D. An expert system for brain tumor detection: fuzzy c-means with super resolution and convolutional neural network with extreme learning machine. Medical Hypotheses 2020; 134: 109433 doi:10.1016/j.mehy.2019.109433

[17] Zhu C, Zou B, Zhao R, Cui J, Duan X, Chen Z, et al. Retinal vessel segmentation in colour fundus images using extreme learning machine. Computerized Medical Imaging and Graphics 2017;55:68–77. doi:10.1016/j.compmedimag.2016.05.004.

[18] Zeng YZ, Zhao YQ, Liao M, Zou BJ, Wang XF, Wang W. Liver vessel segmentation based on extreme learning machine. Physica Medica 2016;32(5):709–16. doi:10.1016/j.ejmp.2016.04.003.

Journal Pre-proofs

Journal Pre-proofs

[19] Huang H, Ma H, JW van Triest H, Wei Y, Qian W. Automatic detection of neovascularization in retinal images using extreme learning machine. Neurocomputing 2018;277:218–27. doi:10.1016/j.neucom.2017.03.093.

[20] Xie W, Li Y, Ma Y. Breast mass classification in digital mammography based on extreme learning machine. Neurocomputing 2016;173:930–41. doi:10.1016/j.neucom.2015.08.048.

[21] Sai P-K, Wang J-G, Teoh E-K. Facial age range estimation with extreme learning machines. Neurocomputing 2015;149:364–72. doi:10.1016/j.neucom.2014.03.074.

[22] Nguyen TV, Mirza B. Dual-layer kernel extreme learning machine for action recognition. Neurocomputing 2017;260:123–30. doi:10.1016/j.neucom.2017.04.007.

[23] Zhao L, Wang X. Seemingly unrelated extreme learning machine. Neurocomputing 2019;355:134–42. doi:10.1016/j.neucom.2019.04.067.

[24] Shang Z, He J. Confidence-weighted extreme learning machine for regression problems. Neurocomputing 2015;148:544–50. doi:10.1016/j.neucom.2014.07.009.

[25] Anandhalli M, Baligar VP. A novel approach in real-time vehicle detection and tracking using raspberry pi. Alexandria Engineering Journal 2018;57(03):1597–607. doi:10.1016/j.aej.2017.06.008.

[26] Sajjad M, Nasir M, Ullah FUM, Muhammad K, Sangaiah AK, Baik SW. Raspberry pi assisted facial expression recognition framework for smart security in law-enforcement services. Information Sciences 2019;479:416–31. doi:10.1016/j.ins.2018.07.027.

[27] Diallo AD, Gobee S, Durairajah V. Autonomous tour guide robot using embedded system control. Procedia Computer Science 2015;76:126–33. doi:10.1016/j.procs.2015.12.302.

[28] Lam Loong Man CKY, Koonjul Y, Nagowah L. A low cost autonomous unmanned ground vehicle. Future Computing and Informatics Journal 2018;3(2):304–20. doi:10.1016/j.fcij.2018.10.001.

[29] Yildiz MZ, Boyraz ÖF. Development of a low-cost microcomputer based vein imaging system. Infrared Physics & Technology 2019;98:27–35. doi:10.1016/j.infrared.2019.02.010.

Journal Pre-proofs

[30] Wang C, Ko M-C, Chen Y-M, Chen L-Q, Lin C-W. An automatic multi-thread image segmentation embedded system for surface plasmon resonance sensor. Sensors and Actuators A: Physical 2019;285:603–12. doi:10.1016/j.sna.2018.12.007.

[31] Tu GJ, Hansen MK, Kryger P, Ahrendt P. Automatic behaviour analysis system for honeybees using computer vision. Computers and Electronics in Agriculture 2016;122:10–8. doi:10.1016/j.compag.2016.01.011.

[32] Lei T, Jia X, Zhang Y, He L, Meng H, Nandi AK. Significantly fast and robust fuzzy c-means clustering algorithm based on morphological reconstruction and membership filtering. IEEE Trans Fuzzy Syst 2018;26(05):3027–41. doi:10.1109/TFUZZ.2018.2796074.

[33] Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995;20(03):273–97. doi:10.1007/BF00994018.

[34] Bezdek JC, Ehrlich R, Full W. FCM: the fuzzy c-means clustering algorithm. Computers & Geosciences 1984;10(02-03):191–203. doi:10.1016/0098-3004(84)90020-7.

[35] Kohonen T, Huang TS, Schroeder MR. Self-organizing maps. Berlin, Heidelberg: Springer Berlin / Heidelberg; 2012.

[36] Scarpace L, Mikkelsen T, Cha S, Rao S, Tekchandani S, Gutman D, et al. Radiology data from the cancer genome atlas glioblastoma multiforme [TCGA-GBM] collection 2016. doi:10.7937/k9/tcia.2016.rnyfuye9.

[37] Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The cancer ımaging archive (TCIA): maintaining and operating a public ınformation repository. Journal of Digital Imaging 2013;26(06):1045–57. doi:10.1007/s10278-013-9622-7.

[38] Scarpace L, Flanders AE, Jain R, Mikkelsen T, Andrews DW. Data from rembrandt 2015. doi:10.7937/k9/tcia.2015.588ozuzb.

[39] Huang G-B, Zhu Q-Y, Siew C-K. Extreme learning machine: theory and applications. Neurocomputing 2006;70(01-03):489–501. doi:10.1016/j.neucom.2005.12.126.

Journal Pre-proofs

[40] McDonnell MD, Tissera MD, Vladusich T, van Schaik A, Tapson J. Fast, simple and accurate handwritten digit classification by training shallow neural network classifiers with the 'extreme learning machine' algorithm. Plos One 2015;10(08):e0134254. doi:10.1371/journal.pone.0134254.

[41] Tang J, Deng C, Huang G-B, Zhao B. Compressed-domain ship detection on spaceborne optical image using deep neural network and extreme learning machine. IEEE Trans Geosci Remote Sensing 2015;53(03):1174–85. doi:10.1109/TGRS.2014.2335751.

[42] Golestaneh P, Zekri M, Sheikholeslam F. Fuzzy wavelet extreme learning machine. Fuzzy Sets and Systems 2018;342:90–108. doi:10.1016/j.fss.2017.12.006.

[43] Bezdek JC. Pattern Recognition with fuzzy objective function algorithms. Boston, MA: Springer US; 1981.

[44] Haykin SS. Neural networks: a comprehensive foundation. 2nd ed. Upper Saddle River, N.J: Prentice Hall; 1999.

[45] Chi D. Self-Organizing Map-based color image segmentation with k -means clustering and saliency map. ISRN Signal Processing 2011;2011:1–18. doi:10.5402/2011/393891.

[46] Yao H, Duan Q, Li D, Wang J. An improved k -means clustering algorithm for fish image segmentation. Mathematical and Computer Modelling 2013;58(03-04):790–8. doi:10.1016/j.mcm.2012.12.025.

[47] Sert E, Okumus IT. Segmentation of mushroom and cap width measurement using modified k-means clustering algorithm. AEEE 2014;12(04):354–60. doi:10.15598/aeee.v12i4.1200.

[48] Joskowicz L, Cohen D, Caplan N, Sosna J. Automatic segmentation variability estimation with segmentation priors. Medical Image Analysis 2018;50:54–64. doi:10.1016/j.media.2018.08.006.

[49] Pratt WK. Digital image processing. Hoboken, N.J.: Wiley-Liss; 2002.

Journal Pre-proofs

Journal Pre-proofs

- In this study, brain tumor segmentation was performed using ELM and FRFCM approaches.
- BTS-ELM-FRFCM approach was run on Raspberry PI hardware.
- ELM and FRFCM are new algorithms with high performance.
- A system to be used as mobile in hospitals is designed.
- Experimental studies reveal the superiority of the proposed method.

Journal Pre-proofs

Pengembangan Rekayasa dan Teknologi, Vol 15, No. 1, Juni 2019, pp 38-42
p-ISSN: 1410-9840 & e-ISSN: 2580-8850
http://journals.usm.ac.id/index.php/jprt/index

KOMPARASI METODE CLUSTERING K-MEANS DAN FUZZY C-MEANS UNTUK MEMPREDEKSI KETEPATAN WAKTU LULUS

Galet Guntoro Setiaji [1], Khoirudin [2], Vensy Vydia [3]
[1,2]Program Studi Teknik Informatika Universitas Semarang
[3] Program Studi Sistem Informasi Universitas Semarang
[1] galet@usm.ac.id, [2] khoirudin@usm.ac.id, [3] vensy@usm.ac.id

ABSTRAK

Metode data mining merupakan metode yang banyak digunakan oleh peneliti untuk mencari sebuah informasi dari sebuah kumpulan data-data, penelitian ini menggunakan object data lulusan mahasiswa Universitas Semarang. Data yang kita ambil yaitu dari jurusan Teknik Elektro dan Teknik Sipil.

Dengan menggunakan metode data mining terutama algoritma K-Means dan Fuzzy C-Means dapat mengidentifikasi ketepatan lulusan di jurusan Teknik Elektro dan Teknik Sipil. Dimana ada tiga cluster yaitu tepat waktu, tidak tepat waktu dan over studi.

Dari hasil penelitian dihasilkan validitas *K-Means* lebih kecil dan lebih baik, dengan validitas DBI sebesar 2.022, sedangkan FCM dengan validitas XBI didapatkan validitas sebesar 11.948. Dengan kesimpulan lulusan lama studi 4 tahun lebih 2 bulan dengan IPK rata-rata 3.13, sedangkan lama studi 4 tahun dengan ipk 3.42, dan untuk mahasiswa dengan lama studi 5 tahun lebih rata-rata ipk 2.88.

Keywords : Clustering, K-Means, Fuzzy C-Means, Prediksi

1. Pendahuluan

Sebuah data yang bisa diolah dengan sebuah metode algoritma tertentu dan bisa menghasilkan sebuah informasi yang dapat digunakan untuk menarik sebuah keputusan. Begitu juga dengan adanya perkiraan informasi mahasiswa yang tepat kelulusannya. Kita mengambil data dari mahasiswa Universitas Semarang, yaitu Fakultas Teknik terutama prodi Teknik Sipil dan Teknik Elektro.

Dari data tersebut, yaitu data IPK, lama studi dan Program Studi yang nantinya digunakan untuk mengetahui pengelompokan berdasarkan kelulusan mahasiswa dari dua program studi. Kluster merupakan metode dengan mengelompokan data-data kedalam klaster sehingga didalam tiap-tiap kluster berisi data yang sama atau memiliki kemiripan data. Pengelompokan kluster ini dengan cara menghitung jarak tiap data dengan mengunakan perhitungan jarak euclidiant distance (jarak terdekat).

Algoritma K-means merupakan salah satu metode yang pengelompokan data yang nantinya bisa menghasilkan klaster yang lebih stabil, dan algoritma K-Means juga merupakan algoritma yang sering digunakan untuk melakukan pengelompokan. Algoritma K-Means juga mudah melakukan modifikasi disetiap tahap seperti fungsi perhitungan jarak.

2. Metode

a. Data Mining

Data mining adalah suatu proses ekstraksi atau penggalian data yang belum diketahui sebelumnya, namun dapat dipahami dan berguna dari database yang besar serta digunakan untuk membuat suatu keputusan bisnis yang sangat penting (Connolly dan Begg, 2010). Data mining biasa juga disebut dengan “Data atau knowledge discovery” atau menemukan pola

tersembunyi pada data. Data mining adalah proses dari menganalisa data dari prespektif yang berbeda dan menyimpulkannya ke dalam informasi yang berguna.

3. Pembahasan

Data yang dikumpulkan berupa data lulusan wisudawan dengan tahun angkatan masuk mahasiswa tahun 2010 dan nilai IPK mahasiswa program studi Teknik Sipil dan Teknik Elektro, didapatkan tabel seperti dibawah ini :

NIM	NAMA	IPK	WISUDA	SMTLULUS
C.111.11.0002	ARBI SETYOKO	2.89	52	GENAP TH 2015/2016
C.111.11.0008	ANDY ROSADI	3.10	50	GENAP TH 2014/2015
C.111.11.0015	RIMBA AKASIA AJIDARMA	3.13	50	GENAP TH 2014/2015
C.111.11.0016	FAUZI ARIE HARTONO	2.92	52	GENAP TH 2015/2016
C.111.11.0017	NUR WERO WIRAWAN	3.15	50	GENAP TH 2014/2015
C.111.11.0018	RAKHMAD SUBARKAH	3.06	50	GENAP TH 2014/2015
C.111.11.0019	BUDI IRAWAN	3.08	50	GENAP TH 2014/2015
C.111.11.0020	ANDI AULIA DAVID	3.31	50	GENAP TH 2014/2015
C.111.11.0023	MUHAMAD SAMSUL HADI	2.69	52	GENAP TH 2015/2016
C.111.11.0026	SHOLIKUN NUR	3.60	50	GENAP TH 2014/2015
C.111.11.0031	FAIZAL MAHMUD	3.26	50	GENAP TH 2014/2015
C.111.11.0032	AHMAD CHOLIL	3.17	50	GENAP TH 2014/2015
C.111.11.0033	MUCHAMMAD MAULANA RIZKA T	3.03	52	GENAP TH 2015/2016
C.111.11.0034	CAHYANI FAJAR SUKOCO	3.54	52	GENAP TH 2015/2016

Tabel 1 Data mahasiswa

3.1 Persiapan Data

Data yang digunakan untuk melakukan penelitian komparasi ini berupa data nilai dari mahasiswa Universitas Semarang yang diambil sampelnya dari program studi Teknik Sipil dan Teknik Elektro. Didapatkan data sejumlah 840 mahasiswa yang sudah lulus dengan atribut berupa, nim, nilai A - E, Indek Prestasi Kumulatif (IPK) dan Lama studi.

NIM	IPK	hr	bln	thn	A	B	C	D	E	NIM	IPK	hr	bln	thn	A	B	C	D	E
C.431.09.0004	3.1	0	0	7	18	37	10	2	10	C.431.12.0100	3.56	1	6	4	38	23	4	2	1
C.431.12.0049	3.17	30	11	3	22	37	14	5	0	C.431.12.0080	3.23	1	6	4	22	35	9	1	0
C.431.10.0055	2.77	0	0	6	12	31	29	18	9	C.431.12.0055	3.37	30	11	3	28	32	9	0	3
C.431.12.0010	3.28	30	11	3	28	28	10	1	2	C.411.12.0021	3.22	30	11	3	28	27	10	1	2
C.441.11.0041	3.17	2	6	5	26	31	11	1	8	C.411.12.0019	3.15	30	11	3	21	34	10	0	1
C.431.08.0043	2.97	1	6	8	15	38	14	3	7	C.411.12.0032	3.14	30	11	3	25	28	12	2	0
C.431.11.0080	3.01	2	6	5	19	31	16	2	3	C.431.12.0063	3.46	1	6	4	35	26	4	1	3
C.431.11.0039	2.8	0	0	5	13	33	24	5	11	C.411.12.0002	3.32	30	11	3	31	24	9	2	0
C.431.11.0068	3.01	2	6	5	14	38	18	3	3	C.431.11.0064	3.21	2	6	5	23	33	11	0	3
C.431.12.0071	3.24	1	6	4	28	27	10	3	0	C.411.12.0003	3.1	30	11	3	22	29	13	2	0

Table 2 Data Sample Nilai mahasiswa

Normalisasi data dilakukan dengan tujuan untuk mempermudah perhitungan data-data yang beda. Salah satunya dengan cara dengan melakukan normalisasi jumlah huruf A = 18, kita jadikan presentase dengan mengalikan 10% sehingga menjadi 1,80. Sedangkan untuk tahun kita normalisasi dengan membulatkan bulan, apabila bulan diatas 6 maka ditambahkan 1 tahun. Maka data akan menjadi seperti table berikut ini.

NIM	IPK	Hr	Bln	Thn	N=thn	A	B	C	D	E	N=a	N=b	N=c	N=d	N=e
C.431.09.0004	3.1	0	0	7	7	18	37	10	2	10	1.8	3.7	1	0.2	1
C.431.12.0049	3.17	30	11	3	4	22	37	14	5	0	2.2	3.7	1.4	0.5	0
C.431.10.0055	2.77	0	0	6	6	12	31	29	18	9	1.2	3.1	2.9	1.8	0.9
C.431.12.0010	3.28	30	11	3	4	28	28	10	1	2	2.8	2.8	1	0.1	0.2
C.441.11.0041	3.17	2	6	5	5	26	31	11	1	8	2.6	3.1	1.1	0.1	0.8
C.431.08.0043	2.97	1	6	8	8	15	38	14	3	7	1.5	3.8	1.4	0.3	0.7
C.431.11.0080	3.01	2	6	5	5	19	31	16	2	3	1.9	3.1	1.6	0.2	0.3
C.431.11.0039	2.8	0	0	5	5	13	33	24	5	11	1.3	3.3	2.4	0.5	1.1
C.431.11.0068	3.01	2	6	5	5	14	38	18	3	3	1.4	3.8	1.8	0.3	0.3
C.431.12.0071	3.24	1	6	4	4	28	27	10	3	0	2.8	2.7	1	0.3	0
C.431.12.0100	3.56	1	6	4	4	38	23	4	2	1	3.8	2.3	0.4	0.2	0.1
C.431.12.0080	3.23	1	6	4	4	22	35	9	1	0	2.2	3.5	0.9	0.1	0
C.431.12.0055	3.37	30	11	3	4	28	32	9	0	3	2.8	3.2	0.9	0	0.3
C.411.12.0021	3.22	30	11	3	4	28	27	10	1	2	2.8	2.7	1	0.1	0.2
C.411.12.0019	3.15	30	11	3	4	21	34	10	0	1	2.1	3.4	1	0	0.1
C.411.12.0032	3.14	30	11	3	4	25	28	12	2	0	2.5	2.8	1.2	0.2	0
C.431.12.0063	3.46	1	6	4	4	35	26	4	1	3	3.5	2.6	0.4	0.1	0.3
C.411.12.0002	3.32	30	11	3	4	31	24	9	2	0	3.1	2.4	0.9	0.2	0
C.431.11.0064	3.21	2	6	5	5	23	33	11	0	3	2.3	3.3	1.1	0	0.3
C.411.12.0003	3.1	30	11	3	4	22	29	13	2	0	2.2	2.9	1.3	0.2	0

Tabel 3 Normalisasi Data Nilai Mahasiswa

Dan hasil akhir normalisasi adalah sebagai berikut :

NIM	IPK	THN	A	B	C	D	E
C.431.09.0004	3.10	7.00	1.80	3.70	1.00	0.20	1.00
C.431.12.0049	3.17	4.00	2.20	3.70	1.40	0.50	0.00
C.431.10.0055	2.77	6.00	1.20	3.10	2.90	1.80	0.90
C.431.12.0010	3.28	4.00	2.80	2.80	1.00	0.10	0.20
C.441.11.0041	3.17	5.00	2.60	3.10	1.10	0.10	0.80
C.431.08.0043	2.97	8.00	1.50	3.80	1.40	0.30	0.70
C.431.11.0080	3.01	5.00	1.90	3.10	1.60	0.20	0.30
C.431.11.0039	2.80	5.00	1.30	3.30	2.40	0.50	1.10
C.431.11.0068	3.01	5.00	1.40	3.80	1.80	0.30	0.30
C.431.12.0071	3.24	4.00	2.80	2.70	1.00	0.30	0.00
C.431.12.0100	3.56	4.00	3.80	2.30	0.40	0.20	0.10
C.431.12.0080	3.23	4.00	2.20	3.50	0.90	0.10	0.00
C.431.12.0055	3.37	4.00	2.80	3.20	0.90	0.00	0.30
C.411.12.0021	3.22	4.00	2.80	2.70	1.00	0.10	0.20
C.411.12.0019	3.15	4.00	2.10	3.40	1.00	0.00	0.10
C.411.12.0032	3.14	4.00	2.50	2.80	1.20	0.20	0.00
C.431.12.0063	3.46	4.00	3.50	2.60	0.40	0.10	0.30
C.411.12.0002	3.32	4.00	3.10	2.40	0.90	0.20	0.00
C.431.11.0064	3.21	5.00	2.30	3.30	1.10	0.00	0.30
C.411.12.0003	3.10	4.00	2.20	2.90	1.30	0.20	0.00

Tabel 4 Normalisasi Ke 2 Data Nilai Mahasiswa

3.2 Algoritma *K-Means* dan *Fuzzy C-Means*

Tahapan K-means

1. Inisialisasi : tentukan nilai K sebagai jumlah cluster yang di inginkan dan matrik ketidakmiripan (jarak) yang di inginkan. Jika perlu, tetapkan ambang batas perubahan fungsi objektif dan ambang batas perubahan posisi centroid.
2. Pilik K data dari set data X sebagai centroid
3. Alokasikan semua data ke centroid terdekat dengan metrik jarak yang sudah ditetapkan (memperbarui cluster ID setiap data)
4. Hitung kembali centroid C berdasarkan data yang mengikuti cluster masing-masing
5. Ulangi langkah 3 dan 4 hingga kondisi konvergen tercapai yaitu (a) perubahan fungsi objektif sudah dibawah ambang batas yang di inginkan, atau (b) tidak ada data yang berpindah cluster, atau (c) perubahan posisi centroid dibawah ambang batas yang ditetapkan.

C1	2.45	2.99	1.50	0.34	0.30	3.13	0.42
C2	3.72	2.39	0.78	0.15	0.15	3.42	0.39
C3	1.90	3.14	2.40	0.89	1.40	2.88	0.54

Tabel 5 Tabel Centroid Akhir

Pada perhitungan *K-Means* didapatkan hasil fungsi objektif “0” pada interaksi ke **13**, dengan validitas *DBI* sebesar **2.021897468**.

3.3 Fuzzy *C-Means (FCM)*

1. Inisialisasi : tentukan jumlah cluster (k ≥ 2), tentukan bobot pangkat (w > 1, tentukan jumlah maksimal interasi, tentukan ambang batas perubahan nilai fungsi objektif (jika perlu juga perubahan nilai centroid)
2. Berikan nilai awal pada matriks *fuzzy pseudo-partition*, dengan syarat seperti pada persamaan

$$\sum_{j=1}^{k} u_{ij} = 1$$

3. Lakukan langkah 4 sampai 5 selama syarat masih terpenuhi : (1) apabila perubahan pada nilai fungsi objektif masih di atas nilai ambang batas yang ditentukan, (2) perubahan pada nilai centroid masih di atas nilai ambang batas yang ditentukan, (3) iterasi maksimal belum tercapai.
4. Hitung nilai centroid dari masing-masing cluster menggunakan persamaan

$$c_{lj} = \frac{\sum_{i=1}^{N} (u_{il})^w x_{ij}}{\sum_{i=1}^{N} (u_{il})^w}$$

5. Hitung kembali matriks *fuzzy pseudo partition* (derajat keanggotaan setiap data pada setiap cluster) menggunakan persama.

Awal	463.50
Nilai J	463.50
Perubahan J	0

Tabel 6 Tabel Nilai Fungsi Objektif Iterasi ke 27

Dari hasil perhitungan *FCM*, fungsi objektif “0” pada interkasi ke 27 dengan hasil validitas *XBI* sebesar **11.94809958**.

3.4 Validitas

Evaluasi sistem merupakan bagian penting dalam model pekerjaan klasifikasi dengan metode evaluasi diharapkan dapat menghitung keajurasian sebuah metode dengan data tertentu. Disini kita menggunakan dua metode yaitu metode K-means dan FCM, untuk K-Means digunakan validitas Davies-Bouldin Index (DBI) sedangkan FCM kita menggunakan Xie Beni (XBI)

Tahapan perhitungan Validitas ***DBI***

1. Menghitung *Sum of square within Cluster (**SSW**)*

$$SSW_i = \frac{1}{m_i}\sum_{j=1}^{m_i} d(x_j, c_i)$$

2. Menghitung *Sum of square between cluster (**SSB**)*

$$SSSB_{i,j} = d(c_i, c_j)$$

3. Mendefinisikan ukuran rasio seberapa baik nilai atara cluster ($R_{i,j}$)

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}}$$

4. Menghitung *Davis-Bouldin Index (**DBI**)*

$$DBI = \frac{1}{K}\sum_{i=1}^{K} \max_{i \neq j}(R_{i,j})$$

Tahapan perhitungan ***XBI***

$$XBI = \frac{\sum_{j=1}^{K}\sum_{i=1}^{N} u_{ij}^{m} \times d(x_i, c_j)^2}{N \times \min_{i,j}(d()^2)} = \frac{J_m(u,c)/N}{Sep(c)}$$

4. Kesimpulan

a. K-Means Lebih cepat dari pada FCM, karena interasi K-Means hingga interasi ke **13** sedangkan FCM hingga interasi ke **27**
b. Dari validitas K-Means lebih kecil lebih baik, dengan validitas DBI sebesar **2.021897468,** sedangkan FCM dengan validitas XBI didapatkan validitas sebesar **11.94809958**
c. Dengan K-Means data mahasiswa tepat waktu di 3 cluster didapatkan sebagai berikut :

	A	B	C	D	E	IPK	THN
C1	2.45	2.99	1.5	0.34	0.3	3.13	0.42
C2	3.72	2.39	0.78	0.15	0.15	3.42	0.39
C3	1.9	3.14	2.4	0.89	1.4	2.88	0.54

Dengan kesimpulan lulusan lama studi 4 tahun lebih 2 bulan dengan ipk rata-rata 3.13, sedangkan lama studi 4 tahun dengan ipk 3.42. dan untuk mahasiswa dengan lama studi 5 tahun lebih rata-rata ipk 2.88

Daftar Pustaka

Prasetyo, Eko (2014). *Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab*. Andi Publisher

Prasetyo, Eko (2013). *Data Mining (Konsep* dan Aplikasi Menggunakan *Matlab)*. Andi Publisher

Kusrini dan EmhaTaufiq Lutfi (2009). *Algoritma Data Mining*. Andi Publisher

Astuti, Fajar Hermawati (2013). *Data Mining*. Andi Publisher.

Connolly, Thomas, and Carolyn Begg. "Database Systems, 5th." (2010).

Amalia, N., Shaufiah, S., & Sa'adah, S. (2015). Penerapan Teknik Data Mining untuk Klasifikasi Ketepatan Waktu Lulus Mahasiswa Teknik Informatika Universitas Telkom Menggunakan Algoritma Naive Bayes Classifier. eProceedings of Engineering, 2(3).

ISSN:2442-5567 | E-ISSN:2443-289X

PENENTUAN PENERIMA BEASISWA DENGAN ALGORITMA *FUZZY C-MEANS* DI UNIVERSITAS MEGOW PAK TULANG BAWANG

Muhardi[1] Nisar[2]

[1]Universitas Megow Pak [2]MTI Institut Informatika & Bisnis Darmajaya
Email: ardyardy68@yahoo.co.id

ABSTRAK

Beasiswa adalah pemberian berupa bantuan keuangan yang diberikan kepada perorangan yang bertujuan untuk digunakan demi keberlangsungan pendidikan yang ditempuh. Penentuan beasiswa dapat dikelompokan berdasarkan kriteria Indeks Prestasi Kumulatif (IPK), tingkat kemiskinan, tanggungan orang tua dan prestasi mahasiswa dalam proses rekruitmen beasiswa. Algoritma *Fuzzy C-Means* merupakan satu algoritma yang mudah dan sering digunakan dalam pengelompokan data karena membuat suatu perkiraan yang efisien dan tidak memerlukan banyak parameter. Beberapa penelitian telah menghasilkan kesimpulan bahwa metode *Fuzzy C-Means* dapat digunakan untuk mengelompokkan data berdasarkan atribut-atribut tertentu. Penerapan algoritma *Fuzzy C-Means* dalam penentuan beasiswa di kelompokan menjadi tiga *cluster* yaitu menerima, dipertimbangkan dan tidak berhak menerima beasiswa , sempel data sebenyak 75 data mahasiswa diperoleh tiga *cluster* berdasarkan nilai rata-rata penentuan beasiswa kemudian setiap *cluster* diklasifikasikan berdasarkan kriteria mana yang lebih diprioritaskan dengan nilai terbesar pada jarak akhir merupakan *cluster* yang menerima beasiswa, sedangkan *cluster* dengan nilai terkecil merupakan *cluster* yang tidak berhak menerima beasiswa.

Kata Kunci: *Beasiswa, clustering, Fuzzy C- Means.*

ABSTRACT

The scholarship is the grant in the form of financial assistance given to an individual who intends to use for the sake of the continuity of education traveled. Determination of the scholarship can be grouped based on the criteria of the cumulative Achievement Index (IPK), the level of poverty, dependent parent and student achievement in the process of recruitment of scholarship. Fuzzy C-Means algorithm is an algorithm that is easily and often used in grouping data because making an estimate of efficient and doesn't require a lot of parameters. Some research has resulted in the conclusion that the Fuzzy C-Means method can be used to classify data based on certain attributes. Application of Fuzzy C-Means algorithm in determination of scholarship in kelompokan into three clusters, namely receiving, considering and not eligible to receive the scholarship, sempel 75 student data sebenyak data retrieved three clusters based on the average value of the determination of a scholarship then each cluster are classified according to criteria which is more prioritized with the greatest value at the end of the cluster is receiving a scholarship, whereas the smallest value cluster is a cluster that is not eligible to receive the scholarship.

Keywords*: scholarship, clustering, Fuzzy C- Means*

ISSN:2442-5567 | E-ISSN:2443-289X

1. PENDAHULUAN

Beasiswa adalah pemberian berupa bantuan keuangan yang diberikan kepada perorangan yang diberikan kapada perorangan yang bertujuan untuk digunakan demi keberlansungan pendidikan yang ditempuh. Pemberian beasiswa merupakan program kerja yang ada di setiap universitas atau perguruan tinggi. Program beasiswa diadakan untuk meringankan beban mahasiswa dalam menempuh masa studi kuliah khususnya dalam masalah biaya. Pemberian beasiswa kepada mahasiswa dilakukan secara selektif sesuai dengan jenis beasiswa yang diadakan. Univeritas Megow Pak Tulang Bawang menyediakan beberpa program beasiswa, sebagai contoh yaitu beasiswa Peningkatan Akademik (PPA), Beasiswa Bantuan Belajar Mahasiswa (BBM) dan lain sebagainya. Indeks prestasi kumulatif, Tingkat Kemiskinan, tanggungan orang tua dan prestasi menjadi kriteria dalam proses rekruitmen beasiswa.

Proses seleksi penerimaan beasiswa secara manual yaitu dengan menginputkan satu persatu data mahasiswa ke dalam file *excel* kemudian melakukan *sorting* data mahasiswa seringkali melakukan beberapa permasalahan, antara lain membutuhkan waktu yang lama dan ketelitian yang tinggi. Selain itu, transparansi serta ketidak jelasan metodologi yang digunakan dalam proses komputasi penerimaan beasiswa juga menjadi salah satu permasalahan, sehingga dibutuhkan suatu sistem yang dapat membantu dalam proses pengambilan keputusan siapa saja mahasiswa yang menerima beasiswa berdasarkan kriteria-kriteria yang telah ditentukan secara capat dan tepat sasaran.

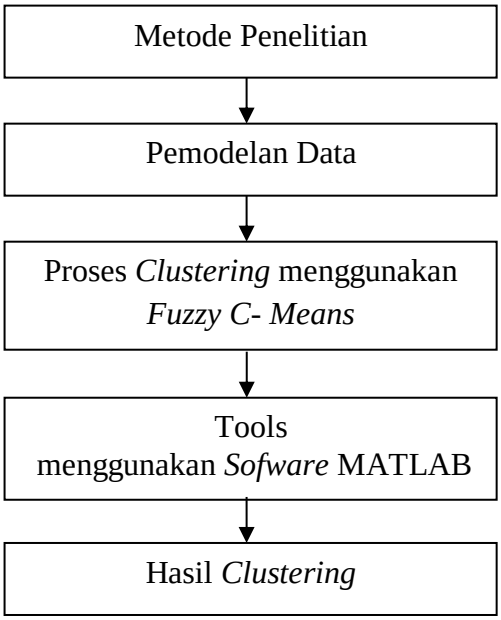
Salah satu metode yang dapat digunakan untuk menyelesaikan permasalahan tersebut adalah metode *Fuzzy C-Means.* Algoritma *Fuzzy C-Means* merupakan satu algoritma yang mudah dan sering digunakan dalam pengelompokan data kerana membuat suatu perkiraan yang efsien dan tidak memerlukan banyak parameter. Beberapa penelitian telah menghasilkan kesimpulan bahwa metode *Fuzzy CMeans* dapat digunakan untuk mengelompokkan data berdasarkan atribut-atribut tertentu. Pada kasus penelitian ini akan menganalisis penerapan metode *Fuzzy C-Means* untuk mengelompokkan data mahasiswa berdasarkan kemampuan mahasiswa dibidang akademik untuk proses penentuan beasiswa.
Pada penelitian sebelumnya, Bahari (2011) melakukan Penerapan metode *Fuzzy C-Means* untuk mengelompokan data siswa dalam penentuan jurusan di Sekolah Menengah Atas pada 81 sampel data siswa yang diuji dalam penelitian menunjukan bahwa metode *Fuzzy C-Means* memiliki tingkat akurasi yang lebih tinggi (rata-rata 78,39%), jika dibandingkan dengan metode penentuan jurusan secara manual yang di lakukan (hanya memiliki tingkat akurasi rata-rata 56,17 %).

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitan

Penelitian ini akan dilaksanakan berdasarkan rancangan penelitian seperti yang di tunjukan pada Gambar 1.

ISSN:2442-5567 | E-ISSN:2443-289X



Gambar 1. Alur Kerangka Penelitian

Penelitian yang dilaksanakan adalah jenis penelitian eksprimen, yaitu melakukan pengujian tingkat akurasi algoritma *Fuzzy C-Means clustering* sebagai model untuk menentukan penerima beasiswa. Data eksperimen diambil dari tempat penelitian yaitu Universitas Megow Pak Tulang Bawang.

2.2 Pengumpulan Data

Pengumpulan data digunakan untuk mengumpulkan data-data dan informasi-informasi yang diperlukan dalam pembuatan sistem pendukung keputusan. Pengumpulan data pada penelitian ini menggunakan metode pengumpulan data *study literature* dan telaah dokumen.

a. *Study Literature*

Study literature dilakukan dengan cara mencari bahan materi yang berhubungan dengan permasalahan, algoritma *Fuzzy C-Means*, sistem pendukung keputusan dan beasiswa, guna mempermudah proses implementasi sistem. Pencarian materi dilakukan melalui pencarian di buku panduan dan internet.

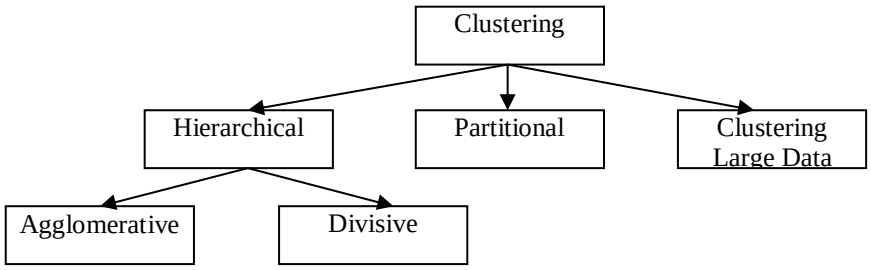
b. Telaah Dokumen

Telaah dokumen adalah pengumpulan data dengan cara mengumpulkan dan mempelajari dokumen-dokumen yang didapatkan dari pihak Jurusan Informaitka FMIPA Universitas Megow Pak Tulang Bawang. Dari metode pengumpulan data ini diperoleh 75 data mahasiswa.

2.3 Algoritma *Clustering*

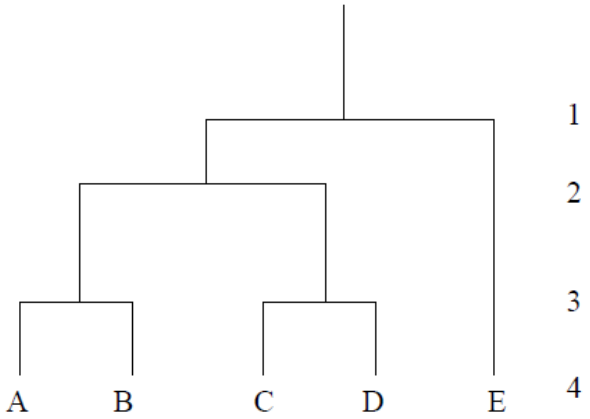
Secara umum pembagian algoritma *clustering* dapat digambarkan sebagai berikut:

ISSN:2442-5567 | E-ISSN:2443-289X



Gambar 2. Kategori Algoritma *Clustering*

Hierarchical clustering menentukan sendiri jumlah *cluster* yang dihasilkan. Hasil dari metode ini adalah suatu struktur data berbentuk pohon yang disebut dendogram dimana data dikelompokkan secara bertingkat dari yang paling bawah dimana tiap *instance* data merupakan satu *cluster* sendiri, hingga tingkat paling atas dimana keseluruhan data membentuk satu *cluster* besar berisi *cluster-cluster* seperti gambar 2.



Gambar 3. *Dendogram*

Divisive hierarchical clustering mengelompokkan data dari kelompok yang terbesar hingga ke kelompok yang terkecil, yaitu masing-masing *instance* dari kelompok data tersebut. Sebaliknya, *agglomerative hierarchical clustering* mulai mengelompokkan data dari kelompok yang terkecil hingga kelompok yang terbesar. Beberapa algoritma yang menggunakan metode ini adalah: *RObust Clustering Using LinKs* (ROCK), *Chameleon, Cobweb, Shared Nearest Neighbor* (SNN).

Partitional clustering yang mengelompokkan data ke dalam k *cluster* dimana k adalah banyaknya *cluster* dari input user. Kategori ini biasanya memerlukan pengetahuan yang cukup mendalam tentang data dan proses bisnis yang memanfaatkannya untuk mendapatkan kisaran nilai input yang sesuai. Beberapa algoritma yang masuk dalam kategori ini antara lain: *K-Means, Fuzzy C-Means, Clustering Large Aplications* (CLARA), *Expectation Maximation* (EM), *Bond Energy Algorithm* (BEA), *algoritma Genetika*, Jaringan Saraf Tiruan.

ISSN:2442-5567 | E-ISSN:2443-289X

a. Algoritma *Fuzzy Clustering C-Means* (FCM)

Pada proses pengklasteran (*clustering*) secara klasik (misalnya pada algoritma *Clustering K-Means*), pembentukan partisi dilakukan sedemikian rupa sehingga setiap obyek berada tepat pada satu partisi. Namun, adakalanya tidak dapat menempatkan suatu obyek tepat pada suatu partisi, karena sebenarnya obyek tersebut terletak di antara 2 atau lebih partisi yang lain. Pada logika *fuzzy*, metode yang dapat digunakan untuk melakukan pengelompokan sejumlah data dikenal dengan nama *fuzzy clustering*. *Fuzzy Clustering* lebih alami jika dibandingkan dengan pengklasteran secara klasik. Suatu algoritma *clustering* dikatakan sebagai *fuzzy clustering* jika algoritma tersebut menggunakan parameter strategi adaptasi secara *soft competitive*. Sebagian besar algoritma *fuzzy clustering* didasarkan atas optimasi fungsi obyektif atau modifikasi dari fungsi obyektif tersebut.

Salah satu teknik *fuzzy clustering* adalah *Fuzzy C-Means* (FCM). FCM adalah suatu teknik pengklasteran data yang keberadaan tiap-tiap data dalam suatu *cluster* ditentukan oleh nilai/derajat keanggotaan tertentu. Teknik ini pertama kali diperkenalkan oleh Jim Bezdek pada tahun 1981. Berbeda dengan teknik pengklasteran secara klasik (dimana suatu obyek hanya akan menjadi anggota suatu klaster tertentu), dalam FCM setiap data bisa menjadi anggota dari beberapa *cluster*.Batas-batas *cluster* dalam FCM adalah lunak (*soft*). Konsep dasar FCM, pertama kali adalah menentukan pusat *cluster* yang akan menandai lokasi rata-rata untuk tiaptiap *cluster*. Pada kondisi awal, pusat *cluster* ini masih belum akurat. Tiap-tiap data memiliki derajat keanggotaan untuk tiap-tiap *cluster*. Dengan cara memperbaiki pusat *cluster* dan nilai keanggotaan tiap-tiap data secara berulang, maa akan terlihat bahwa pusat *cluster* akan bergerak menuju lokasi yang tepat. Perulangan ini didasarkan pada minimasi fungsi obyektif. Fungsi Obyektif yang digunakan pada FCM adalah:

$$J_w(U,V;X) = \sum_{k-1}^{n} \sum_{i-=1}^{c} (\mu_{ik})^w (d_{ik})^2 \quad (1)$$

dengan + ∈ ,1,∞),

$$d_{ik} = d(x_k - v_i) = [\sum_{j=1}^{m} (x_{kj} - x_{ij})]^{\frac{1}{2}} \quad (2)$$

x adalah data yang akan diklaster:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix} \quad (3)$$

dan v adalah matriks pusat *cluster* :

$$V = \begin{bmatrix} v_{11} & \cdots & v_{1m} \\ \vdots & & \vdots \\ v_{m1} & \cdots & v_{mm} \end{bmatrix} \quad (4)$$

nilai Jw terkecil adalah yang terbaik, sehingga:

$$J_w^* (U^*, U^*; x) = \min J(U, U; x) \quad (5)$$

Jika $d_{ik} > 0, \forall i, k$; $w > 1$ dan X setidaknya memiliki m elemen, maka (u, v) ∈ $M_{fm} > 0, \forall i, k$; w dapat meminimasi Jw hanya jika:

$$\mu_{ik}=\frac{[\sum_{j=1}^{m}(x_{ij}-v_{kj})^{2}]^{\frac{-1}{w-1}}}{\sum_{i=1}^{n}[\sum_{j=1}^{m}(x_{ij}-v_{kj})^{2}]^{\frac{-1}{w-1}}}; \quad (6)$$

dan

$$v_{kj}=\frac{\sum_{i=1}^{n}((\mu_{ik})^{w}*X_{ij})}{\sum_{i=1}^{n}(\mu_{ik})^{w}};1 \leq i \leq m;1 \leq J \leq m \quad (7)$$

Algoritma *Fuzzy C-Means* (FCM) diberikan sebagai berikut :

1. Menentukan data yang akan di cluster X, berupa matriks berukuran n x m (n=jumlah sampel data, m = atribut setiap data). Xij=data sampel ke-I (i=1,2,…,n), atribut ke-j (j=1,2,…,m).
2. Menentukan :
   - Jumlah cluster = c
   - Pangkat = w
   - Maksimum interasi = MaxIter
   - Error terkecil yang diharapkan = ξ
   - Fungsi objektif awal = $P_0 = 0$
   - Interasi awal = t = 1

3. Membangkitkan bilangan random μik, i=1,2,3 …,n; k=1,2,3…c; sebagai elemen-elemen matriks partisi awal U.
   Menghitung jumlah setiap kolom :

$$Q_i=\sum_{k=1}^{c}\mu_{ik} \quad (8)$$

dengan j=1,2,…n.

Menghitung :

$$\mu_{ik}=\frac{\mu_{ik}}{Q_i} \quad (9)$$

4. Menghitung pusat cluster ke-k: V_{kj}, dengan k=1,2,…c; dan j=1,2,…m

$$V_{kj}=\frac{\sum_{i=1}^{n}((\mu_{ik})^{w}*X_{ij}}{\sum_{i=1}^{n}((\mu_{ik})^{w}} \quad (10)$$

5. Menghitung fungsi objektif pada interasi ke-t :

$$P_t=\sum_{i=1}^{n}\sum_{k}^{c}([\sum_{j=1}^{m}(X_{ij}-V_{kj})^{2}](\mu_{ik})^{w}) \quad (11)$$

6. Menghitung perubahan matriks partisi :

$$\mu_{ik}=\frac{[\sum_{j=1}^{m}(X_{ij}-V_{kj})^{2}]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}[\sum_{j=1}^{m}(X_{ij}-V_{kj})^{2}]^{\frac{-1}{w-1}}} \quad (12)$$

b. Proses *Clustering* menggunakan *Fuzzy C-Means*

Tahap ini akan diterapkan metode *Fuzzy C-Means* untuk mengelompokkan data. Hasil pengelompokkan ini kemudian akan digunakan untuk pertimbangan menentukan

ISSN:2442-5567 | E-ISSN:2443-289X

mahasiswa yang berhak menerima beasiswa. Adapun algoritma *C-Means Clustering* pada penerima beasiswa adalah sebagai berikut :

1. Menetapkan matriks partisi awal U berupa matriks berukuran n x m (n adalah jumlah sampel data, yaitu=75, dan m adalah parameter/atribut setiap data, yaitu=2). X_{ij}=data sampel ke-i (i=1,2,…,n), atribut ke-j (j=1,2,..,m).
2. Menentukan Nilai Parameter Awal :
   - Jumlah *cluster* (c) = 3
   - Pangkat/bobot (w) = 2
   - Maksimum interasi (MaxIter) = 100
   - Error terkecil yang diharapkan (ξ)= 10^{-5}
   - Fungsi Objektif awal (Po) = 0
   - Interasi awal (t) = 1
3. Membangkitkan bilangan random μik, i=1,2,…c; sebagai elemen-elemen matriks partisi awal (U).

3. HASIL DAN PEMBAHASAN

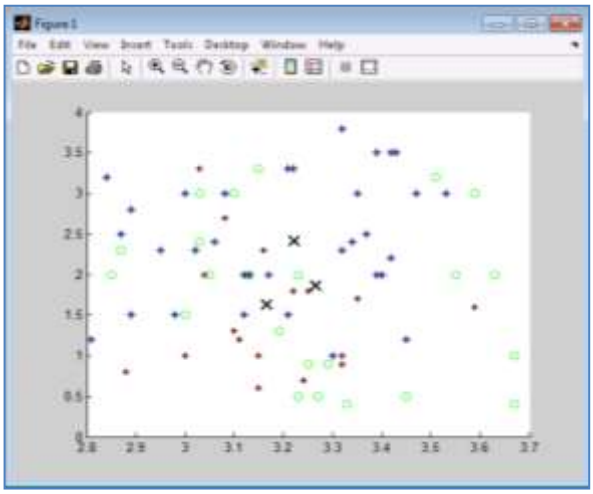
3.1 Tools menggunakan *Sofware* MATLAB

Data yang digunakan dalam penelitian ini adalah data penentuan beasiswa sebanyak 75 mahasiswa. Dalam proses *clustering*, IPK dijadikan sebagai Xi1, penghasilan orang tua dijadikan Xi2, tanggungan orang tua dijadikan Xi3 dan prestasi dijadikan Xi4. Pengolahan data pada penelitian ini menggunakan bantuan program yaitu Matlab. Dari *clustering* yang dilakukan diperoleh hasil yaitu nilai fungsi obyektif selama iterasi, pusat *cluster* atau *center* serta derajat keanggotaan lulusan untuk setiap *cluster* pada iterasi terakhir. Dalam penelitian ini, proses iterasinya berhenti pada iterasi ke-37 karena nilai |Pt - Pt-1|< ξ. Nilai fungsi obyektif pada iterasi terakhir yang diperoleh adalah 111.949781. Untuk penelitian ini *clustering* menggunakan *sofware* Matlab denga menjalankan di *Command Window* yaitu: >> fcm_k3.m

```
Iteration count = 1, obj. fcn = 185.869420
Iteration count = 2, obj. fcn = 146.530736
Iteration count = 3, obj. fcn = 141.726376
Iteration count = 4, obj. fcn = 134.833704
Iteration count = 5, obj. fcn = 130.961108
Iteration count = 6, obj. fcn = 129.145894
Iteration count = 7, obj. fcn = 127.219322
Iteration count = 8, obj. fcn = 124.903487
Iteration count = 9, obj. fcn = 122.880711
Iteration count = 10, obj. fcn = 121.359551
Iteration count = 11, obj. fcn = 119.898753
Iteration count = 12, obj. fcn = 118.236843
Iteration count = 13, obj. fcn = 116.491898
Iteration count = 14, obj. fcn = 114.963865
Iteration count = 15, obj. fcn = 113.865257
Iteration count = 16, obj. fcn = 113.180881
Iteration count = 17, obj. fcn = 112.770379
Iteration count = 18, obj. fcn = 112.513345
Iteration count = 19, obj. fcn = 112.342297
Iteration count = 20, obj. fcn = 112.223512
Iteration count = 21, obj. fcn = 112.139405
```

```
Iteration count = 22, obj. fcn = 112.079712
Iteration count = 23, obj. fcn = 112.037682
Iteration count = 24, obj. fcn = 112.008476
Iteration count = 25, obj. fcn = 111.988487
Iteration count = 26, obj. fcn = 111.975011
Iteration count = 27, obj. fcn = 111.966053
Iteration count = 28, obj. fcn = 111.960170
Iteration count = 29, obj. fcn = 111.956348
Iteration count = 30, obj. fcn = 111.953887
Iteration count = 31, obj. fcn = 111.952314
Iteration count = 32, obj. fcn = 111.951315
Iteration count = 33, obj. fcn = 111.950683
Iteration count = 34, obj. fcn = 111.950285
Iteration count = 35, obj. fcn = 111.950035
Iteration count = 36, obj. fcn = 111.949879
Iteration count = 37, obj. fcn = 111.949781
C =
   3.2649   1.8625   2.2484   4.2701
   3.1653   1.6271   5.1550   1.1770
   3.2225   2.4133   3.0759   0.7967
```

Penyebaran masing-masing anggota *cluster* pada iterasi terakhir dapat dilihat pada *cluster interface* gambar dibawah ini :



Gambar 4. Posisi *Claster* Iterasi terakhir

Keterangan :
Tanda * = *cluster* 1
Tanda + = *cluster* 2
Tanda ○ = *cluster* 3
Tanda X = pusat ke 3 *cluster*

Detail Matriks partisi U yang dihasilkan pada iterasi terakhir (iterasi ke-37) menggunakan fungsi Matlab adalah :

ISSN:2442-5567 | E-ISSN:2443-289X

Columns 1 through 10

0.1634	0.5724	0.0523	0.0467	0.9198	0.0643	0.0525	0.1063	0.0438	0.0605
0.1810	0.1653	0.0954	0.3021	0.0345	0.7542	0.2756	0.1726	0.3300	0.1892
0.6557	0.2623	0.8522	0.6512	0.0457	0.1815	0.6719	0.7211	0.6262	0.7503

Columns 11 through 20

0.0507	0.0576	0.6670	0.0923	0.6497	0.1132	0.6893	0.0108	0.7094	0.0427
0.7378	0.2660	0.1091	0.0971	0.1801	0.6406	0.1642	0.9393	0.0919	0.3048
0.2115	0.6763	0.2239	0.8107	0.1701	0.2463	0.1465	0.0500	0.1987	0.6525

Columns 21 through 30

0.0463	0.0299	0.8531	0.3603	0.0327	0.0592	0.0435	0.0470	0.8602	0.0845
0.7249	0.0572	0.0618	0.3797	0.8855	0.1855	0.3987	0.3511	0.0572	0.7507
0.2288	0.9129	0.0852	0.2600	0.0818	0.7553	0.5578	0.6018	0.0826	0.1648

Columns 31 through 40

0.8104	0.0560	0.1747	0.0183	0.3397	0.6931	0.0611	0.0556	0.2106	0.6219
0.0792	0.8198	0.6184	0.0430	0.3074	0.1693	0.1484	0.1314	0.5727	0.1510
0.1105	0.1242	0.2069	0.9386	0.3530	0.1376	0.7904	0.8130	0.2167	0.2271

Columns 41 through 50

0.0513	0.1739	0.5558	0.8307	0.8829	0.1085	0.0982	0.0307	0.0732	0.0897
0.8324	0.6221	0.2029	0.0788	0.0535	0.2561	0.1566	0.0896	0.4973	0.0934
0.1163	0.2040	0.2413	0.0905	0.0636	0.6354	0.7452	0.8797	0.4295	0.8170

Columns 51 through 60

0.0050	0.1026	0.0297	0.0466	0.0378	0.0673	0.0446	0.0707	0.5855	0.5157
0.9734	0.1039	0.0565	0.2871	0.1144	0.1508	0.3947	0.5757	0.1587	0.2346
0.0216	0.7934	0.9139	0.6663	0.8478	0.7819	0.5607	0.3536	0.2558	0.2497

Columns 61 through 70

0.0825	0.0826	0.0433	0.1208	0.6158	0.0877	0.1063	0.7764	0.0346	0.8865
0.2752	0.2753	0.3423	0.1954	0.1421	0.1292	0.1073	0.0887	0.0653	0.0543
0.6423	0.6421	0.6144	0.6838	0.2421	0.7831	0.7864	0.1349	0.9000	0.0592

Columns 71 through 75

0.8488	0.0548	0.7033	0.6666	0.7703
0.0635	0.4938	0.0943	0.1236	0.1006
0.0877	0.4514	0.2024	0.2097	0.1291

Pada iterasi terakhir (iterasi ke-37), pusat *cluster Vkj* yang dihasilkan dengan k=1,2,3,4 dan j=1,2 adalah:

$$V_{kj} = \begin{pmatrix} 3.2649 & 1.8625 & 2.2484 & 4.2701 \\ 3.1653 & 1.6271 & 5.1550 & 1.1770 \\ 3.2225 & 2.4133 & 3.0759 & 0.7967 \end{pmatrix}$$

ISSN:2442-5567 | E-ISSN:2443-289X

Nilai ini merupakan nilai dari koordinat ketiga titik pusat *cluster* dan memberikan garis besar tiap *cluster* yaitu:

1. Kelompok pertama (*cluster* ke-1), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.2649; memiliki rata-rata penghasilan orang tua sekitar 1.8625; tanggungan orang tua 2.2484 dan memiliki prestasi 4.2701.
2. Kelompok kedua (*cluster* ke-2), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.1653; memiliki rata-rata penghasilan orang tua sekitar 1.6271; tanggungan orang tua 5.1550 dan memiliki prestasi 1.1770.
3. Kelompok ketiga (*cluster* ke-3), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.2225; memiliki rata-rata penghasilan orang tua sekitar 2.4133; tanggungan orang tua 3.0759 dan memiliki prestasi 0.7967.

Derajat keanggotaan penentuan beasiswa untuk sertiap *cluster* pada iterasi terakhir (iterasi ke-37) dapat dilihat pada Tabel 1. di bawah ini:

Tabel 1. Derajat Keanggotaan Penentuan Beasiswa Untuk Setiap *Cluster* Pada Iterasi Terakhir

No	Derajat Keanggotaan (μ) pada Iterasi Terakhir			Kecenderungan data masuk pada *cluster*		
	(μ_1)	(μ_2)	(μ_3)	C1	C2	C3
1	0.1634	0.1810	0.6557			*
2	0.5724	0.1653	0.2623	*		
3	0.0523	0.0954	0.8522			*
4	0.0467	0.3021	0.6512			*
5	0.9198	0.0345	0.0457	*		
6	0.0643	0.7542	0.1815		*	
7	0.0525	0.2756	0.6719			*
8	0.1063	0.1726	0.7211			*
9	0.0438	0.3300	0.6262			*
10	0.0605	0.1892	0.7503			*
11	0.0507	0.7378	0.2115		*	
12	0.0576	0.2660	0.6763			*
13	0.6670	0.1091	0.2239	*		
14	0.0923	0.0971	0.8107			*
15	0.6497	0.1801	0.1701	*		
16	0.1132	0.6406	0.2463		*	
17	0.6893	0.1642	0.1465	*		
18	0.0108	0.9393	0.0500		*	
19	0.7094	0.0919	0.1987	*		
20	0.0427	0.3048	0.6525			*
21	0.0463	0.7249	0.2288		*	
22	0.0299	0.0572	0.9129			*
23	0.8531	0.0618	0.0852	*		
24	0.3603	0.3797	0.2600		*	
25	0.0327	0.8855	0.0818		*	
26	0.0592	0.1855	0.7553			*

ISSN:2442-5567 | E-ISSN:2443-289X

Nilai ini merupakan nilai dari koordinat ketiga titik pusat *cluster* dan memberikan garis besar tiap *cluster* yaitu:

1. Kelompok pertama (*cluster* ke-1), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.2649; memiliki rata-rata penghasilan orang tua sekitar 1.8625; tanggungan orang tua 2.2484 dan memiliki prestasi 4.2701.
2. Kelompok kedua (*cluster* ke-2), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.1653; memiliki rata-rata penghasilan orang tua sekitar 1.6271; tanggungan orang tua 5.1550 dan memiliki prestasi 1.1770.
3. Kelompok ketiga (*cluster* ke-3), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.2225; memiliki rata-rata penghasilan orang tua sekitar 2.4133; tanggungan orang tua 3.0759 dan memiliki prestasi 0.7967.

Derajat keanggotaan penentuan beasiswa untuk sertiap *cluster* pada iterasi terakhir (iterasi ke-37) dapat dilihat pada Tabel 2. di bawah ini:

Tabel 2. Derajat Keanggotaan Penentuan Beasiswa Untuk Setiap *Cluster* Pada Iterasi Terakhir

No	Derajat Keanggotaan (μ) pada Iterasi Terakhir			Kecenderungan data masuk pada *cluster*		
	(μ_1)	(μ_2)	(μ_3)	C1	C2	C3
1	0.1634	0.1810	0.6557			*
2	0.5724	0.1653	0.2623	*		
3	0.0523	0.0954	0.8522			*
4	0.0467	0.3021	0.6512			*
5	0.9198	0.0345	0.0457	*		
6	0.0643	0.7542	0.1815		*	
7	0.0525	0.2756	0.6719			*
8	0.1063	0.1726	0.7211			*
9	0.0438	0.3300	0.6262			*
10	0.0605	0.1892	0.7503			*
11	0.0507	0.7378	0.2115		*	
12	0.0576	0.2660	0.6763			*
13	0.6670	0.1091	0.2239	*		
14	0.0923	0.0971	0.8107			*
15	0.6497	0.1801	0.1701	*		
16	0.1132	0.6406	0.2463		*	
17	0.6893	0.1642	0.1465	*		
18	0.0108	0.9393	0.0500		*	
19	0.7094	0.0919	0.1987	*		
20	0.0427	0.3048	0.6525			*
21	0.0463	0.7249	0.2288		*	
22	0.0299	0.0572	0.9129			*
23	0.8531	0.0618	0.0852	*		
24	0.3603	0.3797	0.2600		*	
25	0.0327	0.8855	0.0818		*	
26	0.0592	0.1855	0.7553			*

Nilai ini merupakan nilai dari koordinat ketiga titik pusat *cluster* dan memberikan garis besar tiap *cluster* yaitu:

1. Kelompok pertama (*cluster* ke-1), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.2649; memiliki rata-rata penghasilan orang tua sekitar 1.8625; tanggungan orang tua 2.2484 dan memiliki prestasi 4.2701.
2. Kelompok kedua (*cluster* ke-2), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.1653; memiliki rata-rata penghasilan orang tua sekitar 1.6271; tanggungan orang tua 5.1550 dan memiliki prestasi 1.1770.
3. Kelompok ketiga (*cluster* ke-3), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.2225; memiliki rata-rata penghasilan orang tua sekitar 2.4133; tanggungan orang tua 3.0759 dan memiliki prestasi 0.7967.

Derajat keanggotaan penentuan beasiswa untuk sertiap *cluster* pada iterasi terakhir (iterasi ke-37) dapat dilihat pada Tabel 3. di bawah ini:

Tabel 3. Derajat Keanggotaan Penentuan Beasiswa Untuk Setiap *Cluster* Pada Iterasi Terakhir

No	Derajat Keanggotaan (μ) pada Iterasi Terakhir			Kecenderungan data masuk pada *cluster*		
	(μ_1)	(μ_2)	(μ_3)	C1	C2	C3
1	0.1634	0.1810	0.6557			*
2	0.5724	0.1653	0.2623	*		
3	0.0523	0.0954	0.8522			*
4	0.0467	0.3021	0.6512			*
5	0.9198	0.0345	0.0457	*		
6	0.0643	0.7542	0.1815		*	
7	0.0525	0.2756	0.6719			*
8	0.1063	0.1726	0.7211			*
9	0.0438	0.3300	0.6262			*
10	0.0605	0.1892	0.7503			*
11	0.0507	0.7378	0.2115		*	
12	0.0576	0.2660	0.6763			*
13	0.6670	0.1091	0.2239	*		
14	0.0923	0.0971	0.8107			*
15	0.6497	0.1801	0.1701	*		
16	0.1132	0.6406	0.2463		*	
17	0.6893	0.1642	0.1465	*		
18	0.0108	0.9393	0.0500		*	
19	0.7094	0.0919	0.1987	*		
20	0.0427	0.3048	0.6525			*
21	0.0463	0.7249	0.2288		*	
22	0.0299	0.0572	0.9129			*
23	0.8531	0.0618	0.0852	*		
24	0.3603	0.3797	0.2600		*	
25	0.0327	0.8855	0.0818		*	
26	0.0592	0.1855	0.7553			*

ISSN:2442-5567 | E-ISSN:2443-289X

1. Kelompok pertama (*cluster* ke-1), akan berisi data mahasiswa ke : 2, 5, 13, 15, 17, 19, 23, 29, 31, 36, 40, 43, 44, 45, 59, 60, 65, 68, 70, 71, 73, 74 dan 75.
2. Kelompok kedua (*cluster* ke-2), akan berisi data mahasiswa ke : 6, 11, 16, 18, 21, 24, 25, 30, 32, 33, 39, 41, 42, 49, 51, 58 dan 72.
3. Kelompok ketiga (*cluster* ke-3), akan berisi data mahasiswa ke : 1, 3, 4, 7, 8, 9, 10, 12, 14, 20, 22, 26, 27, 28, 34, 35, 37, 38, 46, 47, 48, 50, 52, 53, 54, 55, 56, 57, 61, 62, 63, 64, 66, 67 dan 69.

3.2 PEMBAHAAN

Proses *Clustering* memerlukan klasifikasi untuk menentukan kelompok (*cluster*) mana yang berhak untuk menerima beasiswa. Dalam penelitian akan mengelompokkan mahasiswa mejadi tiga (3) *cluster* yaitu :

1. *Cluster* yang menerima beasiswa
2. *Cluster* yang dipertimbangkan menerima beasiswa.
3. *Cluster* yang tidak menerima beasiswa.

Kemudian setiap *cluster* dibagi berdasarkan kriteria mana yang lebih diprioritaskan (berdasarkan IPK atau PO (Penghasilan Orang tua), Tanggungan Orang Tua (TO), Prestasi mahasiswa selama menjadi mahasiswa).

Iterasi pada percobaan ini berhenti pada iterasi ke-37. Hasil akhir *clutering* yang di perolah adalah :

1. *Cluster* pertama memiliki pusat *cluster* (3.2649; 1.8625; 2.2484; 4.2701)
2. *Cluster* kedua memiliki pusat *cluster* (3.1653; 1.6271; 5.1550; 1.1770)
3. *Cluster* ketiga memiliki pusat *cluster* (3.2225; 2.4133; 3.0759; 0.7967)

Hasil klasifikasi *clustering* dapat dilihat pada Tabel 4.

Tabel 4. Hasil Kalasifikasi

Proritas IPK		
Menerima	Dipertimbangkan	Tidak Menerima
Cluster 1	*Cluster* 3	*Cluster* 2
3.2649	3.2225	3.1653
Prioritas TK		
Menerima	Dipertimbangkan	Tidak Menerima
Cluster 3	*Cluster* 1	*Cluster* 2
2.4133	1.8625	1.6271
Prioritas TO		
Menerima	Dipertimbangkan	Tidak Menerima
Cluster 2	*Cluster* 3	*Cluster* 1
5.1550	3.0759	2.2484
Prioritas Prestasi		
Menerima	Dipertimbangkan	Tidak Menerima
Cluster 1	*Cluster* 2	*Cluster* 3
4.2701	1.1770	0.7967

ISSN:2442-5567 | E-ISSN:2443-289X

Hasil klasifikasi mahasiswa berdasarkan Tabel 4 untuk 75 data dapat dilihat pada Tabel 5 di bawah ini.

Tabel 5. Hasil klasifikasi prioritas mahasiswa

No	IPK	PO	TO	Prestasi	*C*	IPK	PO	TO	Prestasi
1	2.81	1.2	2	1	3	DP	DP	M	TM
2	2.85	2	3	3	1	M	TM	DP	M
3	2.84	3.2	3	1	3	DP	DP	M	TM
4	2.87	2.5	4	1	3	DP	DP	M	TM
5	2.87	2.3	2	5	1	M	TM	DP	M
6	2.88	0.8	6	0	2	TM	M	TM	DP
7	2.89	2.8	4	1	3	DP	DP	M	TM
8	2.89	1.5	2	0	3	DP	DP	M	TM
9	2.95	2.3	4	1	3	DP	DP	M	TM
10	2.98	1.5	3	0	3	DP	DP	M	TM
11	3	1	5	0	2	TM	M	TM	DP
12	3	3	4	1	3	DP	DP	M	TM
13	3	1.5	1	3	1	M	TM	DP	M
14	3.02	2.3	2	1	3	DP	DP	M	TM
15	3.03	3	4	5	3	DP	DP	M	TM
16	3.03	3.3	7	1	2	TM	M	TM	DP
17	3.03	2.4	4	5	1	M	TM	DP	M
18	3.04	2	5	1	2	TM	M	TM	DP
19	3.05	2	2	3	1	M	TM	DP	M
20	3.06	2.4	4	1	3	DP	DP	M	TM
21	3.08	2.7	5	1	2	TM	M	TM	DP
22	3.08	3	3	1	3	DP	DP	M	TM
23	3.1	3	2	5	1	M	TM	DP	M
24	3.1	1.3	4	3	2	TM	M	TM	DP
25	3.11	1.2	6	1	2	TM	M	TM	DP
26	3.12	1.5	3	0	3	DP	DP	M	TM
27	3.12	2	4	1	3	DP	DP	M	TM
28	3.13	2	4	0	1	M	TM	DP	M
29	3.13	2	1	5	2	TM	M	TM	DP
30	3.15	1	7	1	2	TM	M	TM	DP
31	3.15	3.3	2	5	1	M	TM	DP	M
32	3.15	0.6	6	1	2	TM	M	TM	DP
33	3.16	2.3	6	3	2	TM	M	TM	DP
34	3.17	2	3	1	3	DP	DP	M	TM
35	3.22	3.3	4	3	3	DP	DP	M	TM
36	3.19	1.3	4	5	1	M	TM	DP	M
37	3.21	1.5	3	1	3	DP	DP	M	TM

ISSN:2442-5567 | E-ISSN:2443-289X

38	3.21	3.3	3	0	3	DP	DP	M	TM
39	3.22	1.8	5	3	2	TM	M	TM	DP
40	3.23	0.5	2	3	1	M	TM	DP	M
41	3.24	0.7	6	1	2	TM	M	TM	DP
42	3.25	1.8	7	3	2	TM	M	TM	DP
43	3.25	0.9	3	3	1	M	TM	DP	M
44	3.27	0.5	2	5	1	M	TM	DP	M
45	3.29	0.9	2	5	1	M	TM	DP	M
46	3.3	1	3	1	3	DP	DP	M	TM
47	3.32	3.8	3	1	3	DP	DP	M	TM
48	3.32	2.3	3	0	3	DP	DP	M	TM
49	3.32	0.9	4	0	2	TM	M	TM	DP
50	3.34	2.4	2	1	3	DP	DP	M	TM
51	3.35	1.7	5	1	2	TM	M	TM	DP
52	3.35	3	2	1	3	DP	DP	M	TM
53	3.35	3	3	1	3	DP	DP	M	TM
54	3.37	2.5	4	0	3	DP	DP	M	TM
55	3.39	2	3	0	3	DP	DP	M	TM
56	3.39	3.5	3	0	3	DP	DP	M	TM
57	3.4	2	4	1	3	DP	DP	M	TM
58	3.32	1	4	1	2	TM	M	TM	DP
59	3.23	2	3	3	2	TM	M	TM	DP
60	3.33	0.4	3	3	1	M	TM	DP	M
61	3.42	3.5	4	1	3	DP	DP	M	TM
62	3.43	3.5	4	1	3	DP	DP	M	TM
63	3.42	2.2	4	1	3	DP	DP	M	TM
64	3.45	1.2	2	0	3	DP	DP	M	TM
65	3.45	0.5	1	3	1	M	TM	DP	M
66	3.47	3	2	0	3	DP	DP	M	TM
67	3.53	3	2	1	3	DP	DP	M	TM
68	3.51	3.2	1	5	1	M	TM	DP	M
69	3.53	3	3	1	3	DP	DP	M	TM
70	3.55	2	3	5	1	M	TM	DP	M
71	3.59	3	2	5	1	M	TM	DP	M
72	3.59	1.6	4	1	2	TM	M	TM	DP
73	3.63	2	2	3	1	M	TM	DP	M
74	3.67	1	2	3	1	M	TM	DP	M
75	3.67	0.4	1	5	1	M	TM	DP	M

Keterangan :
M : menerima.
DP : dipertimbangkan.

TM : tidak menerima.

Berdasarkan tabel prioritas di atas terdapat 3 *cluster* yaitu menerima, dipertimbangkan dan tidak menerima yang di hasilkan dari tabel kalsifikasi.

4. KESIMPULAN DAN SARAN

4.1 KESIMPULAN

Berdasarkan pembahasan maka dapat diberikan beberapa simpulan sebagai berikut:

1. Algoritma *Fuzzy C-means* dapat digunakan untuk mengelompokan data mahasiswa lebih halus dengan menerap- kan besarnya derajat keanggotaan setiap elemen untuk masuk ke dalam kelom- pok-kelompok yang ada.
2. Pengujian data dilakukan sebanyak 37 iterasi, diperoleh tiga kelompok berdasarkan nilai rata-rata penentuan beasiswa, yaitu :
   a. Kelompok pertama (*cluster* ke-1), berisi mahasiswa yang memiliki rata-rata ipk sekitar 3.2649; memiliki rata-rata pendapatan orang tua sekitar 1.8625; tanggungan orang tua 2.2484 dan memiliki prestasi 4.2701.
   b. Kelompok kedua (*cluster* ke-2), berisi mahasiswa yang memiliki rata-rata ipk sekitar 3.1653; memiliki rata-rata pendapatan orang tua sekitar 1.6271; tanggungan orang tua 5.1550 dan memiliki prestasi 1.1770.
   c. Kelompok ketiga (*cluster* ke-3), berisi mahasiswa yang memiliki rata-rata ipk sekitar 3.2225; memiliki rata-rata pendapatan orang tua sekitar 2.4133; tanggungan orang tua 3.0759 dan memiliki prestasi 0.7967.
3. Kemudian setiap *cluster* diklasifikasikan berdasarkan kriteria mana yang lebih diprioritaskan dengan nilai terbesar pada jarak akhir merupakan *cluster* yang menerima beasiswa, sedangkan *cluster* dengan nilai terkecil merupakan *cluster* yang tidak berhak menerima beasiswa.

4.2 SARAN

Saran yang dapat digunakan untuk mendapat tingkat akurasi yang lebih baik dalam penerapan algoritma *Fuzzy C-Means* untuk penentuan beasiswa di Univeristas Megou Pak Tulang Bawang, serta untuk memberikan konstribusi yang lebih besar di dunia riset, disarankan agar hasil penelitian ini dikembangkan dengan cara memodifikasi/*updating* algoritma *Fuzzy C-means* yang digunakan saat ini, atau dengan menggabungkan algoritma *Fuzzy C-means* dengan algoritma lain.

REFERENSI

[1] Ahmad Mauliyadi M, dkk. (2013) , " Perbandingan Metode *Fuzzy C-Means* (Fcm) Dan *Fuzzy Gustafson-Kessel* (Fgk) Menggunakan Data Citra Satelit *Quickbird", Matematika 00 (2013) 01–05,*

[2] Ali Keshavarzi, Fereydoon Sarmadian (2012)," Fuzzy Clustering Analysis For Modeling Of Soil Cation Exchange Capacity", AJAE 3(1):27-33 (2012), ISSN:1836-9448.

[3] Apriansyah Putra, Dinna Yunika Hardiyanti (2011) "Penentuan Penerima Beasiswa Dengan Menggunakan *Fuzzy Multiple Atribute Decission Makin ",Vol. 3, No. 1, April 2011, Halaman 286-293.*

ISSN:2442-5567 | E-ISSN:2443-289X

[4] Bahar, Vincent Suhartono, dkk (2011), "Penentuan Jurusan Sekolah Menengah Atas Dengan Algoritma Fuzzy C-Means", Jurnal Teknologi Informasi, Issn 1414-9999.

[5] Dimas Wahyu Wibowo, dkk (2013) " Perhitungan Jumlah dan Jenis Kendaraan Menggunakan Metode Fuzzy C-means dan Segmentasi Deteksi Tepi Canny" , *Vol. 7, No. 2, Desember 2013, Halaman 103-110.*

[6] Eko Sediyono, dkk (2006), " Penentuan Lokasi Fasilitas Gudang Menggunakan *Fuzzy C – Means* (FCM)", Vol.2, No. 2, Desember 2006: 155 – 166.

[7] Ivan Hardiyanto, dkk (2012), " Implementasi Segmentasi Citra Dengan Menggunakan Metode Generalized Fuzzy Cmeans Clustering Algorithm With Improved Fuzzy Partitions", *Vol. 1, No. 1, (2012) 1-5*

[8] Heru Dibyo Laksono, Abraham Arief, (2012)," Perkiraan Kebutuhan Energi Listrik Jangka Panjang Di Provinsi Sumatera Barat Sampai Tahun 2021 Dengan Logika Fuzzy Clustering ",Vol. 5 No. 2 September 2012, Issn : 2086 – 4981.

[9] Kusumadewi, S., Purnomo, H., (2010), *Aplikasi Fuzzy Untuk Pendukung Keputusan*, Graha Ilmu, Jakarta

[10] Kusrini, (2006), Algoritma Data Mining, Penerbit ANDI, Yogyakarta.

[11] Noor Fitriana Hastuti, (2010), Pemanfaatan Metode *K-Means Clustering* dalam Penentuan Penerima Beasiswa, Jurusan Informatika. Fakultas MIPA. Universitas Sebelas Maret.

[12] Prabowo Pudjo Widodo, dkk (2013), " Penerapan *Data Mining* dengan *Matlab",* Penerbit Rekayasa Sains, Bandung.

[13] Turban, Efraim, et al . (2005). Decision Support Systems and Intelligent Systems. Yogyakarta : Penerbit Andi

FUZZY C-MEANS UNTUK CLUSTERING DATA (STUDI KASUS : DATA PERFORMANCE MENGAJAR DOSEN)

Emha Taufiq Luthfi
STMIK AMIKOM Yogyakarta
e-mail : emha_tl@yahoo.com

Abstraksi

Clustering merupakan proses pengelompokan data dalam kelas-kelas atau cluster-cluster sehingga data dalam suatu cluster memiliki tingkat persamaan yang tinggi satu dengan lainnya tetapi sangat berbeda dengan data dalam cluster lain. Dalam tulisan ini dilakukan percobaan penggunaan metode Fuzzy C-Means untuk mengetahui kemungkinan adanya cluster-cluster dari data performance mengajar dosen.

Kata kunci : Clustering, Fuzzy C-Means, Data Performance Mengajar Dosen

1. PENDAHULUAN

Untuk evaluasi proses pengajaran oleh dosen selama satu semester, salah satu cara dengan meminta mahasiswa untuk memberikan penilaian terhadap performance dosen dalam mengampu suatu mata kuliah selalu satu semester dalam beberapa kriteria penilaian. Kriteria yang digunakan antara lain :

- N1 : Penguasaan dan kemampuan dalam menjelaskan materi
- N2 : Kemampuan dalam menjawab pertanyaan
- N3 : Kemampuan dalam memberi motivas mahasiswa
- N4 : kemampuan membuat suasana kelas menyenangkan
- N5 : Kedisiplinan hadir dalam perkuliahan

Dari nilai kelima kriteria tersebut, didapat nilai indeks prestasi kumulatif dosen dalam mengampu suatu mata kuliah selama satu semester.

Seorang dosen dalam satu semester mungkin harus mengampu lebih dari satu mata kuliah. Atau satu mata kuliah mungkin harus diampu oleh lebih dari seorang dosen.

Proses clustering yang dilakukan, akan mencoba mengetahui kemungkinan adanya kelompok-kelompok yang mungkin belum diketahui dari data performance mengajar dosen terhadap suatu mata kuliah.

2. FUZZY C-MEANS

Fuzzy C-means Clustering (FCM), atau dikenal juga sebagai Fuzzy ISODATA, merupakan salah satu metode clustering yang merupakan bagian dari metode Hard K-Means. FCM menggunakan model pengelompokan fuzzy sehingga data dapat menjadi anggota dari semua kelas atau cluster terbentuk dengan derajat atau tingkat keanggotaan yang berbeda antara 0 hingga 1.

Tingkat keberadaan data dalam suatu kelas atau cluster ditentukan oleh derajat keanggotaannya. Teknik ini pertama kali diperkenalkan oleh Jim Bezdek pada tahun 1981.

Konsep dasar FCM, pertama kali adalah menentukan pusat cluster yang akan menandai lokasi rata-rata untuk tiap-tiap cluster. Pada kondisi awal, pusat cluster ini masih belum akurat. Tiap-tiap data memiliki derajat keanggotaan untuk tiap-tiap cluster. Dengan cara memperbaiki pusat cluster dan nilai keanggotaan tiap-tiap data secara berulang, maka dapat dilihat bahwa pusat cluster akan menujui lokasi yang tepat. Perulangan ini didasarkan pada minimasi fungsi obyektif (Gelley, 2000).

Fungsi obyektif yang digunakan FCM adalah (Ross, 2005) :

$$J(U,V;X) = \sum_{k=1}^{n}\sum_{i=1}^{C}(\mu_{ik})^{w}(d_{ik})^{2}$$

dengan w є [1,∞] ,

$$d_{ik} = d(x_k - v_i) = \left[\sum_{j=1}^{m}(x_{kj} - v_{ij})\right]^{1/2}$$

X adalah data yang dicluster :

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}$$

dan V adalah matriks pusat cluster :

$$V = \begin{bmatrix} v_{11} & \cdots & v_{1m} \\ \vdots & & \vdots \\ v_{c1} & \cdots & v_{cm} \end{bmatrix}$$

Nilai J_w terkecil adalah yang terbaik, sehingga :

$$J_w^*(U^*,V^*;X) = \min_{M_{fc}} J(U,V,X)$$

Algoritma FCM secara lengkap diberikan sebagai berikut (Zimmerman, 1991); (Yan, 1994); (Ross, 2005) :

1. Tentukan :
   a. Matriks X berukuran n x m, dengan n = jumlah data yang akan di cluster; dan m = jumlah variabel (kriteria).
   b. Jumlah cluster yang akan dibentuk (C ≥ 2)
   c. Pangkat (pembobot w > 1)
   d. Maksimum iterasi
   e. Kriteria penghentian (ε = nilai positif yang sangat kecil)
2. Bentuk matriks partisi awal U (derajat keanggotaan dalam cluster); matriks partisi awal biasanya dibuat secara acak

$$U = \begin{bmatrix} \mu_{11}(x_1) & \mu_{12}(x_2) & \cdots & \mu_{1n}(x_n) \\ \mu_{21}(x_1) & \mu_{22}(x_2) & \cdots & \mu_{2n}(x_n) \\ \vdots & & & \vdots \\ \mu_{c1}(x_1) & \mu_{c2}(x_2) & \cdots & \mu_{cn}(x_n) \end{bmatrix}$$

3. Hitung pusat cluster V untuk setiap cluster

$$V_{ij} = \frac{\sum_{k=1}^{n} (\mu_{ik})^w \cdot x_{kj}}{\sum_{k=1}^{n} (\mu_{ik})^w}$$

4. Perbaiki derajat keanggotaan setiap data pada setiap cluster (perbaiki matriks partisi)

$$\mu_{ik} = \left[\sum_{j=1}^{C} \left(\frac{d_{ik}}{d_{jk}} \right)^{2/(w-1)} \right]^{-1}$$

dengan :

$$d_{ik} = d(x_k - v_i) = \left[\sum_{j=1}^{m} (x_{kj} - v_{ij}) \right]^{1/2}$$

5. Tentukan kriteria penghentian iterasi, yaitu perubahan matriks partisi pada iterasi sekarang dan iterasi sebelumnya

$$\Delta = \left\| U^t - U^{t-1} \right\|$$

Apabila $\Delta < \varepsilon$ maka iterasi dihentikan.

3. PERCOBAAN

Clustering dengan metode FCM terhadap data performance mengajar dosen di suatu semester dilakukan untuk mengetahui cluster-cluster yang mungkin ada dan memiliki pola tertentu dengan parameter sebagai berikut :

1. Matriks input satu dimensi berupa nilai indeks prestasi mengajar seorang dosen pada mata kuliah tertentu pada program studi tertentu.
2. Jumlah cluster yang dicoba dibentuk = C = 4
3. Pangkat (pembobot) = w = 2
4. Maksimum Iterasi = 1000
5. Kriteria penghentian = $\varepsilon = 10^{-6}$

Implementasi dilakukan dengan menggunakan software matlab.

4. HASIL

Penghitungan pusat cluster dan perbaikan nilai keanggotaan akan dihentikan pada iterasi ke 42. Pada iterasi tersebut, nilai Δ = 7.8161e-007.

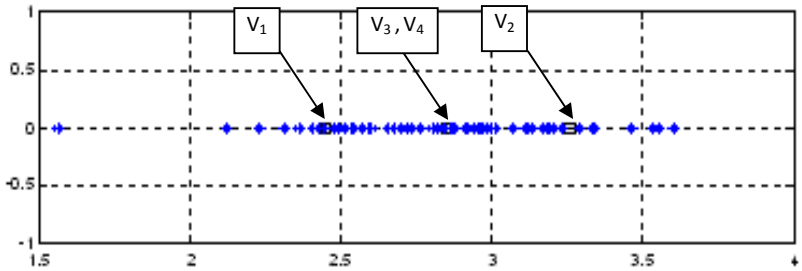
Nilai Pusat cluster adalah V = [2.4485 3.2564 2.8508 2.8508]

Nilai pusat cluster 3 dan 4 memiliki nilai sama, sehingga terbentuk hanya 3 cluster. Pada Percobaan lain mungkin didapat hasil yang sedikit berbeda, dikarenakan inisialisasi awal matriks partisi yang dilakukan secara acak.

Mata Kuliah	Dosen	Jenjang	IP_MK	Kecenderungan Cluster			
				c1	c2	c3	c4
Akuntansi I	I	D3	**2.87**	0.001	0.0012	**0.4989**	**0.4989**
Akuntansi I	SDMM	S1 REG	**2.41**	**0.9829**	0.002	0.0075	0.0075
Aljabar Linier	E	D3	**2.68**	0.2068	0.0333	**0.3799**	**0.3799**
Aljabar Linier	RJ	D3	**3.20**	0.0053	**0.9454**	0.0247	0.0247
Aljabar Linier	SM	S1 REG	**2.76**	0.0401	0.0158	**0.4721**	**0.4721**
Analisis Kinerja Sistem	MDI	S1 TR	**1.56**	**0.45**	0.1235	0.2133	0.2133
Analisis Kinerja Sistem	MDI	S1 REG	**2.12**	**0.6721**	0.0562	0.1359	0.1359
Analisis Laporan Keuangan	SH	S1 REG	**2.80**	0.0103	0.0061	**0.4918**	**0.4918**
Analisis Laporan Keuangan	SH	D3	**2.87**	0.001	0.0012	**0.4989**	**0.4989**
Analisis Sistem Informasi	HAF	D3	**3.01**	0.0322	0.1672	**0.4003**	**0.4003**
Analisis Sistem Informasi	MHT	S1 REG	**3.11**	0.029	**0.5928**	0.1891	0.1891
Analisis Sistem Informasi	SMT	S1 REG	**2.59**	**0.6122**	0.0276	0.1801	0.1801
Analisis Sistem Informasi	WW	D3	**2.94**	0.0156	0.0376	**0.4734**	**0.4734**
Bahasa Inggris II	TS	D3	**3.55**	0.0499	**0.7024**	0.1238	0.1238
Bahasa Inggris II	TS	S1 REG	**3.60**	0.059	**0.6624**	0.1393	0.1393
Broadcasting TV	MSMM	S1 TR	**2.86**	0.0003	0.0003	**0.4997**	**0.4997**
Broadcasting TV	MSMM	S1 REG	**2.91**	0.0081	0.0143	**0.4888**	**0.4888**
Broadcasting TV	MSMM	S1 REG	**2.98**	0.026	0.096	**0.439**	**0.439**
E-Business	MSMM	S1 TR	**2.97**	0.0235	0.0779	**0.4493**	**0.4493**
E-Commerce	J	S1 REG	**2.49**	**0.9715**	0.0028	0.0128	0.0128
E-Commerce	J	D3	**2.72**	0.1013	0.0259	**0.4364**	**0.4364**
Interaksi Manusia dan Komputer	PSM	S1 TR	**3.12**	0.0265	**0.6432**	0.1651	0.1651
Interaksi Manusia dan Komputer	PSM	S1 REG	**3.29**	0.0016	**0.9869**	0.0058	0.0058
Jaringan Komputer	AP	D3	**2.51**	**0.933**	0.0063	0.0303	0.0303
Jaringan Komputer	MS	S1 REG	**3.46**	0.0321	**0.7912**	0.0884	0.0884
Keamanan Komputer	MD	S1 TR	**3.00**	0.0304	0.1404	**0.4146**	**0.4146**
Kepemimpinan	ASSH	D3	**2.66**	0.2791	0.0351	**0.3429**	**0.3429**
Kepemimpinan	MI	D3	**2.82**	0.0034	0.0025	**0.4971**	**0.4971**
Kepemimpinan	SPDMM	S1 REG	**2.23**	**0.7733**	0.0351	0.0958	0.0958
Kewirausahaan	AMT	S1 REG	**2.36**	**0.9304**	0.0091	0.0303	0.0303
Kewirausahaan	AMT	S1 REG	**2.73**	0.0823	0.0235	**0.4471**	**0.4471**
Kewirausahaan	AMT	S1 REG	**2.76**	0.0401	0.0158	**0.4721**	**0.4721**

Mata Kuliah	Dosen	Jenjang	IP_MK	Kecenderungan Cluster			
				c1	c2	c3	c4
Kewirausahaan	KP	D3	**2.66**	0.2791	0.0351	**0.3429**	**0.3429**
Kewirausahaan	KP	D3	**2.70**	0.1478	0.0302	**0.411**	**0.411**
Komputer Grafis	AFS	S1 REG	**2.70**	0.1478	0.0302	**0.411**	**0.411**
Komputer Grafis	AY	D3	**2.87**	0.001	0.0012	**0.4989**	**0.4989**
Komunikasi Data	K	D3	**2.88**	0.0023	0.003	**0.4974**	**0.4974**
Komunikasi Data	K	S1 REG	**2.95**	0.0183	0.0489	**0.4664**	**0.4664**
Komunikasi Data	RMAKR	D3	**2.76**	0.0401	0.0158	**0.4721**	**0.4721**
Komunikasi Data	RMAKR	S1 REG	**3.01**	0.0322	0.1672	**0.4003**	**0.4003**
Manajemen Stratejik	FA	S1 TR	**3.19**	0.0074	**0.922**	0.0353	0.0353
Matematika Diskret	SM	S1 TR	**2.60**	**0.5609**	0.0299	0.2046	0.2046
Metodelogi Penelitian	HS	D3	**3.07**	0.0355	**0.3943**	0.2851	0.2851
Metodelogi Penelitian	ES	S1 REG	**2.43**	**0.9956**	0.0005	0.0019	0.0019
Pemp. Basis Data	AA	D3	**2.98**	0.026	0.096	**0.439**	**0.439**
Pemp. Basis Data	AS	D3	**2.76**	0.0401	0.0158	**0.4721**	**0.4721**
Pemp. Basis Data	ETL	S1 REG	**2.87**	0.001	0.0012	**0.4989**	**0.4989**
Pemp. Basis Data	MRA	S1 REG	**2.57**	**0.7115**	0.0223	0.1331	0.1331
Pemp. Berorientasi Objek II	AB	D3	**2.44**	**0.999**	0.0001	0.0004	0.0004
Pemp. Berorientasi Objek II	ASMT	D3	**2.53**	**0.876**	0.011	0.0565	0.0565
Pemp. Berorientasi Objek II	WW	D3	**3.00**	0.0304	0.1404	**0.4146**	**0.4146**
Pemp. Berorientasi Obyek II	KMT	S1 REG	**2.53**	**0.876**	0.011	0.0565	0.0565
Pemp. Terstruktur	AA	S1 REG	**2.92**	0.0104	0.0205	**0.4845**	**0.4845**
Pemp. Terstruktur	EHS	D3	**3.53**	0.0461	**0.7202**	0.1169	0.1169
Pemp. Terstruktur	HP	S1 REG	**2.54**	**0.8407**	0.0137	0.0728	0.0728
Pengelolaan Instalasi Komputer	RW	D3	**3.19**	0.0074	**0.922**	0.0353	0.0353
Pengelolaan Instalasi Komputer	MHP	S1 REG	**2.91**	0.0081	0.0143	**0.4888**	**0.4888**
Pengelolaan Proyek SI	MHT	S1 TR	**3.18**	0.0098	**0.894**	0.0481	0.0481
Pengelolaan Proyek SI	MHT	S1 REG	**3.07**	0.0355	**0.3943**	0.2851	0.2851
Perancangan Multimedia	MSMM	D3	**2.96**	0.0209	0.0623	**0.4584**	**0.4584**
Perancangan Multimedia	MSMM	S1 REG	**3.23**	0.0011	**0.9893**	0.0048	0.0048
Perancangan Multimedia	TI	D3	**2.84**	0.0004	0.0003	**0.4996**	**0.4996**
Perancangan Multimedia	TI	S1 REG	**2.95**	0.0183	0.0489	**0.4664**	**0.4664**
Perilaku Organisasi	AR	S1 REG	**3.18**	0.0098	**0.894**	0.0481	0.0481
Perilaku Organisasi	BS	D3	**3.34**	0.0082	**0.937**	0.0274	0.0274
Rekayasa Perangkat Lunak	SP	S1 TR	**2.48**	**0.9842**	0.0016	0.0071	0.0071

Mata Kuliah	Dosen	Jenjang	IP_MK	Kecenderungan Cluster			
				c1	c2	c3	c4
Sistem Informasi Akuntansi	SH	S1 REG	**2.70**	0.1478	0.0302	**0.411**	**0.411**
Sistem Informasi Akuntansi	SH	D3	**2.88**	0.0023	0.003	**0.4974**	**0.4974**
Sistem Pakar	HAF	S1 REG	**3.17**	0.0124	**0.8614**	0.0631	0.0631
Sistem Pakar	HAF	S1 TR	**3.33**	0.0066	**0.9486**	0.0224	0.0224
Sistem Penunjang Keputusan	MD	S1 REG	**3.18**	0.0098	**0.894**	0.0481	0.0481
Sistem Penunjang Keputusan	SP	S1 TR	**2.31**	**0.8675**	0.0186	0.0569	0.0569
Sistem Penunjang Keputusan	SP	S1 REG	**2.54**	**0.8407**	0.0137	0.0728	0.0728
Statistik	MCM	D3	**3.13**	0.0238	**0.6924**	0.1419	0.1419
Statistik	S	S1 REG	**2.96**	0.0209	0.0623	**0.4584**	**0.4584**
Statistik	SS	D3	**2.72**	0.1013	0.0259	**0.4364**	**0.4364**
Statistik	ZM	S1 REG	**2.57**	**0.7115**	0.0223	0.1331	0.1331
Testing dan Implementasi Sistem	SR	S1 REG	**2.66**	0.2791	0.0351	**0.3429**	**0.3429**
Testing dan Implementasi Sistem	SR	S1 TR	**2.82**	0.0034	0.0025	**0.4971**	**0.4971**



Gambar 1 Hasil FCM pada Iterasi akhir

Cluster c1 didapat untuk nilai matriks input (IPK) antara 1.56 sampai dengan 2.60. Cluster c2 didapat untuk nilai matriks input (IPK) antara 3.07 sampai dengan 3.60. Sedangkan cluster c3 dan c4 untuk nilai matriks input (IPK) 2.66 sampai dengan 3.01.

Contoh hal yang dapat dilihat misal, untuk mata kuliah Pemrograman Basis Data yang diampu oleh 4 orang dosen memiliki kecenderungan berada pada cluster c3,c4 dengan derajat keanggotaan yang rendah serta 1 dosen berada pada cluster c1 dengan derajat keanggotaan cukup tinggi.

Begitu pula mata kuliah Pemrograman Berorientasi Obyek 2, untuk 4 orang dosen pengampu memiliki kecenderungan berada pada cluster c3,c4 serta cluster c1.

Atau mungkin dapat diamati mata kuliah atau dosen yang memiliki kecenderungan berada di cluster c2.

Untuk percobaan lebih lanjut, perlu dicoba clustering dengan lebih dari 1 variabel input. Semisal, untuk kasus data performance dosen dapat dilakukan clustering dengan kombinasi input kriteria penilaian N1, N2, N3, N4 dan N5.

5. KESIMPULAN

Clustering merupakan proses pengelompokan obyek atau data tidak berlabel kedalam suatu kelas atau cluster dengan obyek yang memiliki kesamaan. Clustering dengan menggunakan metode Fuzzy C-Means terhadap data performance mengajar dosen, dapat memunculkan beberapa cluster data yang dapat dianalisa lebih lanjut persamaan dan perbedaannya.

6. DAFTAR PUSTAKA

Lin, Ching-Teng; Lee, George. 1996. *Neural Fuzzy Systems : A Neuro-Fuzzy Synergism to Intelligent Systems*. United States of America. Prentice Hall International Inc.

Kusumadewi, Sri; Hartati, Sri. 2006. *Neuro-Fuzzy : Integrasi Sistem Fuzzy dan Jaringan Syaraf.* Yogyakarta. Graha Ilmu.

http://www.cs.bris.ac.uk/home/tr1690/documentation/fuzzy_clustering_initial_report/node11.html

Comparative Analysis of K-Means and Fuzzy C-Means Algorithms

Soumi Ghosh
Department of Computer Science and Engineering,
Amity University, Uttar Pradesh
Noida, India

Sanjay Kumar Dubey
Department of Computer Science and Engineering,
Amity University, Uttar Pradesh
Noida, India

***Abstract*—In the arena of software, data mining technology has been considered as useful means for identifying patterns and trends of large volume of data. This approach is basically used to extract the unknown pattern from the large set of data for business as well as real time applications. It is a computational intelligence discipline which has emerged as a valuable tool for data analysis, new knowledge discovery and autonomous decision making. The raw, unlabeled data from the large volume of dataset can be classified initially in an unsupervised fashion by using cluster analysis i.e. clustering the assignment of a set of observations into clusters so that observations in the same cluster may be in some sense be treated as similar. The outcome of the clustering process and efficiency of its domain application are generally determined through algorithms. There are various algorithms which are used to solve this problem. In this research work two important clustering algorithms namely centroid based K-Means and representative object based FCM (Fuzzy C-Means) clustering algorithms are compared. These algorithms are applied and performance is evaluated on the basis of the efficiency of clustering output. The numbers of data points as well as the number of clusters are the factors upon which the behaviour patterns of both the algorithms are analyzed. FCM produces close results to K-Means clustering but it still requires more computation time than K-Means clustering.**

Keywords—clustering; k-means; fuzzy c-means; time complexity

I. INTRODUCTION

In the field of software data analysis is considered as a very useful and important tool as the task of processing large volume of data is rather tough and it has accelerated the interest of application of such analysis. To be precise data mining is the analysis of datasets that are observational, aiming at finding out unsuspected relationships among datasets and summarizing the data in such a noble fashion that are both understandable and useful to the data users [9].

It also makes data description possible by means of clustering visualization, association and sequential analysis. Data clustering is primarily a method of data description which is used as a common technique for data analysis in various fields like machine learning, data mining, pattern recognization, image analysis and bio-informatics. Cluster analysis is also recognised as an important technique for classifying data, finding clusters of a dataset based on similarities in the same cluster and dissimilarities between different clusters [13]. Putting each point of the dataset to exactly one cluster is the basic of the conventional clustering method where as clustering algorithm actually partitions unlabeled set of data into different groups according to the similarity. As compare to data classification, data clustering is considered as an unsupervised learning process which does not require any labelled dataset as training data and the performance of data clustering algorithm is generally considered as much poorer. Although data classification is better performance oriented but it requires a labelled dataset as training data and practically classification of labelled data is generally very difficult as well as expensive. As such there are many algorithms that are proposed to improve the clustering performance. Clustering is basically considered as classification of similar objects or in other words, it is precisely partitioning of datasets into clusters so that data in each cluster shares some common trait. The hierarchical, partitioning and mixture model methods are the three major types of clustering processes that are applied for organising data. The choice of application of a particular method generally depends on the type of output desired, the known performance of the method with particular type of data, available hardware and software facilities and size of the dataset [13].

In this research paper, K-Means and Fuzzy C-Means clustering algorithms are analyzed based on their clustering efficiency.

II. K-MEANS CLUSTERING

K-Means or Hard C-Means clustering is basically a partitioning method applied to analyze data and treats observations of the data as objects based on locations and distance between various input data points. Partitioning the objects into mutually exclusive clusters (K) is done by it in such a fashion that objects within each cluster remain as close as possible to each other but as far as possible from objects in other clusters.

Each cluster is characterized by its centre point i.e. centroid. The distances used in clustering in most of the times do not actually represent the spatial distances. In general, the only solution to the problem of finding global minimum is exhaustive choice of starting points. But use of several replicates with random starting point leads to a solution i.e. a global solution [2, 6, 14]. In a dataset, a desired number of clusters K and a set of k initial starting points, the K-Means clustering algorithm finds the desired number of distinct clusters and their centroids. A centroid is the point whose co-

ordinates are obtained by means of computing the average of each of the co-ordinates of the points of samples assigned to the clusters.

Algorithmic steps for K-Means clustering [12]

1) Set K – To choose a number of desired clusters, K.

2) Initialization – To choose k starting points which are used as initial estimates of the cluster centroids. They are taken as the initial starting values.

3) Classification – To examine each point in the dataset and assign it to the cluster whose centroid is nearest to it.

4) Centroid calculation – When each point in the data set is assigned to a cluster, it is needed to recalculate the new k centroids.

5) Convergence criteria – The steps of (iii) and (iv) require to be repeated until no point changes its cluster assignment or until the centroids no longer move.

The actual data samples are to be collected before the application of the clustering algorithm. Priority has to be given to the features that describe each data sample in the database [3, 10]. The values of these features make up a feature vector (F_{i1}, F_{i2}, F_{i3},……., F_{im}) where F_{im} is the value of the M-dimensional space [12]. As in the other clustering algorithms, k- means requires that a distance metric between points is to be defined. This distance metric is used in the above mentioned step (iii) of the algorithm. A common distance metric is the Euclidean distance. In case, the different features used in the feature vector have different relative values and ranges then the distance computation may be distorted and so may be scaled.

The input parameters of the clustering algorithm are the number of clusters that are to be found along with the initial starting point values. When the initial starting values are given, the distance from each sample data point to each initial starting value is found using equation. Then each data point is placed in the cluster associated with the nearest starting point. After all the data points are assigned to a cluster, the new cluster centroids are calculated. For each factor in each cluster, the new centroid value is then calculated. The new centroids are then considered as the new initial starting values and steps (iii) and (iv) of the algorithm are repeated. This process continues until no more data point changes or until the centroids no longer move.

III. Fuzzy C-Means Clustering

Bezdek [5] introduced Fuzzy C-Means clustering method in 1981, extend from Hard C-Mean clustering method. FCM is an unsupervised clustering algorithm that is applied to wide range of problems connected with feature analysis, clustering and classifier design. FCM is widely applied in agricultural engineering, astronomy, chemistry, geology, image analysis, medical diagnosis, shape analysis and target recognition [16].

With the development of the fuzzy theory, the FCM clustering algorithm which is actually based on Ruspini Fuzzy clustering theory was proposed in 1980's. This algorithm is used for analysis based on distance between various input data points. The clusters are formed according to the distance between data points and the cluster centers are formed for each cluster.

Infact, FCM is a data clustering technique [11, 7] in which a data set is grouped into n clusters with every data point in the dataset related to every cluster and it will have a high degree of belonging (connection) to that cluster and another data point that lies far away from the center of a cluster which will have a low degree of belonging to that cluster.

Algorithmic steps for Fuzzy C-Means clustering [13]

We are to fix c where c is (2<=c<n) and then select a value for parameter 'm' and there after initialize the partition matrix $U^{(0)}$. Each step in this algorithm will be labelled as 'r' where r = 0, 1, 2 …

1) We are to calculate the c center vector $\{V_{ij}\}$ for each step.

$$v_{ij} = \frac{\sum_{k=1}^{n} (\mu_{ik})^m x_{kj}}{\sum_{k=1}^{n} (\mu_{ij})^m} \tag{1}$$

2) Calculate the distance matrix $D_{[c,n]}$.

$$D_{ij} = \left(\sum_{j=1}^{m} (x_{kj} - v_{ij})^2 \right)^{1/2} \tag{2}$$

3) Update the partition matrix for the rth step, $U^{(R)}$ as

$$\mu_{ij}^{r-1} = \left(1 \Big/ \sum_{j=1}^{c} (d_{ik}^{r} \Big/ d_{jk}^{r})^{2/(m-1)} \right) \tag{3}$$

If $\|U^{(k+1)}-U^{(k)}\|<\delta$ then we are to stop otherwise we have to return to step 2 by updating the cluster centers iteratively and also the membership grades for data point [13].

FCM iteratively moves the cluster centers to the right location within a dataset. To be specific introducing the fuzzy logic in K-Means clustering algorithm is the Fuzzy C-Means algorithm in general. Infact, FCM clustering techniques are based on fuzzy behaviour and they provide a technique which is natural for producing a clustering where membership weights have a natural interpretation but not probabilistic at all. This algorithm is basically similar in structure to K-Means algorithm and it also behaves in a similar fashion.

IV. Implementation Methodology

For the purpose of testing the efficiency of K-Means and FCM in matlab [8], the well known UCI Machine Learning Repository [1] is used and it is actually a collection of databases which is widely used by the researchers of Machine Learning, especially for the empirical algorithms analysis of this discipline [1]. Iris plant Dataset: Total number of attributes is five of which four (Sepal Length, Sepal Width, Petal Length and Petal Width) are numeric and one is non-numeric. This non-numeric attribute has three classes. The total numbers of instances are 150 in this attribute. The three classes are Iris Setosa, Iris Versicolour, and Iris Virginica. One class is linearly separable from the other 2, the latter are not linearly separable from each other.

A. Implementation of K-Means Clustering

The matlab function kmeans used for K-Means clustering to partitions the points in the n-by-p data matrix data into k clusters [8]. This iterative partitioning minimises the overall sum of clusters, within cluster sums of point- to cluster centroid distances. Rows of data correspond to points, columns correspond to variables and kmeans return an n-by-1 vector idx containing the cluster indices of each point. By default, k-means uses squared Euclidean distances. When data is a vector, k-means treats it as an n-by-1 data matrix, regardless of its orientation. The iris dataset for three clusters, five 'replicates' have been specified and the 'display' parameters are used to print out the final sum of distances for each of the solutions. The sum total of distances covering 13 iterations that have taken into considerations in this paper comes to 7897.88. The total **elapsed time is 0.443755 seconds.** Following scattered K-Means graph for iris data set (sepal length, sepal width and petal length) represents three clusters.

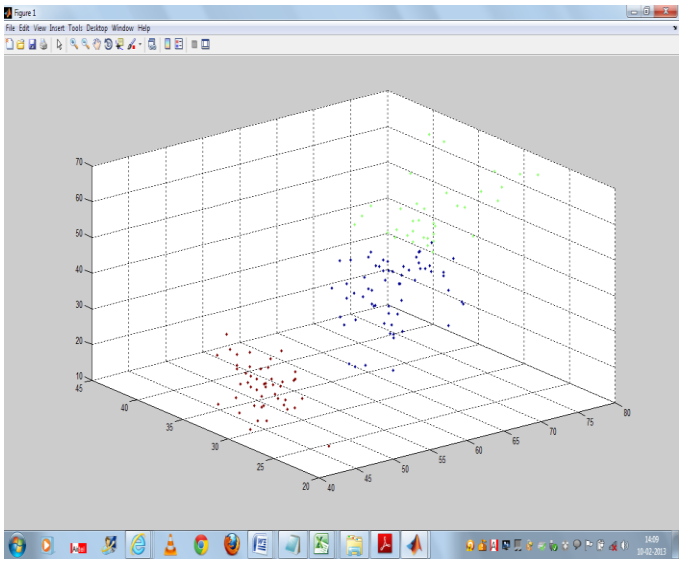


Fig.1. Scattered K-Means graph of iris dataset for three clusters

B. Implementation of Fuzzy C-Means Clustering

The mat lab function fcm performs FCM clustering [8]. The function fcm takes a data set and a desired number of clusters and returns optimal cluster centers and membership grades for each data point. It starts with an initial guess for the cluster centers, which are intended to mark the mean location of each cluster. The initial guess for these cluster centers is most likely incorrect. Next, fcm assigns every data point a membership grade for each cluster.

By iteratively updating the cluster centers and the membership grades for each data point, fcm iteratively moves the cluster centers to the right location within a data set. This iteration is based on minimizing an objective function that represents the distance from any given data point to a cluster center weighted by that data point's membership grade. The dataset is obtained from the data file 'iris.dat'[1]. From each of the three groups (setosa, versicolor and virginica), two characteristics (for example, sepal length vs. sepal width) of the flowers are plotted in a 2-dimensional plot.

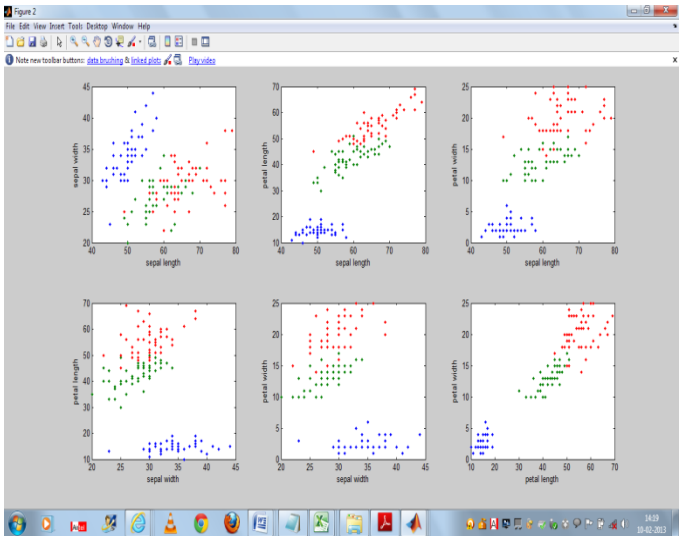


Fig.2. Scattered Fuzzy C-Means graph of iris dataset for three clusters

FCM clustering is an iterative process. The process stops when the maximum number of iterations is reached, or when the objective function improvement between two consecutive iterations is less than the minimum amount of improvement specified. For iris dataset comprising of 30 total iteration count results a total objective function equals to 6058.689983. The total **elapsed time is 0.781679 seconds.** The figure shows the initial and final fuzzy cluster centers. The bold numbers represent the final fuzzy cluster centers obtained by updating them iteratively.

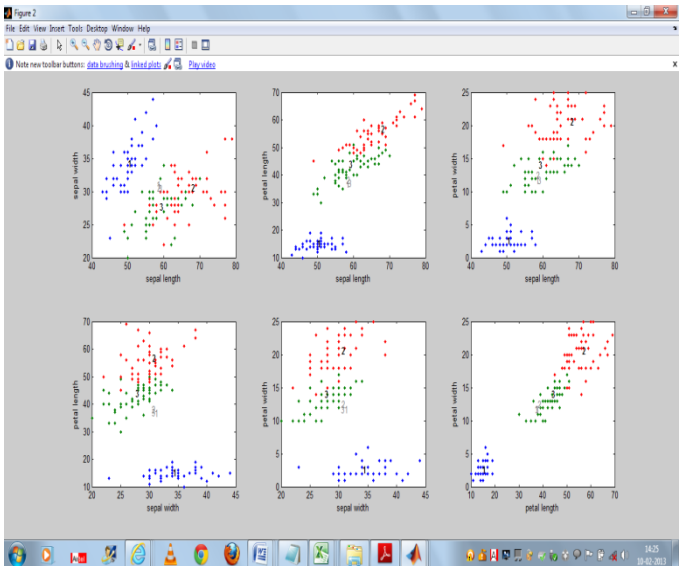


Fig.3. Scattered Fuzzy C-Means graph with initial and final fuzzy cluster centers

V. Experimental Results

This experiment reveals the fact that K-Means clustering algorithm consumes less elapsed time i.e. 0.443755 seconds than FCM clustering algorithm which takes 0.781679 seconds. On the basis of the result drawn by this experiment it may be

safely stated that K-Means clustering algorithm less time consuming than FCM algorithm and hence superior.

A. Comparison of Time Complexity of K-Means and FCM

The time complexity of K-means [15] is O(ncdi) and time complexity of FCM [4] is $O(ndc^2i)$. Keeping the number of data points constant we may assume that n = 100, d = 3, i = 20 and varying number of clusters where n = number of data points, c = number of cluster, d = number of dimension and i = number of iterations. The following table and graph represents the comparison in details.

TABLE I. COMPARATIVE ANALYSIS OF K-MEANS AND FCM

Algorithm	Time Complexity	Elapsed Time (Seconds)
K-Means	O(ncdi)	0.443755
FCM	$O(ndc^2i)$	0.781679

TABLE II. TIME COMPLEXITY OF K-MEANS AND FCM WHEN NUMBER OF CLUSTERS VARYING

S.No.	Number of Clusters	K-Means Time Complexity	FCM Time Complexity
1	1	6000	6000
2	2	12000	24000
3	3	18000	54000
4	4	24000	96000

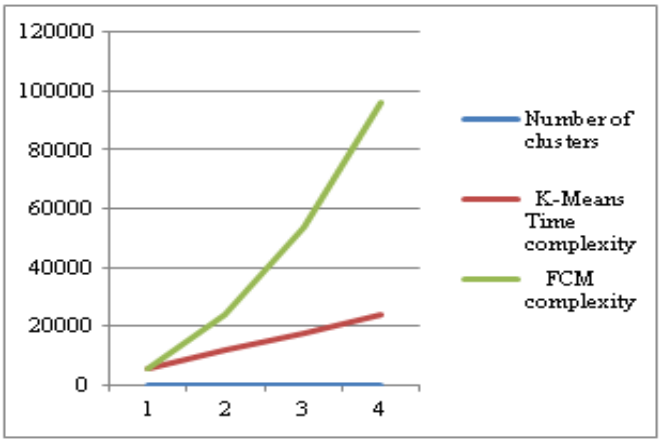


Fig.4. Time complexity of K-Means and FCM by varying number of clusters

Now keeping no. of cluster constant, lets assume n=150, d=2, c=2 and varying no. of iteration, we obtain the following table and graph.

TABLE III. TIME COMPLEXITY OF K-MEANS AND FCM WHEN NUMBER OF ITERATIONS VARYING

S.No.	Number of Iterations	K-Means Time Complexity	FCM Time Complexity
1	5	3000	6000
2	10	6000	12000
3	15	9000	18000
4	20	12000	24000

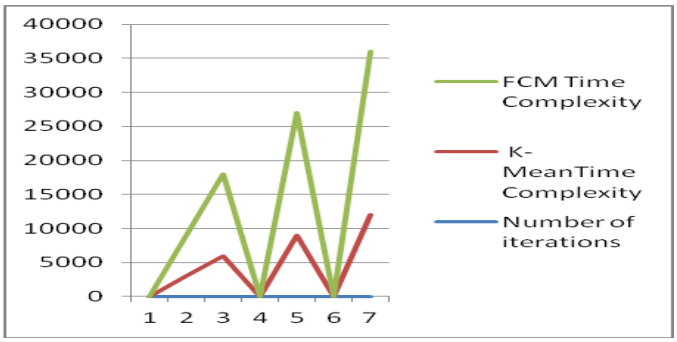


Fig.5. Time complexity of K-Means and FCM by varying number of iterations

VI. CONCLUSION

K-Means partitioning based clustering algorithm required to define the number of final cluster (k) beforehand. Such algorithms are also having problems like susceptibility to local optima, sensitivity to outliers, memory space and unknown number of iteration steps that are required to cluster. The time complexity of the K-Means algorithm is O(ncdi) and the time complexity of FCM algorithm is $O(ndc^2i)$. From the obtained results we may conclude that K-Means algorithm is better than FCM algorithm. FCM produces close results to K-Means clustering but it still requires more computation time than K-Means because of the fuzzy measures calculations involvement in the algorithm. Infact, FCM clustering which constitute the oldest component of software computing, are really suitable for handling the issues related to understand ability of patterns, incomplete/noisy data, mixed media information, human interaction and it can provide approximate solutions faster. They have been mainly used for discovering association rules and functional dependencies as well as image retrieval. So, overall conclusion is that K-Means algorithm seems to be superior than Fuzzy C-Means algorithm.

REFERENCES

[1] A. Asuncion and D. J. Newman, UCI Machine Learning Repository Irvine, CA: University of California, School of Information and Computer Science, 2013.

[2] A. K. Jain, M. N. Murty and P. J. Flynn, “Data Clustering: A review”, ACM Computing Surveys, vol. 31, no. 3, 1999.

[3] A. Rakhlin and A. Caponnetto, “Stability of K-Means clustering”, Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 2007, pp. 216–222.

[4] A. Rui and J. M. C. Sousa, "Comparison of fuzzy clustering algorithms for Classification", International Symposium on Evolving Fuzzy Systems, 2006 , pp. 112-117.

[5] J. C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms", New York: Plenum Press, 1981.

[6] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, 2nd ed., New Delhi, 2006.

[7] L. Hui, "Method of image segmentation on high-resolution image and classification for land covers", Fourth International Conference on Natural Computation, vol. 5, 2008, pp. 563-566.

[8] Mathworks. http: //www.mathworks.com

[9] R. Mosley, "The Use of Predictive Modeling in the Insurance Industry", Pinnacle actuarial resources, 2005.

[10] S. Borah and M. K. Ghose, "Performance analysis of AIM-K-Means and K-Means in quality cluster generation", Journal of Computing, vol. 1, Issue-1, 2009.

[11] S. Chen and D. Zhang, "Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure", IEEE Transactions on Systems, Man and Cybernetics, vol. 34, 1998, pp. 1907-1916,.

[12] T. Kanungo and D. M. Mount, "An Efficient K-means Clustering Algorithm: Analysis and Implementatio*n",* Pattern Analysis and Machine Intelligence, IEEE Transactions on Pattern Analysis and Machine Intelligence. vol. 24, no. 7, 2002.

[13] V. S. Rao and Dr. S. Vidyavathi, "Comparative Investigations and Performance Analysis of FCM and MFPCM Algorithms on Iris data", Indian Journal of Computer Science and Engineering, vol.1, no.2, 2010 pp. 145-151.

[14] X. Hui, J. Wu and C. Jian, "K-Means clustering versus validation measures: A data distribution perspective", IEEE Transactions on Systems, Man, and cybernetics, vol. 39, Issue-2, 2009 , pp.319-331.

[15] X. Rui, D. Wunsch II, "Survey of Clustering Algorithms", IEEE Transactions on Neural Networks, vol.16, no.3, 2005.

[16] Y. Yong, Z. Chongxun and L. Pan, "A Novel Fuzzy C-Means Clustering Algorithm for Image Thresholding", Measurement Science Review, vol. 4, no.1, 2004.

AUTHORS PROFILE

Soumi Ghosh is pursuing M. Tech (CS&E) at Amity University, Uttar Pradesh, India. Her research areas include Software Engineering and Fuzzy Logic.

Mr. Sanjay Kumar Dubey is Assistant Professor and Proctor in Amity University, Uttar Pradesh, India. He has submitted his Ph. D. thesis in Object Oriented Software Engineering. He has published more than 73 papers in International Journals. He has presented 14 research papers at various National/International conferences.He is member of IET and IEANG. His research areas include Human Computer Interaction, Soft Computing and Usability Engineering.

Kumpulan jurnaL Ilmu Komputer (KLIK)
Volume 01, No.01 September 2014
ISSN: 2406-7857

IMPLEMENTASI METODE FUZZY C-MEANS PADA SISTEM CLUSTERINGDATA VARIETAS PADI

Nurjanah[1], Andi Farmadi[2], Fatma Indriani[2]
[1,2,3] Prodi Ilmu Komputer Fakultas MIPA Universitas Lambung Mangkurat

Jl. A. Yani Km 36 Banjarbaru, Kalimantan Selatan
Email: nurjanah_abs@yahoo.co.id

Abstract

Mutations with gamma rays conducted on five local rice varieties tidal South Kalimantan produce a lot of data availability. In order for these data not only become a graveyard of useless data required a method that could be used to probe the hidden information from the data. The method known as data mining. Data mining is a technique to gain knowledge from the data by looking for certain patterns or rules of a number of large amounts of data. One method of data mining is clustering, where clustering is usually used to group objects that are similar in the same class or segment. By utilizing the data of local rice varieties tidal South Kalimantan mutated by gamma rays, data mining process is done by grouping the data based on the harvest age, productive tillers, and weight of 1000 seeds into 4 groups using fuzzy c-means algorithm. From that cluster information, carried ranking using the Simple Additive Weighting method and acquired knowledge about improved varieties by harvest age, productive tillers, and a weight of 1000 is kuatek with a dose of 30 krad.

Keywords *: Data Mining, Cluster, Fuzzy C-Means, Local Rice Varieties*

Abstrak

Mutasi dengan sinar gamma yang dilakukan terhadap lima varietas padi lokal pasang surut kalimantan selatan menghasilkan tersedianya banyak data. Agar data-data tersebut tidak hanya menjadi kuburan data yang tidak berguna dibutuhkan sebuah metode yang bisa digunakan untuk menggali informasi–informasi tersembunyi dari data tersebut. Metode tersebut dikenal dengan data mining. Data mining merupakan suatu teknik untuk menggali pengetahuan dari data dengan mencari pola atau aturan tertentu dari sejumlah data dalam jumlah besar. Salah satu metode data mining adalah klastering, dimana klastering biasanya digunakan untuk mengelompokan objek-objek yang memiliki kemiripan dalam kelas atau segmen yang sama. Dengan memanfaatkan data varietas padi hasil mutasi dengan sinar gamma dilakukan proses penggalian data dengan cara mengelompokkan data umur panen, anakan produktif, dan bobot 1000 menjadi 4 klaster menggunakan algoritma fuzzy c-means. Dari informasi 4 klaster tersebut, dilakukan perangkingan klaster menggunakan metode Simple Additive Weighting dan diperoleh pengetahuan tentang varietas unggul berdasarkan umur panen, anakan produktif, dan bobot 1000 yaitu kuatek dengan dosis 30 krad.

Kata kunci *: Data Mining, Klaster, Fuzzy C-Means, Varietas Padi Lokal*

1. PENDAHULUAN

Varietas padi lokal Kalimantan Selatan masih menjadi pilihan para petani untuk ditanam, yang dalam penanamannya tidak terlalu bergantung pada input pertanian. Selain itu, varietas padi lokal juga bersifat adaptif dengan lingkungan Kalimantan Selatan yang didominasi oleh lahan rawa, meskipun berumur panjang dan produktivitas relatif rendah daripada varietas padi unggul. Akan tetapi sifat "pera" nasi yang dihasilkan membuat harga jual varietas padi lokal lebih tinggi dibandingkan dengan varietas padi unggul [3].

Perjuangan dalam peningkatan produksi pangan untuk memenuhi kebutuhan masyarakat Kalimantan Selatan merupakan tantangan bagi masyarakat pertanian yang dapat melahirkan peluang berupa pengembangan teknologi alternatif untuk menerobos kebuntuan dan stagnasi produksi beras lokal. Alternatif itu dapat diarahkan pada pengembangan varietas lokal yang dapat beradaptasi pada lingkungan produksi yang buruk dengan produktivitas yang tinggi dan umur panen yang pendek. Peningkatan keragaman populasi dapat dilakukan salah satunya melalui induksi mutasi secara fisik dengan iradiasi sinar gamma.

Salah satu upaya untuk meningkatkan produksi pangan varietas padi lokal pasang surut kalimantan selatan yaitu dengan cara meningkatkan keragaman populasi, dilakukan radiasi dengan sinar gamma terhadap lima varietas padi. Dari upaya tersebut dihasilkan banyak data salah satunya yaitu data varietas padi hasil mutasi dengan sinar gammamaka dibutuhkan sebuah metode yang bisa digunakan untuk menggali informasi–informasi tersembunyi dari data tersebut. Metode tersebut dikenal dengan *data mining*.

Data mining adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar. Salah satu tugas dalam data mining adalah klastering. Tujuan utama dari klastering adalah pengelompokan sejumlah data/obyek ke dalam klaster sehingga dalam setiap klaster akan berisi data yang semirip mungkin [1].

Penggalian suatu data menjadi pengetahuan adalah kegiatan yang sangat berguna, karena pengetahuan tersebut bisa digunakan untuk memecahkan suatu masalah yang ada. *Data mining* merupakan salah satu teknik untuk menggali pengetahuan dari data. Agar data-data tersebut tidak menjadi kuburan data yang tidak bergunamaka dibuat sistem yang mampu menggali data varietas padi yang dimutasi dengan sinar gamma yang dapat mengorganisir data varietas padi yang dimutasi dengan sinar gamma dalam suatu database serta dapat menghasilkan pengelompokan sejumlah data/obyek ke dalam klaster yang memberikan pengetahuan mengenai varietas unggulmenggunakan metode *Fuzzy C-Means* (FCM).

2. METODOLOGI PENELITIAN

2.1. Analisa Data

Pada tahap ini akan dilakukan analisis data dengan KDD (*knowledge Discovery in Databases*). Keseluruhan proses KDD dari mulai memasukan data mengalami 3 tahap secara global sebelum menjadi hasil yang berupa informasi, yaitu:

(a) *Preprocessing*

Pada tahap ini struktur basis data akan dipersiapkan sehingga mempermudah proses mining. Proses *preprocessing* ini mencakup tiga hal utama yaitu :

1) seleksi data
   untuk memilih data yang akan digunakan dalam proses data mining. Dalam penelitian ini dipilih 3 data varietas padi yaitu umur panen, anakan produktif, dan bobot 1000 biji;
2) *Cleaning*
   Memastikan kualitas data yang telah dipilih pada tahap seleksi data, pada tahap ini masalah yang harus dihadapi adalah noisy data dan missing values. Proses cleaning data perlu dilakukan agar data bersih dari duplikasi data, data yang tidak konsisten, atau kesalahan cetak. Sehingga data yang telah melewati proses ini siap untuk di proses di data mining. Pada penelitian ini proses pembersihan data hanya dilakukan pada data yang missing value;
3) integrasi data
   Setelah melakukan pembersihan data yang menyeluruh, maka akan dilakukan tahapan selanjutnya yaitu integrasi data, karena data lebih dari satu tabel maka dilakukan tahap penggabungan data dari berbagai tabel. Kemudian data padi diberi kode agar tidak ada data yang memiliki nama yang sama. Kode padi terdiri dari nomor padi, baris padi, dosis radiasi, dan varietas padi.

(b) Data Mining

Proses data mining yang dilakukan pada penelitian ini adalah menerapkan teknik *clustering* dengan algoritma *Fuzzy C-Means* untuk mendapatkan karakteristik dari data. *Cluster* yang digunakan pada penelitian ini sebanyak empat klaster.

Konsep dasar FCM (*Fuzzy C-Means*) pertama kali adalah menentukan pusat *cluster* yang akan menandai lokasi rata-rata untuk tiap *cluster*. Pada kondisi awal, pusat *cluster* ini masih belum akurat. Tiap-tiap titik data memiliki derajat keanggotan untuk tiap-tiap *cluster*. Dengan cara memperbaiki pusat *cluster* dan derajat keanggotaan tiap-tiap titik data secara berulang, maka akan dapat dilihat bahwa pusat *cluster* akan bergerak menuju lokasi yang tepat. Perulangan ini didasarkan pada derajat keanggotaan yang menggambarkan jarak dari titik data yang diberikan ke pusat *cluster* yang terbobot oleh derajat keanggotaan titik data tersebut [6].

Pada perhitungan menggunakan FCM dengan menggunakan data yang sama tetapi diolah dengan jumlah *cluster* yang berbeda, maka hasil pengelompokannya akan sedikit berbeda, karena data tidak diolah dengan satu variabel saja tetapi dengan semua variabel. Perbedaan hasil pengelompokan itu dikarenakan data pada kelompok tertentu kemungkinan akan berpindah pada kelompok lain jika diolah dengan jumlah *cluster* yang berbeda, ini menunjukkan bahwa sistem aplikasi sudah berjalan dengan benar [6]

Output dari FCM bukan merupakan fuzzy inference sistem, namun merupakan derajat pusat *cluster* dan beberapa derajat keanggotaan untuk

tiap-tiap titik data. Algoritma *Fuzzy C-Means* disusun dengan langkah sebagai [6]:

(1) Input data
input data yang akan di*cluster*yaitu data varietas padi (X) berupa matrik berukuran n × m (n = jumlah data, m = atribut setiap data). X_{ij} = data ke- i (i = 1, 2, …, n), atribut ke- j (j = 1, 2, …, m).

(2) Batasan
(a) Jumlah *cluster* = c = 4
(b) Pangkat = w = 2
(c) Maksimum iterasi = Maxit = 100
(d) Error terkecil yang diharapkan = ξ = 10^{-3}
(e) Fungsi obyektif awal = Po = 0
(f) Iterasi awal = t = 1

(3) Membangkitkan bilangan random μ_{ik}, i = 1, 2, …, n; k = 1, 2, …,c; sebagai elemen-elemen matrik partisi awal U, dengan jumlah setiap nilai elemen kolom dalam satu baris adalah 1 (satu)

$$\sum_{i=1}^{c} \mu_{ci} = 1 \qquad \dots (1)$$

(4) Menghitung pusat *cluster* ke- k : V_{kj} dengan k = 1, 2, …, c; dan j = 1, 2, …,m.

$$V_{kj} = \frac{\sum_{i=1}^{n}\left((\mu_{ik})^{w} * X_{ij}\right)}{\sum_{i-1}^{n}(\mu_{ik})^{w}} \qquad \dots (2)$$

(5) Menghitung fungsi obyektif pada iterasi ke- t, P_t :

$$P_t = \sum_{i=1}^{n}\sum_{k=1}^{c}\left(\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right](\mu_{ik})^{w}\right) \qquad \dots (3)$$

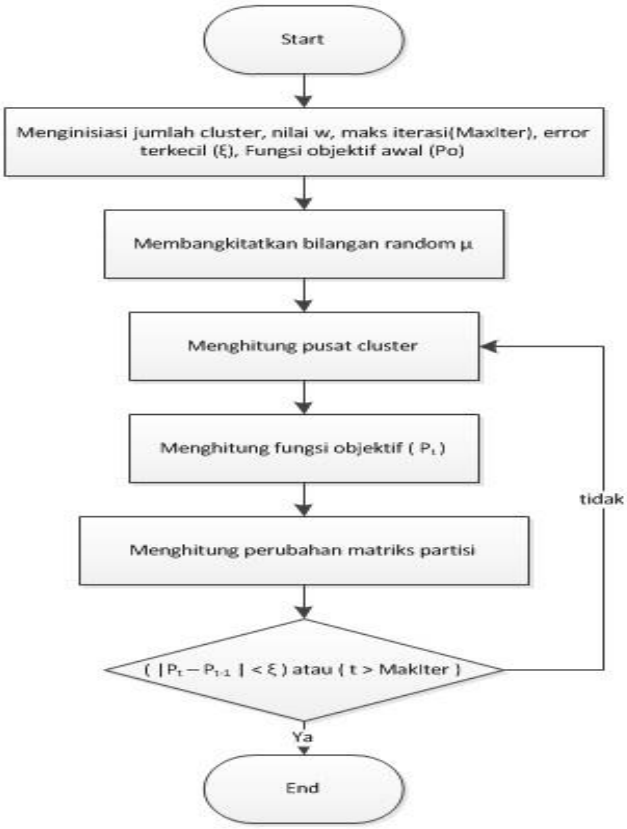
(6) Menghitung perubahan matrik partisi:

$$\mu_{ik} = \frac{\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right]^{\frac{-1}{w-1}}} \dots (4)$$

dengan i = 1, 2, …, n; dan k = 1, 2, …, c

(7) Mengecek kondisi berhenti:
1. Jika : ($|P_t - P_{t-1}| < \xi$) atau (t > MakIter) maka berhenti;
2. Jika tidak : t = t + 1, ulangi langkah ke- d (menghitung V_{kj}).

Bagan alir untuk metode *Fuzzy C-Means* yaitu sebagai berikut :



Gambar 1. Bagan Alir Metode *Fuzzy C-Means*

(c) ***Postprocessing***

Postprocessing bertujuan untuk membantu pengguna dalam memahami informasi, Karena pusat *cluster* masih tidak bisa dienterpretasikan secara langsung, maka diperlukan metode *Simple Additive Weighting* yang bisa melakukan perangkingan nilai pusat *cluster* terakhir tersebut.

Metode SAW (*Simple Additive Weighting*) sering juga dikenal istilah metode penjumlahan terbobot. Konsep dasar metode SAW adalah mencari penjumlahan terbobot dari rating kinerja pada setiap alternatif pada semua atribut. Langkah-langkah metode dalam metode SAW adalah [7] :

1. Membuat matriks keputusan Z berukuran m x n, dimana m = alternatif yang akan dipilih dan n = kriteria .
2. Memberikan nilai x setiap alternatif (i) pada setiap kriteria (j) yang sudah ditentukan, dimana, i=1,2,…m dan j=1,2,…n pada matriks keputusan Z,

$$\mathbf{Z} = \begin{vmatrix} x_{11} & x_{y12} & \dots & x_{1j} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ x_{i1} & x_{i2} & & x_{ij} \end{vmatrix} \dots (5)$$

3. Memberikan nilai bobot preferensi (W) oleh pengambil keputusan untuk masing-masing kriteria yang sudah ditentukan.

$$W = [W_1\ W_2 W_3 \dots\ W_j] \dots(6)$$

4. Melakukan normalisasi matriks keputusan Z dengan cara menghitung nilai rating kinerja ternormalisasi (r_{ij}) dari alternatif A_i pada atribut C_j

$$r_{ij} = \begin{cases} \frac{x_{ij}}{\underset{i}{MAX}(x_{ij})} & \text{jika j adalah atribut keuntungan} \\ \frac{\underset{i}{MIN}(x_{ij})}{x_{ij}} & \text{jika j adalah atribut biaya} \end{cases} \dots(7)$$

Dengan ketentuan :

a. Dikatakan atribut keuntungan apabila atribut banyak memberikan keuntungan bagi pengambil keputusan, sedangkan atribut biaya merupakan atribut yang banyak memberikan pengeluaran jika nilainya semakin besar bagi pengambil keputusan.
b. Apabila berupa atribut keuntungan maka nilai (x_{ij}) dari setiap kolom atribut dibagi dengan nilai (MAX x_{ij}) dari tiap kolom, sedangkan untuk atribut biaya, nilai (MIN x_{ij}) dari tiap kolom atribut dibagi dengan nilai (x_{ij}) setiap kolom.

5. Hasil dari nilai rating kinerja ternormalisasi (r_{ij}) membentuk matriks ternormalisasi (N)

$$\mathbf{N} = \begin{vmatrix} r_{11} & r_{y12} & \dots & r_{1j} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ r_{i1} & r_{i2} & & r_{ij} \end{vmatrix} \dots (8)$$

6. Melakukan proses perankingan dengan cara mengalikan matriks ternormalisasi (N) dengan nilai bobot preferensi (W).
7. Menentukan nilai preferensi untuk setiap alternatif (V_i) dengan cara menjumlahkan hasil kali antara matriks ternormalisasi (N) dengan nilai bobot preferensi (W).

$$V_i = \sum_{j=1}^{n} w_j r_{ij} \dots (9)$$

Nilai Vi yang lebih besar mengindikasikan bahwa alternatif Ai merupakan alternatif terbaik.

2.3. Implementasi

Implementasi dilakukan dengan metode Fuzzy C-Means (FCM) pada sistem clustering data varietas padilokal pasang surut kalimantan selatan yang dimutasi dengan sinar gamma menggunakan delphi XE2

2.4. Pengujian

Untuk memastikan apakah perhitungan dengan metode fuzzy c-means sudah benar dilakukan uji coba perhitungan secara manual meggunakan MS. Excel dan Matlab.Pengujian dilakukan untuk menghindari kesalahan dari sistem yang dibuat. Apabila terjadi kesalahan, maka sistem akan diperbaiki kembali sampai hasil proses sesuai dengan apa yang diharapkan.

3. HASIL DAN PEMBAHASAN

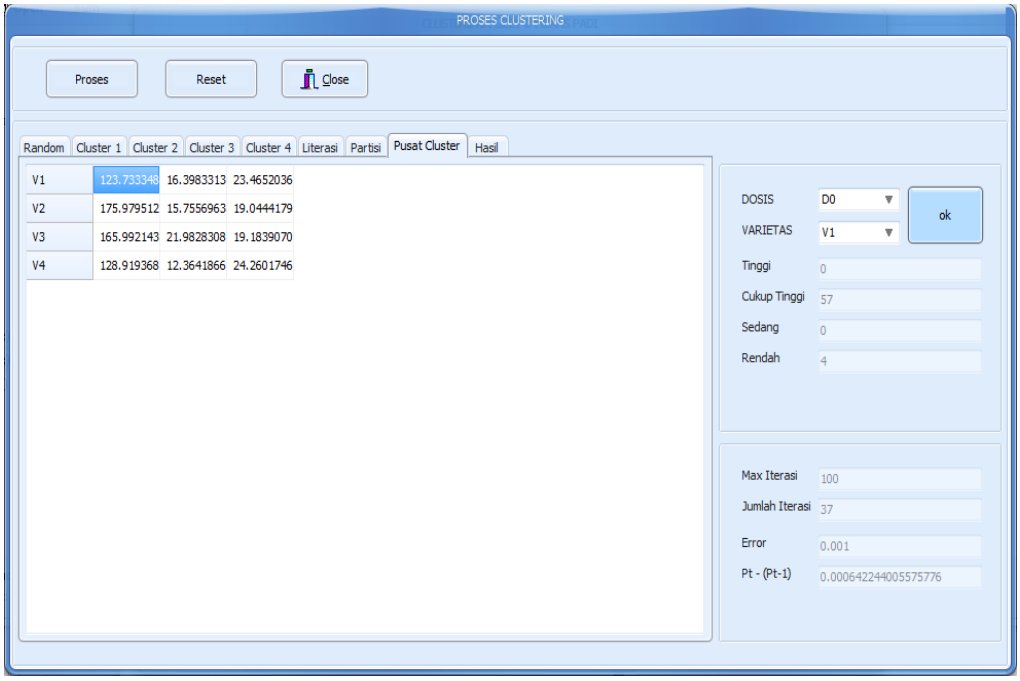
Dariproses clustering Fuzzy C-Means dengan batasan jumlah cluster sebanyak 4 cluster, pangkat 2, masimal iterasi 100, error 0,001 proses berhenti pada iterasi ke-37 dengan error 0,000642 dan diperoleh pusat cluster iterasi terakhir yaitu pada Tabel 1 atau pada Gambar 2 dengan menggunakan rumus dibawah ini (persamaan 2):

$$V_{kj} = \frac{\sum_{i=1}^{2420}\left((\mu_{ik})^2 * X_{ij}\right)}{\sum_{i=1}^{2420}(\mu_{ik})^2}$$

Keterangan :
V_{kj}: pusat cluster dengan k=1,2,...,c dan j = 1,2,...,n
X_{ij} : data ke-i (i=1,2,...m), atribut ke-j (j=1,2,...n)
μ_{ik} : bilangan acak dengan i=1,2,...m dan k=1,2,...c

Tabel 1. Pusat Cluster Delphi

	Umur Panen	Anakan Produktif	Bobot 1000
Klaster 1	165.992143	21.9828308	19.1839070
Klaster 2	123.733348	16.3983313	23.4652036
Klaster 3	128.919368	12.3641866	24.2601746
Klaster 4	175.979512	15.7556963	19.0444179



Gambar 2. Form Proses Fuzzy C-Means

Dari Tabel 1 dapat dilihat bahwa klaster 1 memiliki umur panen terendah ketiga, anakan produktif terbanyak pertama, dan bobot 1000 terberat ketiga. Klaster 2 memiliki umur panen terendah pertama, anakan produktif terbanyak kedua, dan bobot 1000 terberat kedua. Klaster 3 memiliki umur panen terendah kedua, anakan produktif terbanyak keempat, dan bobot 1000 terberat pertama.

Klaster 4 memiliki umur panen terendah keempat, anakan produktif terbanyak ketiga, dan bobot 1000 terberat keempat.

Dilihat dari umur panen klaster 2 lebih unggul karena memiliki umur panen terendah dibandingkan klaster 1, klaster 3, dan klaster 4. Untuk anakan produktif klaster 1 lebih unggul karena memiliki anakan produktif terbanyak. Dan untuk bobot 1000 klaster 3 lebih unggul karena memiliki bobot seribu terberat apabila dibandingkan dengan klaster lainnya.

Dari hasil pusat klaster (Tabel 1) tersebut dilakukan perangkingan dengan metode *Simple Additive Weighting* (SAW) untuk melihat klaster mana yang memiliki umur panen, anakan produktif dan bobot 1000 terbaik. Perhitungan nilai prefrensi dilakukan menggunakan rumus dibawah ini (persamaam 9) :

$$V_i = \sum_{j=1}^{n} w_j r_{ij}$$

Keterangan :
V_i : nilaipreferensiuntuksetiapalternatif
w_j : nilai bobot preferensi
r_{ij} : matriks ternormalisasi

Berikut tabel nilai prefrensi dan tabel hasil pengelompokan per varietas :

Tabel 2. Nilai Prefensi Klaster

Klaster	Nila Prefensi	Kategori
Klaster 1	2.535898966	Cukup Tinggi
Klaster 2	2.712760587	Tinggi
Klaster 3	2.53314209	Sedang
Klaster 4	2.209208036	Rendah

Tabel 3. Hasil Pengelompokan Varietas Siam Harli

Kategori	Varietas Siam Harli (V1)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	-	-	-	46
Cukup Tinggi	57	77	32	-
Sedang	-	-	-	44
Rendah	4	2	3	-

Tabel 4. Hasil Pengelompokan Varietas Siam Unus

Kategori	Varietas Siam Unus (V2)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	-	1	-	-
Cukup Tinggi	48	205	48	86
Sedang	-	1	4	-
Rendah	4	-	5	8

Tabel 5. Hasil Pengelompokan Varietas Kuatek

Kategori	Varietas Kuatek (V3)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	-	-	-	371
Cukup Tinggi	7	152	18	176
Sedang	-	-	-	-
Rendah	41	5	132	-

Tabel 6. Hasil Pengelompokan Varietas Siam Sebelas

Kategori	Varietas Siam Sebelas (V4)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	5	-	-	25
Cukup Tinggi	1	28	9	-
Sedang	1	-	-	22
Rendah	25	158	78	-

Tabel 7. Hasil Pengelompokan Varietas Siam Gumpal

Kategori	Varietas Siam Gumpal (V5)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	-	-	-	13
Cukup Tinggi	-	-	-	-
Sedang	-	-	-	116
Rendah	54	173	105	-

Pengetahuan yang didapat dari hasil penambangan data/klastering ini berupa varietas unggul yaitu varietas yang masuk kategori tinggi. Varietas yang masuk kategori tinggi terdiri atas rata-rata umur panen sebesar 123,7665737; rata-rata anakan produktif sebesar 16,2834365; dan rata-rata bobot 1000 sebesar 23,46984343 dengan nilai preferensi sebesar 2,712760587. Diantara kelima varietas tersebut yang paling banyak masuk kategori tinggi yaitu varietas kuatek dengan dosis 30 krad yakni sebanyak 371 data padi atau sebesar 67,82 % data padi masuk kategori tinggi

Tabel 8. Presentase Data Padi Yang Termasuk Kategori Tinggi

Varietas	Dosis			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Siam Harli (V1)	-	-	-	51,11%
Siam Unus (V2)	-	0,93%	-	-
Kuatek (V3)	-	-	-	67,82%
Siam Sebelas (V4)	15,61%	-	-	53,19%
Siam Gumpal (V5)	-	-	-	10,08%

4. KESIMPULAN

Dari hasil penelitian dan pengamatan dari sistem yang telah dibuat, maka dapat diperoleh kesimpulan sebagai berikut :

1. Metode Fuzzy C-Means berhasil diimplemantasikan pada clustering data varietas padi lokal pasang surut kalimantan selatan yang dimutasi dengan sinar gamma.
2. Dilihat dari umur panen klaster 2 lebih unggul karena memiliki umur panen terendah, untuk anakan produktif klaster 1 lebih unggul karena memiliki anakan produktif terbanyak, dan untuk bobot 1000 klaster 3 lebih unggul karena memiliki bobot seribu terberat apabila dibandingkan dengan klaster lainnya.
3. Pengetahuan yang didapat mengenai varietas unggul berdasarkan umur panen, anakan produktif dan bobot 1000 terbaik menggunakan SAW memiliki nilai preferensi sebesar 23,46984343 yaitu kategori tinggi.
4. Diantara kelima varietas yang paling banyak masuk kategori tinggi yaitu varietas kuatek dengan dosis 30 krad yakni sebanyak 371 data padi atau sebesar 67,82 % data padi masuk kategori tinggi.

DAFTAR PUSTAKA

[1]. Budi Santosa, "**Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis***",* Graha Ilmu, Yogyakarta, 2007.

[2]. Iko Pramudiono. "***Pengantar Data Mining : Menambang Permata Pengetahuan di Gunung Data***" 2003.

[3]. Noor, M., Y. Rina dan Noorginayuwati, "**Persepsi petani tentang lahan gambut dan pengelolaannya",** Laporan Penelitian, Balai Penelitian Pertanian Lahan Rawa, Banjarbaru, 2007.

[4]. Nurjanah, "**Clustering Varietas Padi Lokal Pasang Surut Kalimantan Selatan Yang Dimutasi Dengan Sinar Gamma Menggunakan Metode Fuzzy C-Means**" Naskah Skripsi Universitas Lambung Mangkurat, Banjarbaru, 2014.

[5]. Nuqson Masykur Huda, "***Aplikasi Data Mining Untuk Menampilkan Informasi Tingkat Kelulusan Mahasiswa (Studi Kasus Di Fakultas MIPA Universitas Diponegoro)",*** Naskah Skripsi Universitas Diponegoro, Semarang, 2010.

[6]. Sri Kusumadewi and Hari Purnomo, "**Aplikasi Logika Fuzzy Untuk Pendukung Keputusan***",* Yogyakarta, Graha Ilmu, 2010.

[7]. Wibowo S, Henry., Amalia, Riska., Fadlun M, Andi., Arivanty, Kurnia, "**Sistem Pendukung Keputusan Untuk MenentukanPenerima Beasiswa Bank BRI Menggunakan FMADM (Studi Kasus: Mahasiswa Fakultas Teknologi Industri Unversitas Islam Indonesia)***",* Seminar Nasional Aplikasi Teknologi Informasi, Yogyakarta, 2008.

International Journal of Business and Management; Vol. 11, No. 7; 2016
ISSN 1833-3850 E-ISSN 1833-8119
Published by Canadian Center of Science and Education

Customer Clustering Using a Combination of Fuzzy C-Means and Genetic Algorithms

Azarnoush Ansari[1] & Arash Riasi[2]

[1] Department of Management, University of Isfahan, Isfahan, Iran

[2] Institute for Financial Services Analytics, University of Delaware, Newark, DE, USA

Correspondence: Arash Riasi, Alfred Lerner College of Business & Economics, Amstel Ave., Newark, DE, 19716, USA. Tel: 1-(302)-898-6249. E-mail: riasi@udel.edu

Received: April 23, 2016 Accepted: May 20, 2016 Online Published: June 18, 2016

doi:10.5539/ijbm.v11n7p59 URL: http://dx.doi.org/10.5539/ijbm.v11n7p59

Abstract

This study intends to combine the fuzzy c-means clustering and genetic algorithms to cluster the customers of steel industry. The customers were divided into two clusters by using the variables of the LRFM (length, recency, frequency, monetary value) model. Results indicated that customers belonging to the first cluster had a higher length of the relationship, recency of trade, and frequency of trade but lower monetary value compared to the average values of these criteria for all customers. The results also showed that customers belonging to the second cluster had a higher recency of trade and monetary value but lower length of the relationship and frequency of trade compared to the average values of these criteria for all customers. It was also found that the combined algorithm (i.e., fuzzy c-means clustering and genetic algorithm) used in this study had a lower mean squared error (MSE) compared to fuzzy c-means clustering.

Keywords: customer clusterin, fuzzy c-means clustering, genetic algorithms, LRFM model, steel industry, customer relationship management

1. Introduction

Identifying customers and customers' needs is an important factor which helps service providers and producers to gain competitive advantage in providing their products/services to the customers. Managers should prioritize their customers and focus on key customers and they should also try to gain a better intuition regarding the costs associated with losing their customers on a daily basis. This is important because when the customers discontinue doing business with us and move to our competitors, various negative consequences such as losing current revenues due to discontinued business relations or losing good reputation and credibility are expected (Tarokh & Sharifian, 2010). Our current customers will probably share their negative experiences with prospective customers. This loss of credibility can lead to a loss of current and potential customers' trust in our products/services. Experts suggest firms to design policies and competitive strategies, therefore organizations cannot disregard their fundamental goals including the achievement of competitive advantage. Identifying different groups of customers and their needs can lead to customer satisfaction, which in turn contributes to customer loyalty. In the long term, it is more beneficial to identify key customers and retain them rather than acquiring new customers to fill the empty place of those who have decided to discontinue doing business with the organization. This is mainly because the cost of new key customer acquisition is five times more than the cost of current customer retention (Griffin & Lowenstein, 2001). Companies have a higher probability of success in selling their products/services to their current customers compared to their prospective customers; the probability of successfully selling a product/service to current active customers is roughly 60-70 percent, while this probability is only 5-20 percent for prospective customers (Tarokh & Sharifian, 2010).

Classification is one of the most important topics in customer relationship management (CRM). In customer classification, the entire population of customers is divided into smaller groups, such that customers in the same group have similar characteristics. Ideally, organizations should have a good understanding about all of their customers, but this is not feasible in real world. Customer classification and clustering enable the firms to group similar customers together and help managers to better understand the customers' needs; because it is much easier to identify and analyze the characteristics of groups of customers rather than studying each customer individually. Data mining has various tools for customer classification. One of the most famous clustering

techniques is fuzzy c-means clustering which has problems with high dimensional data sets and a large number of prototypes (Winkler et al., 2012). Additionally, the performance of fuzzy c-means algorithm is strongly affected by the selection of the initial centroid clusters (Arnaldo & Bedregal, 2013). Therefore, in order to obtain better clusters, this study combines genetic algorithms and fuzzy c-means clustering. The main goal of this study is to use this combined clustering method to divide customers of the steel industry into two different clusters.

2. Research Question and Problem Statement

Effective relationship marketing is an art in today's business environment. Constructing relations with clients is an important skill which is required in order to retain the key customers. Classical marketing theories mostly focused on trade and did not consider customer retention as an important aspect of firms' marketing strategies. But as the markets became more competitive and saturated, and as population combinations of different areas changed over time, companies realized that they are no longer facing a growing economic system with growing markets. In order to survive in these competitive markets, firms need to gain competitive advantage (Amiri Aghdaie et al., 2012; Riasi, 2015a, 2015b). One of the most important determinants of competitive advantage is demand condition (Porter, 1990; Porter 1991) meaning that firms should focus on retaining their customers in order to ensure that their business is economically viable (Riasi &Amiri Aghdaie, 2013; Riasi & Pourmiri, 2015). Nowadays, every single customer has a special value because firms are struggling to obtain a larger share of their fixed or diminishing markets. As a result of this, the cost of new customer acquisition has been significantly increased. Companies have now realized that losing a customer does not simply mean that they are unable to sell a single item to that customer at a specific time, rather it means that they will lose all future purchases of that customer during his/her entire life (Kotler, 1994). Krasnikov et al. (2009) examined the impact of CRM on cost efficiency and profit efficiency and found that although CRM implementation leads to a decline in cost efficiency, it is associated with an increase in profit efficiency. The impact of CRM on long term profitability of the firms is very important, because the main objective of most firms is to maximize the value delivered to the shareholders. In order to maximize the shareholder value it is required to measure the value which is generated from different customers during a specific period and to identify the customers or groups of customers that contribute the most to the firm's value. After identifying these types of customers the firm should motivate them to establish long term relations. As a result of these actions, the customer loyalty will be enhanced, customer life-cycle will be optimized, and eventually the firm will become more profitable. In addition to these benefits, customer satisfaction and relationship commitment have a positive impact on firm's brand loyalty and brand awareness (Kim et al., 2008). It is clear that a firm's relation with its customers can affect its prosperity and it should be considered as an important issue when designing customer retention strategies. By dividing their customers into different clusters, firms can better decide how to effectively allocate their limited resources to different groups of customers based on their value. Also, by using customer clustering techniques, firms can effectively design their customer retention strategies and maximize their overall profitability.

Organizations collect tons of data about their customers, suppliers, and their business partners, but their inability to discover the latent knowledge which is inherent in these valuable data can make the whole data collection process useless. Many business owners are willing to discover this untapped knowledge from their data sets in order to increase their profitability. Data mining techniques including clustering methods enable the firms to exploit this hidden knowledge. Using clustering techniques, customers are divided into homogenous clusters in which customers with similar needs and characteristics are grouped together (Ghazanfari et al., 2010). After identifying the needs and values of its customers, a company must provide valuable products and services to its clients in order to increase their satisfaction and to ensure that they will remain loyal to the firm (Moslehi et al., 2012). But the main question is how to identify the key clients of the organization and how to analyze their behavioral attributes?

3. Literature Review

Various studies have been performed regarding customer segmentation and classification. Ghazanfari et al. (2010) studied the concept of customer segmentation in apparel industry using clustering algorithms. The goal of their study was to classify the countries based on the value of Iran's apparel exports during a 14-year period (1992-2006). To analyze the clusters and to measure their value, and also in order to design an appropriate strategy for each segment, they used RFM (recency, frequency, and monetary) model. RFM is a renowned model for customer value analysis and it has been used by many researchers to perform customer segmentation (Cheng & Chen, 2009; Coussement et al., 2014; Kafashpour et al., 2012; Khajvand & Tarokh, 2011; McCarty and Hastak, 2007; Michopoulou & Buhalis, 2008; Rastegar, 2009; Tao & Yeh, 2003; Wu et al.; 2009). RFM analyzes the behavior of the customers and therefore it is considered as a behavior-based model (Wei et al, 2012; Yeh et al., 2008). Hanafizadeh and Rastkhiz Paydar (2011) used neural networks to classify the customers of auto

insurance into 4 different groups based on their risk. The goal of their study was to use these clusters to identify the appropriate insurance premium for each group of customers. Wei et al. (2012) adopted self-organizing maps (SOM) technique to develop an extended RFM model, called LRFM (length, recency, frequency, and monetary) model for a children's dental clinic in Taiwan to segment its dental patients. They recommended 12 clusters and labeled the 3 clusters having the above average LRFM values as core patients. SOM is considered as a neural network technique and can be used for clustering problems, visualization, and market screening (Fish & Ruby, 2009; Hanafizadeh & Mirzazadeh, 2011; Hsu et al., 2009; Hung & Tsai, 2008; Kiang & Fisher, 2008; Mostafa & El-Masry, 2013; Wang, 2001; Wei et al., 2012). The advantage of SOM is that it can automatically detect strong features in large data sets (Wei et al., 2012). Gholamian and Niknam (2012) proposed an adjusted model for clustering customers in retail banking industry using RFM model. They tested their model on checking accounts of customers of a private bank and compared their results with the benchmark model. Chang and Tsay (2004) used LRFM model to increase the length of communication with customers. After extracting the data, they classified the customers into 16 groups. They believe that adding additional factors to the model, helps to identify loyal customers that provide the firm with higher profits. Li et al. (2011) used LRFM model to classify customers of textile industry into different clusters. Their results indicated that factories with domestic sales had more long-term and loyal customers, even though monetary value and frequency were not high. RFM model is unable to divide customers into different segments based on the length of the relationship with the company (Reinartz & Kumar, 2000). Additionally, the customer loyalty is dependent on both the type of the relationship that exists between the customer and the company (Chow & Holden, 1997) and the length of this relationship (Chang & Tsay, 2004). Therefore, in the current study, the LRFM model which is the extended version of RFM model is used to take the length of the relationship into account.

4. Methodology

This study can be classified as an applied research based on its objective and as a descriptive study based on its data collection method. The statistical society of this research included the managers and experts of steel industry. Since the size of the statistical society was known and all the measures were qualitative, the sample size was calculated to be 120 using the Cochran's formula.

4.1 Genetic Algorithms

Since clustering is an NP-hard problem, heuristic algorithms are used for data clustering. The most notable type of these algorithms is genetic algorithm, which has a strong capability in parallel implementations and has a low probability of being trapped in local minimums. Genetic algorithms became popular through the work of Holland (1975), and they are considered as well-developed search heuristics that mimic the natural selection mechanism. Genetic algorithms are iterative processes based on evolutions which initiate from a population of randomly generated individuals with the ultimate goal of finding comprehensive optimized solutions (Holland, 1975). The flowchart of genetic algorithm (Shyr, 2010) is depicted in Figure 1.

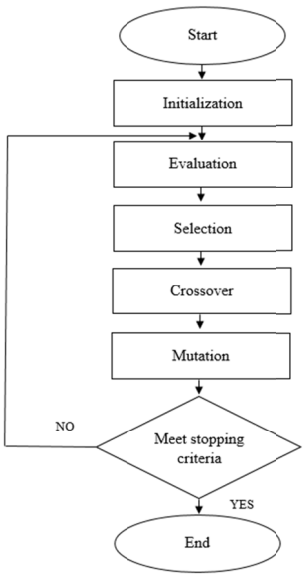


Figure 1. Genetic algorithm flowchart

4.2 Fuzzy C-Means Clustering

In classical clustering each input data point belongs to exactly one cluster and cannot be part of two or more clusters; in other words, the clusters do not overlap each other. Imagine that there is a data point which has characteristics similar to two or more clusters, in classical clustering we have to decide which cluster is more suitable for this data point and allocate it to only a single cluster. The main difference between classical and fuzzy clustering is that in fuzzy clustering a data point can belong to more than one cluster. Similar to classical c-means algorithm, in the fuzzy c-means algorithm the number of clusters is also known beforehand. The objective function of this algorithm is as follows:

$$J = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} d_{ik}^{2} = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} \left\| x_k - v_i \right\|^2 \tag{1}$$

Where m is a real number larger than 1, x_k is the kth data point, v_i is the centroid of the ith cluster, u_{ik} is the degree to which data point k belongs to cluster i, and $\|x_k - v_i\|$ is the Euclidean distance between k^{th} data point and i^{th} cluster center. Using u_{ik} a matrix U with c rows and n columns can be defined, where each element of the matrix can have a value between 0 and 1. If all the elements of matrix U are either 0 or 1 then the algorithm is similar to classical c-means. Although the elements of matrix U can have any value between 0 and 1, but the elements in each column should sum to 1. In other words:

$$\sum_{i=1}^{c} u_{ik} = 1 \quad , \quad \forall k = 1, \ldots, n \tag{2}$$

This indicates that the sum of the proportions that each data point belongs to each of the c different clusters should be equal to 1. Using the above condition and by minimizing the objective function we will have:

$$V_i = \frac{\sum_{k=1}^{n} u_{ik}^{m} x_k}{\sum_{k=1}^{n} u_{ik}^{m}} \, , u_{ik} = \frac{1}{\sum_{j=1}^{c} (\frac{d_{ik}}{d_{jk}})^{2/(m-1)}} \tag{3}$$

5. Data Analysis and Results

In order to perform the clustering, the data from 120 customers were collected and normalized. The data included four different variables, namely, length of the relationship, recency of trade, frequency of trade, and monetary value. In order to perform the fuzzy clustering, GA-Fuzzy Clustering software was used. This software performs the clustering by combining fuzzy c-means clustering and genetic algorithm. The parameters of these two algorithms are displayed in Table 1.

Table 1. Parameters of fuzzy c-means clustering and genetic algorithm

Algorithm	Parameter	Value
Fuzzy c-means clustering	Number of clusters	2
	Stop parameter	0.1
Genetic algorithm	Size of initial population	10
	Mutation frequency	0.5
	Crossover frequency	1
	Crossover percentage	0.5

After running the software, the customers were divided into two clusters. The cluster centers for each criterion were calculated as shown in Table 2. According to Table 2, customers belonging to the first cluster had a higher length of the relationship, recency of trade, and frequency of trade but lower monetary value compared to the average values of these criteria for all customers. Customers belonging to the second cluster had a higher recency of trade and monetary value but lower length of the relationship and frequency of trade compared to the average values of these criteria for all customers.

Table 2. Cluster centers for each criterion

Cluster	Length of the relationship (L)	Recency of trade (R)	Frequency of trade (F)	Monetary value (M)
First cluster ↑	0.8347 ↑	0.7945 ↑	0.6531 ↓	0.0212
Second cluster ↓	0.1352 ↑	0.4634 ↓	0.0319 ↑	0.7049

In order to compare the efficiency of the combined algorithm (i.e., fuzzy c-means clustering and genetic algorithm) with fuzzy c-means clustering, mean squared error (MSE) and run time criteria were used. Table 3 shows the MSE and run time for each of these two algorithms. According to table 3, the combined algorithm that was used in our study had a lower MSE but a higher run time compared to fuzzy c-means clustering algorithm. However, it is clear that in performing customer clustering, the MSE criteria is relatively more important than run time, because customer clustering is usually a time consuming process which needs to be done with high accuracy.

Table 3. MSE and run time comparison

Algorithm	MSE	Run time (s)
Fuzzy c-means clustering	0.0875	0.0546
Combined algorithm (fuzzy c-means clustering and genetic algorithm)	0.0152	0.8474

After running the combined algorithm, the customers were divided into two different clusters as shown in figure 2.

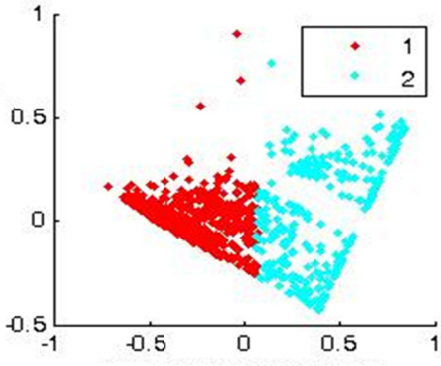


Figure 2. Clusters obtained from the combined algorithm

6. Discussion and Conclusions

As the importance of customer satisfaction became more evident in business environment, many organizations started to focus on customer identification, customer loyalty, and profitability analysis in order to increase their market share and to enhance the level of customer satisfaction. Appropriate customer relationship management is considered as a competitive advantage for organizations, and it is obvious that clustering is one of the most important issues in CRM. Therefore, the goal of this study was to use an appropriate model for clustering the customers of steel industry based on length of the relationship, recency of trade, frequency of trade, and monetary value criteria. As mentioned earlier, various algorithms can be used for data clustering including but not limited to fuzzy c-means algorithm. Although fuzzy c-means clustering is relatively easy to implement and has a low run time, its efficiency is not very high due to its sensitivity to initial centroid values and the possibility of being trapped into local optimum. The known deficiencies of classical clustering techniques, accentuates the need for an analytical clustering method which is both capable of providing optimal clusters and is expandable to complicated systems modelling. In this study, a combined algorithm which was a combination of fuzzy c-means clustering and genetic algorithms was used in order to analyze the data and to perform the

customer clustering. This combined algorithm was used because it had a higher efficiency and it did not have the deficiencies associated with classical clustering algorithms.

The customers of steel industry were divided into two clusters according to the variables of the LRFM model. The results revealed that customers belonging to the first cluster had a higher length of the relationship, recency of trade, and frequency of trade but lower monetary value compared to the average values of these criteria for all customers. On the other hand, the results indicated that customers belonging to the second cluster had a higher recency of trade and monetary value but lower length of the relationship and frequency of trade compared to the average values of these criteria for all customers. Therefore, based on the findings of Chang and Tsay (2004), the customers in the first cluster can be considered as loyal customers according to loyalty measures and they can be considered as uncertain customers based on value measures. Also the customers in the second cluster can be considered as newcomers based on the loyalty matrix and they can be considered as uncertain customers based on the value matrix. Additionally, the findings revealed that the combined clustering algorithm had a lower MSE and a higher run time compared to the fuzzy c-means algorithm. Since the accuracy is more important than speed in customer clustering, the authors suggest that future studies use a combination of fuzzy c-means clustering and genetic algorithm in order to obtain the most accurate clusters. This study clearly contributes to the literature in this field by showing that a combination of fuzzy c-means clustering and genetic algorithms is an efficient way to perform customer clustering in steel industry.

References

Amiri Aghdaie, S. F., Seidi, M., & Riasi, A. (2012). Identifying the Barriers to Iran's Saffron Exportby Using Porter's Diamond Model. *International Journal of Marketing Studies*, *4*(5), 129-138. http://dx.doi.org/10.5539/ijms.v4n5p129

Arnaldo, H. A., & Bedregal, B. R. C. (2013). A new way to obtain the initial centroid clusters in Fuzzy C-Means algorithm. *2013 2nd Workshop-School on Theoretical Computer Science (WEIT) - IEEE,* 139-144.

Chang, H., & Tsay, S. (2004). Integrating of SOM and K-mean in data mining clustering: An empirical study of CRM and profitability evaluation. *Journal of Information Management*, *11*, 161-203.

Cheng, C. H., & Chen, Y. S. (2009). Classifying the segmentation of customer value via RFM model and RS theory. *Expert Systems with Applications*, *36*(3), 4176-4184.

Chow, S., & Holden, R. (1997). Toward an understanding of loyalty: The moderating role of trust. *Journal of Managerial Issues*, *9*(3), 275-298.

Coussement, K., Van den Bossche, F. A., & De Bock, K. W. (2014). Data accuracy's impact on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees. *Journal of Business Research*, *67*(1), 2751-2758. http://dx.doi.org/10.1016/j.jbusres.2012.09.024

Fish, K., & Ruby, P. (2009). An artificial intelligence foreign market screening method for small businesses. *International Journal of Entrepreneurship*, *13*, 65-81.

Ghazanfari, M., Malek Mohammadi, S., & Alizadeh, S. (2010). Customer Segmentation in Clothing Exports Based on Clustering Algorithm. *Iranian Business Studies Quarterly*, *14*(56), 59-86.

Gholamian, M. R., & Niknam, Z. (2012). Providing an adjusted model for clustering bank customers based on their life cycle value. *Iranian Journal of Executive Management*, *4*(7), 59-76.

Griffin, J., & Lowenstein, M. W. (2001). *Customer winback: How to recapture lost customers and keep them loyal*. San Francisco, CA: Jossey-Bass.

Hanafizadeh, P., & Mirzazadeh, M. (2011). Visualizing market segmentation using self-organizing maps and Fuzzy Delphi method–ADSL market of a telecommunication company. *Expert Systems with Applications*, *38*(1), 198-205.

Hanafizadeh, P., & Rastkhiz Paydar, N. (2011). A model for risk-based clustering of auto insurance customers based on their risk using data mining techniques (Case study: auto insurance in an insurance company). *Iranian Insurance Research (Insurance Industry)*, *26*(2), 55-81.

Holland, J. (1975). *Adaptation in Natural and Artificial Systems.* Ann Arbor, MI: University of Michigan Press.

Hsu, S. H., Hsieh, J. P. A., Chih, T. C., & Hsu, K. C. (2009). A two-stage architecture for stock price forecasting by integrating self-organizing map and support vector regression. *Expert Systems with Applications*, *36*(4), 7947-7951.

Hung, C., & Tsai, C. F. (2008). Market segmentation based on hierarchical self-organizing map for markets of

multimedia on demand. *Expert Systems with Applications*, *34*(1), 780-787.

Kafashpour, A., Tavakoli, A., & Alizadeh, A. (2012). Clustering customers based on the life cycle value using data mining techniques and RFM model. *Iranian Journal of Public Management Research*, *5*(15), 63-84.

Khajvand, M., & Tarokh, M. J. (2011). Estimating customer future value of different customer segments based on adapted RFM model in retail banking context. *Procedia Computer Science*, *3*, 1327-1332.

Kiang, M. Y., & Fisher, D. M. (2008). Selecting the right MBA schools–An application of self-organizing map networks. *Expert Systems with Applications*, *35*(3), 946-955.

Kim, K. H., Kim, K. S., Kim, D. Y., Kim, J. H., & Kang, S. H. (2008). Brand equity in hospital marketing. *Journal of Business Research*, *61*(1), 75-82.

Kotler, P. (1994). *Marketing Management: Analysis, Planning, Implementation, and Control*. Upper Saddle River, NJ: Prentice Hall.

Krasnikov, A., Jayachandran, S., & Kumar, V. (2009). The impact of customer relationship management implementation on cost and profit efficiencies: evidence from the US commercial banking industry. *Journal of Marketing*, *73*(6), 61-76.

Li, D. C., Dai, W. L., & Tseng, W. T. (2011). A two-stage clustering method to analyze customer characteristics to build discriminative customer management: A case of textile manufacturing business. *Expert Systems with Applications*, *38*(6), 7186-7191.

McCarty, J. A., & Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. *Journal of Business Research*, *60*(6), 656-662.

Michopoulou, E., & Buhalis, D. (2008). Performance measures of net-enabled hypercompetitive industries: The case of tourism. *International Journal of Information Management*, *28*(3), 168-180.

Moslehi, N. (2012). *Using LRFM model for clustering customers based on their life cycles in order to improve customer relationship management* (Master's thesis). Ferdowsi University of Mashhad, Mashhad, Iran.

Mostafa, M. M., & El-Masry, A. A. (2013). Citizens as consumers: Profiling e-government services' users in Egypt via data mining techniques. *International Journal of Information Management*, *33*(4), 627-641.

Porter, M. E. (1990). *The Competitive Advantage of Nations*. New York, NY: The Free Press.

Porter, M. E. (1991). *Canada at the crossroads: The reality of a new competitive environment*. Ottawa: Business Council on National Issues and Minister of Supply and Services of the Government of Canada.

Rastegar, N. (2009). *Presenting the developed RFM model for clustering customers of banking services*. Paper presented at the first international conference of banking services marketing, Tehran, Iran. Retrieved January 15, 2016, from http://www.ensani.ir/fa/14822/magazine.aspx

Reinartz, W. J., & Kumar, V. (2000). On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing. *Journal of Marketing*, *64*(4), 17-35.

Riasi, A. (2015a). Barriers to international supply chain management in Iranian flower industry. *Management Science Letters*, *5*(4), 363-368. http://dx.doi.org/10.5267/j.msl.2015.2.005

Riasi, A. (2015b). Competitive Advantages of Shadow Banking Industry: An Analysis Using Porter Diamond Model. *Business Management and Strategy*, *6*(2), 15-27.

Riasi, A., & Amiri Aghdaie, S. F. (2013). Effects of a Hypothetical Iranian Accession to the World Trade Organization on Iran's Flower Industry. *Consilience: The Journal of Sustainable Development, 10*(1), 99-110. DOI: 10.7916/D8HQ3ZK8

Riasi, A., & Pourmiri, S. (2015). Effects of online marketing on Iranian ecotourism industry: Economic, sociological, and cultural aspects. *Management Science Letters*, *5*(10), 915-926.

Shyr, W. J. (2010). Parameters Determination for Optimum Design by Evolutionary Algorithm. In M. Crisan (Ed.), *Convergence and Hybrid Information Technologies*. Intech. Retrieved January 15, 2016, from http://www.intechopen.com/books/convergence-and-hybrid-information-technologies/parameters-determination-for-optimum-design-by-evolutionary-algorithm

Tao, Y. H., & Yeh, C. C. R. (2003). Simple database marketing tools in customer analysis and retention. *International Journal of Information Management*, *23*(4), 291-301.

Tarokh, M., & Sharifian, K. (2010). Applications of data mining in improving customer communication

management. *Iranian Industrial Management Studies Quarterly*, *6*(17), 153-181.

Wang, S. (2001). Cluster analysis using a validated self‐organizing method: Cases of problem identification. *Intelligent Systems in Accounting, Finance and Management*, *10*(2), 127-138.

Wei, J. T., Lin, S. Y., Weng, C. C., & Wu, H. H. (2012). A case study of applying LRFM model in market segmentation of a children's dental clinic. *Expert Systems with Applications*, *39*(5), 5529-5533.

Winkler, R., Klawonn, F., & Kruse, R. (2012). Problems of fuzzy c-means clustering and similar algorithms with high dimensional data sets. In W. A. Gaul, A. Geyer-Schulz, L. Schmidt-Thieme, & Y. Kunze (Eds.), *Challenges at the Interface of Data Analysis, Computer Science, and Optimization* (pp. 79-87). Heidelberg, Germany: Springer Berlin Heidelberg.

Wu, H. H., Chang, E. C., & Lo, C. F. (2009). Applying RFM model and K-means method in customer value analysis of an outfitter. In S. Y. Chou, A. Trappey, J. Pokojski, & S. Smith (Eds.), *Global Perspective for Competitive Enterprise, Economy and Ecology* (pp. 665-672). London, U.K.: Springer London.

Yeh, I. C., Yang, K. J., & Ting, T. M. (2009). Knowledge discovery on RFM model using Bernoulli sequence. *Expert Systems with Applications*, *36*(3), 5866-5871.

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).

ISSN:2442-5567 | E-ISSN:2443-289X

PENENTUAN PENERIMA BEASISWA DENGAN ALGORITMA *FUZZY C-MEANS* DI UNIVERSITAS MEGOW PAK TULANG BAWANG

Muhardi[1] Nisar[2]

[1]Universitas Megow Pak [2]MTI Institut Informatika & Bisnis Darmajaya
Email: ardyardy68@yahoo.co.id

ABSTRAK

Beasiswa adalah pemberian berupa bantuan keuangan yang diberikan kepada perorangan yang bertujuan untuk digunakan demi keberlangsungan pendidikan yang ditempuh. Penentuan beasiswa dapat dikelompokan berdasarkan kriteria Indeks Prestasi Kumulatif (IPK), tingkat kemiskinan, tanggungan orang tua dan prestasi mahasiswa dalam proses rekruitmen beasiswa. Algoritma *Fuzzy C-Means* merupakan satu algoritma yang mudah dan sering digunakan dalam pengelompokan data karena membuat suatu perkiraan yang efisien dan tidak memerlukan banyak parameter. Beberapa penelitian telah menghasilkan kesimpulan bahwa metode *Fuzzy C-Means* dapat digunakan untuk mengelompokkan data berdasarkan atribut-atribut tertentu. Penerapan algoritma *Fuzzy C-Means* dalam penentuan beasiswa di kelompokan menjadi tiga *cluster* yaitu menerima, dipertimbangkan dan tidak berhak menerima beasiswa , sempel data sebenyak 75 data mahasiswa diperoleh tiga *cluster* berdasarkan nilai rata-rata penentuan beasiswa kemudian setiap *cluster* diklasifikasikan berdasarkan kriteria mana yang lebih diprioritaskan dengan nilai terbesar pada jarak akhir merupakan *cluster* yang menerima beasiswa, sedangkan *cluster* dengan nilai terkecil merupakan *cluster* yang tidak berhak menerima beasiswa.

Kata Kunci: *Beasiswa, clustering, Fuzzy C- Means.*

ABSTRACT

The scholarship is the grant in the form of financial assistance given to an individual who intends to use for the sake of the continuity of education traveled. Determination of the scholarship can be grouped based on the criteria of the cumulative Achievement Index (IPK), the level of poverty, dependent parent and student achievement in the process of recruitment of scholarship. Fuzzy C-Means algorithm is an algorithm that is easily and often used in grouping data because making an estimate of efficient and doesn't require a lot of parameters. Some research has resulted in the conclusion that the Fuzzy C-Means method can be used to classify data based on certain attributes. Application of Fuzzy C-Means algorithm in determination of scholarship in kelompokan into three clusters, namely receiving, considering and not eligible to receive the scholarship, sempel 75 student data sebenyak data retrieved three clusters based on the average value of the determination of a scholarship then each cluster are classified according to criteria which is more prioritized with the greatest value at the end of the cluster is receiving a scholarship, whereas the smallest value cluster is a cluster that is not eligible to receive the scholarship.

Keywords*: scholarship, clustering, Fuzzy C- Means*

ISSN:2442-5567 | E-ISSN:2443-289X

1. PENDAHULUAN

Beasiswa adalah pemberian berupa bantuan keuangan yang diberikan kepada perorangan yang diberikan kapada perorangan yang bertujuan untuk digunakan demi keberlansungan pendidikan yang ditempuh. Pemberian beasiswa merupakan program kerja yang ada di setiap universitas atau perguruan tinggi. Program beasiswa diadakan untuk meringankan beban mahasiswa dalam menempuh masa studi kuliah khususnya dalam masalah biaya. Pemberian beasiswa kepada mahasiswa dilakukan secara selektif sesuai dengan jenis beasiswa yang diadakan. Univeritas Megow Pak Tulang Bawang menyediakan beberpa program beasiswa, sebagai contoh yaitu beasiswa Peningkatan Akademik (PPA), Beasiswa Bantuan Belajar Mahasiswa (BBM) dan lain sebagainya. Indeks prestasi kumulatif, Tingkat Kemiskinan, tanggungan orang tua dan prestasi menjadi kriteria dalam proses rekruitmen beasiswa.

Proses seleksi penerimaan beasiswa secara manual yaitu dengan menginputkan satu persatu data mahasiswa ke dalam file *excel* kemudian melakukan *sorting* data mahasiswa seringkali melakukan beberapa permasalahan, antara lain membutuhkan waktu yang lama dan ketelitian yang tinggi. Selain itu, transparansi serta ketidak jelasan metodologi yang digunakan dalam proses komputasi penerimaan beasiswa juga menjadi salah satu permasalahan, sehingga dibutuhkan suatu sistem yang dapat membantu dalam proses pengambilan keputusan siapa saja mahasiswa yang menerima beasiswa berdasarkan kriteria-kriteria yang telah ditentukan secara capat dan tepat sasaran.

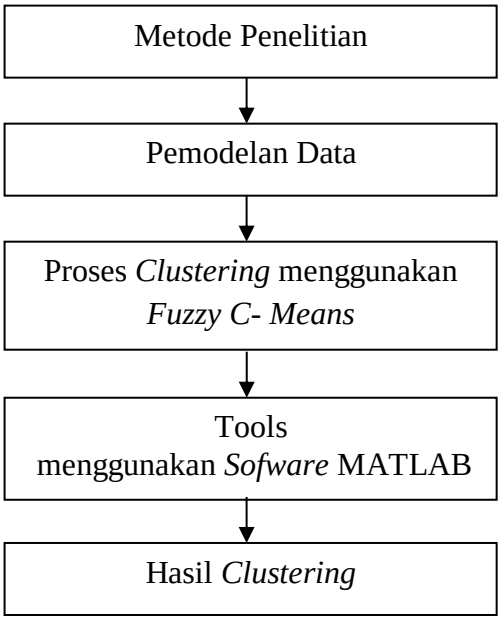
Salah satu metode yang dapat digunakan untuk menyelesaikan permasalahan tersebut adalah metode *Fuzzy C-Means.* Algoritma *Fuzzy C-Means* merupakan satu algoritma yang mudah dan sering digunakan dalam pengelompokan data kerana membuat suatu perkiraan yang efsien dan tidak memerlukan banyak parameter. Beberapa penelitian telah menghasilkan kesimpulan bahwa metode *Fuzzy CMeans* dapat digunakan untuk mengelompokkan data berdasarkan atribut-atribut tertentu. Pada kasus penelitian ini akan menganalisis penerapan metode *Fuzzy C-Means* untuk mengelompokkan data mahasiswa berdasarkan kemampuan mahasiswa dibidang akademik untuk proses penentuan beasiswa.
Pada penelitian sebelumnya, Bahari (2011) melakukan Penerapan metode *Fuzzy C-Means* untuk mengelompokan data siswa dalam penentuan jurusan di Sekolah Menengah Atas pada 81 sampel data siswa yang diuji dalam penelitian menunjukan bahwa metode *Fuzzy C-Means* memiliki tingkat akurasi yang lebih tinggi (rata-rata 78,39%), jika dibandingkan dengan metode penentuan jurusan secara manual yang di lakukan (hanya memiliki tingkat akurasi rata-rata 56,17 %).

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitan

Penelitian ini akan dilaksanakan berdasarkan rancangan penelitian seperti yang di tunjukan pada Gambar 1.

ISSN:2442-5567 | E-ISSN:2443-289X



Gambar 1. Alur Kerangka Penelitian

Penelitian yang dilaksanakan adalah jenis penelitian eksprimen, yaitu melakukan pengujian tingkat akurasi algoritma *Fuzzy C-Means clustering* sebagai model untuk menentukan penerima beasiswa. Data eksperimen diambil dari tempat penelitian yaitu Universitas Megow Pak Tulang Bawang.

2.2 Pengumpulan Data

Pengumpulan data digunakan untuk mengumpulkan data-data dan informasi-informasi yang diperlukan dalam pembuatan sistem pendukung keputusan. Pengumpulan data pada penelitian ini menggunakan metode pengumpulan data *study literature* dan telaah dokumen.

a. *Study Literature*

Study literature dilakukan dengan cara mencari bahan materi yang berhubungan dengan permasalahan, algoritma *Fuzzy C-Means*, sistem pendukung keputusan dan beasiswa, guna mempermudah proses implementasi sistem. Pencarian materi dilakukan melalui pencarian di buku panduan dan internet.

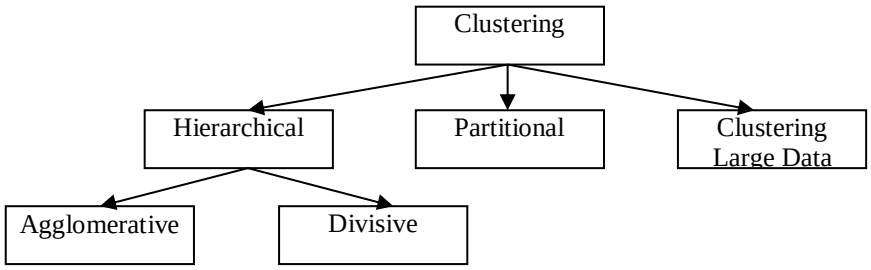
b. Telaah Dokumen

Telaah dokumen adalah pengumpulan data dengan cara mengumpulkan dan mempelajari dokumen-dokumen yang didapatkan dari pihak Jurusan Informaitka FMIPA Universitas Megow Pak Tulang Bawang. Dari metode pengumpulan data ini diperoleh 75 data mahasiswa.

2.3 Algoritma *Clustering*

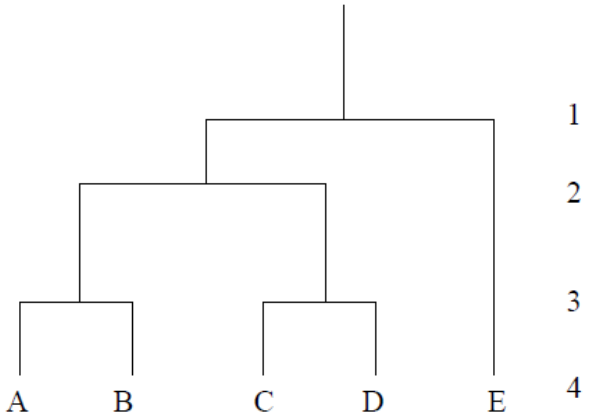
Secara umum pembagian algoritma *clustering* dapat digambarkan sebagai berikut:

ISSN:2442-5567 | E-ISSN:2443-289X



Gambar 2. Kategori Algoritma *Clustering*

Hierarchical clustering menentukan sendiri jumlah *cluster* yang dihasilkan. Hasil dari metode ini adalah suatu struktur data berbentuk pohon yang disebut dendogram dimana data dikelompokkan secara bertingkat dari yang paling bawah dimana tiap *instance* data merupakan satu *cluster* sendiri, hingga tingkat paling atas dimana keseluruhan data membentuk satu *cluster* besar berisi *cluster-cluster* seperti gambar 2.



Gambar 3. *Dendogram*

Divisive hierarchical clustering mengelompokkan data dari kelompok yang terbesar hingga ke kelompok yang terkecil, yaitu masing-masing *instance* dari kelompok data tersebut. Sebaliknya, *agglomerative hierarchical clustering* mulai mengelompokkan data dari kelompok yang terkecil hingga kelompok yang terbesar. Beberapa algoritma yang menggunakan metode ini adalah: *RObust Clustering Using LinKs* (ROCK), *Chameleon, Cobweb, Shared Nearest Neighbor* (SNN).

Partitional clustering yang mengelompokkan data ke dalam k *cluster* dimana k adalah banyaknya *cluster* dari input user. Kategori ini biasanya memerlukan pengetahuan yang cukup mendalam tentang data dan proses bisnis yang memanfaatkannya untuk mendapatkan kisaran nilai input yang sesuai. Beberapa algoritma yang masuk dalam kategori ini antara lain: *K-Means, Fuzzy C-Means, Clustering Large Aplications* (CLARA), *Expectation Maximation* (EM), *Bond Energy Algorithm* (BEA), *algoritma Genetika*, Jaringan Saraf Tiruan.

ISSN:2442-5567 | E-ISSN:2443-289X

a. Algoritma *Fuzzy Clustering C-Means* (FCM)

Pada proses pengklasteran (*clustering*) secara klasik (misalnya pada algoritma *Clustering K-Means*), pembentukan partisi dilakukan sedemikian rupa sehingga setiap obyek berada tepat pada satu partisi. Namun, adakalanya tidak dapat menempatkan suatu obyek tepat pada suatu partisi, karena sebenarnya obyek tersebut terletak di antara 2 atau lebih partisi yang lain. Pada logika *fuzzy*, metode yang dapat digunakan untuk melakukan pengelompokan sejumlah data dikenal dengan nama *fuzzy clustering*. *Fuzzy Clustering* lebih alami jika dibandingkan dengan pengklasteran secara klasik. Suatu algoritma *clustering* dikatakan sebagai *fuzzy clustering* jika algoritma tersebut menggunakan parameter strategi adaptasi secara *soft competitive*. Sebagian besar algoritma *fuzzy clustering* didasarkan atas optimasi fungsi obyektif atau modifikasi dari fungsi obyektif tersebut.

Salah satu teknik *fuzzy clustering* adalah *Fuzzy C-Means* (FCM). FCM adalah suatu teknik pengklasteran data yang keberadaan tiap-tiap data dalam suatu *cluster* ditentukan oleh nilai/derajat keanggotaan tertentu. Teknik ini pertama kali diperkenalkan oleh Jim Bezdek pada tahun 1981. Berbeda dengan teknik pengklasteran secara klasik (dimana suatu obyek hanya akan menjadi anggota suatu klaster tertentu), dalam FCM setiap data bisa menjadi anggota dari beberapa *cluster*.Batas-batas *cluster* dalam FCM adalah lunak (*soft*). Konsep dasar FCM, pertama kali adalah menentukan pusat *cluster* yang akan menandai lokasi rata-rata untuk tiaptiap *cluster*. Pada kondisi awal, pusat *cluster* ini masih belum akurat. Tiap-tiap data memiliki derajat keanggotaan untuk tiap-tiap *cluster*. Dengan cara memperbaiki pusat *cluster* dan nilai keanggotaan tiap-tiap data secara berulang, maa akan terlihat bahwa pusat *cluster* akan bergerak menuju lokasi yang tepat. Perulangan ini didasarkan pada minimasi fungsi obyektif. Fungsi Obyektif yang digunakan pada FCM adalah:

$$J_w(U,V;X) = \sum_{k-1}^{n} \sum_{i-=1}^{c} (\mu_{ik})^w (d_{ik})^2 \quad (1)$$

dengan + ∈ ,1,∞),

$$d_{ik} = d(x_k - v_i) = [\sum_{j=1}^{m} (x_{kj} - x_{ij})]^{\frac{1}{2}} \quad (2)$$

x adalah data yang akan diklaster:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix} \quad (3)$$

dan v adalah matriks pusat *cluster* :

$$V = \begin{bmatrix} v_{11} & \cdots & v_{1m} \\ \vdots & & \vdots \\ v_{m1} & \cdots & v_{mm} \end{bmatrix} \quad (4)$$

nilai Jw terkecil adalah yang terbaik, sehingga:

$$J_w^* (U^*, U^*; x) = \min J(U, U; x) \quad (5)$$

Jika $d_{ik} > 0, \forall i, k$; $w > 1$ dan X setidaknya memiliki m elemen, maka (u, v) ∈ $M_{fm} > 0, \forall i, k$; w dapat meminimasi Jw hanya jika:

$$\mu_{ik}=\frac{[\sum_{j=1}^{m}(x_{ij}-v_{kj})^{2}]^{\frac{-1}{w-1}}}{\sum_{i=1}^{n}[\sum_{j=1}^{m}(x_{ij}-v_{kj})^{2}]^{\frac{-1}{w-1}}}; \quad (6)$$

dan

$$v_{kj}=\frac{\sum_{i=1}^{n}((\mu_{ik})^{w}*X_{ij})}{\sum_{i=1}^{n}(\mu_{ik})^{w}};1 \leq i \leq m;1 \leq J \leq m \quad (7)$$

Algoritma *Fuzzy C-Means* (FCM) diberikan sebagai berikut :

1. Menentukan data yang akan di cluster X, berupa matriks berukuran n x m (n=jumlah sampel data, m = atribut setiap data). Xij=data sampel ke-I (i=1,2,…,n), atribut ke-j (j=1,2,…,m).
2. Menentukan :
   - Jumlah cluster = c
   - Pangkat = w
   - Maksimum interasi = MaxIter
   - Error terkecil yang diharapkan = ξ
   - Fungsi objektif awal = $P_0 = 0$
   - Interasi awal = t = 1

3. Membangkitkan bilangan random μik, i=1,2,3 …,n; k=1,2,3…c; sebagai elemen-elemen matriks partisi awal U.
   Menghitung jumlah setiap kolom :

$$Q_i=\sum_{k=1}^{c}\mu_{ik} \quad (8)$$

dengan j=1,2,…n.

Menghitung :

$$\mu_{ik}=\frac{\mu_{ik}}{Q_i} \quad (9)$$

4. Menghitung pusat cluster ke-k: V_{kj}, dengan k=1,2,…c; dan j=1,2,…m

$$V_{kj}=\frac{\sum_{i=1}^{n}((\mu_{ik})^{w}*X_{ij}}{\sum_{i=1}^{n}((\mu_{ik})^{w}} \quad (10)$$

5. Menghitung fungsi objektif pada interasi ke-t :

$$P_t=\sum_{i=1}^{n}\sum_{k}^{c}([\sum_{j=1}^{m}(X_{ij}-V_{kj})^{2}](\mu_{ik})^{w}) \quad (11)$$

6. Menghitung perubahan matriks partisi :

$$\mu_{ik}=\frac{[\sum_{j=1}^{m}(X_{ij}-V_{kj})^{2}]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}[\sum_{j=1}^{m}(X_{ij}-V_{kj})^{2}]^{\frac{-1}{w-1}}} \quad (12)$$

b. Proses *Clustering* menggunakan *Fuzzy C-Means*

Tahap ini akan diterapkan metode *Fuzzy C-Means* untuk mengelompokkan data. Hasil pengelompokkan ini kemudian akan digunakan untuk pertimbangan menentukan

ISSN:2442-5567 | E-ISSN:2443-289X

mahasiswa yang berhak menerima beasiswa. Adapun algoritma *C-Means Clustering* pada penerima beasiswa adalah sebagai berikut :

1. Menetapkan matriks partisi awal U berupa matriks berukuran n x m (n adalah jumlah sampel data, yaitu=75, dan m adalah parameter/atribut setiap data, yaitu=2). X_{ij}=data sampel ke-i (i=1,2,…,n), atribut ke-j (j=1,2,..,m).
2. Menentukan Nilai Parameter Awal :
   - Jumlah *cluster* (c) = 3
   - Pangkat/bobot (w) = 2
   - Maksimum interasi (MaxIter) = 100
   - Error terkecil yang diharapkan (ξ)= 10^{-5}
   - Fungsi Objektif awal (Po) = 0
   - Interasi awal (t) = 1
3. Membangkitkan bilangan random μik, i=1,2,…c; sebagai elemen-elemen matriks partisi awal (U).

3. HASIL DAN PEMBAHASAN

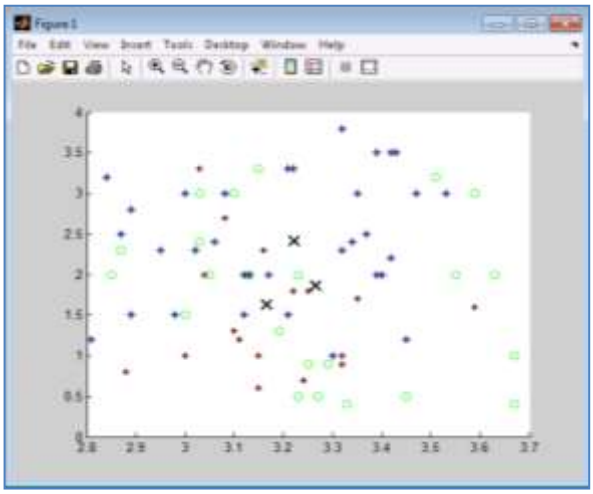
3.1 Tools menggunakan *Sofware* MATLAB

Data yang digunakan dalam penelitian ini adalah data penentuan beasiswa sebanyak 75 mahasiswa. Dalam proses *clustering*, IPK dijadikan sebagai Xi1, penghasilan orang tua dijadikan Xi2, tanggungan orang tua dijadikan Xi3 dan prestasi dijadikan Xi4. Pengolahan data pada penelitian ini menggunakan bantuan program yaitu Matlab. Dari *clustering* yang dilakukan diperoleh hasil yaitu nilai fungsi obyektif selama iterasi, pusat *cluster* atau *center* serta derajat keanggotaan lulusan untuk setiap *cluster* pada iterasi terakhir. Dalam penelitian ini, proses iterasinya berhenti pada iterasi ke-37 karena nilai |Pt - Pt-1|< ξ. Nilai fungsi obyektif pada iterasi terakhir yang diperoleh adalah 111.949781. Untuk penelitian ini *clustering* menggunakan *sofware* Matlab denga menjalankan di *Command Window* yaitu: >> fcm_k3.m

```
Iteration count = 1, obj. fcn = 185.869420
Iteration count = 2, obj. fcn = 146.530736
Iteration count = 3, obj. fcn = 141.726376
Iteration count = 4, obj. fcn = 134.833704
Iteration count = 5, obj. fcn = 130.961108
Iteration count = 6, obj. fcn = 129.145894
Iteration count = 7, obj. fcn = 127.219322
Iteration count = 8, obj. fcn = 124.903487
Iteration count = 9, obj. fcn = 122.880711
Iteration count = 10, obj. fcn = 121.359551
Iteration count = 11, obj. fcn = 119.898753
Iteration count = 12, obj. fcn = 118.236843
Iteration count = 13, obj. fcn = 116.491898
Iteration count = 14, obj. fcn = 114.963865
Iteration count = 15, obj. fcn = 113.865257
Iteration count = 16, obj. fcn = 113.180881
Iteration count = 17, obj. fcn = 112.770379
Iteration count = 18, obj. fcn = 112.513345
Iteration count = 19, obj. fcn = 112.342297
Iteration count = 20, obj. fcn = 112.223512
Iteration count = 21, obj. fcn = 112.139405
```

```
Iteration count = 22, obj. fcn = 112.079712
Iteration count = 23, obj. fcn = 112.037682
Iteration count = 24, obj. fcn = 112.008476
Iteration count = 25, obj. fcn = 111.988487
Iteration count = 26, obj. fcn = 111.975011
Iteration count = 27, obj. fcn = 111.966053
Iteration count = 28, obj. fcn = 111.960170
Iteration count = 29, obj. fcn = 111.956348
Iteration count = 30, obj. fcn = 111.953887
Iteration count = 31, obj. fcn = 111.952314
Iteration count = 32, obj. fcn = 111.951315
Iteration count = 33, obj. fcn = 111.950683
Iteration count = 34, obj. fcn = 111.950285
Iteration count = 35, obj. fcn = 111.950035
Iteration count = 36, obj. fcn = 111.949879
Iteration count = 37, obj. fcn = 111.949781
C =
   3.2649   1.8625   2.2484   4.2701
   3.1653   1.6271   5.1550   1.1770
   3.2225   2.4133   3.0759   0.7967
```

Penyebaran masing-masing anggota *cluster* pada iterasi terakhir dapat dilihat pada *cluster interface* gambar dibawah ini :



Gambar 4. Posisi *Claster* Iterasi terakhir

Keterangan :
Tanda * = *cluster* 1
Tanda + = *cluster* 2
Tanda ○ = *cluster* 3
Tanda X = pusat ke 3 *cluster*

Detail Matriks partisi U yang dihasilkan pada iterasi terakhir (iterasi ke-37) menggunakan fungsi Matlab adalah :

ISSN:2442-5567 | E-ISSN:2443-289X

Columns 1 through 10

0.1634	0.5724	0.0523	0.0467	0.9198	0.0643	0.0525	0.1063	0.0438	0.0605
0.1810	0.1653	0.0954	0.3021	0.0345	0.7542	0.2756	0.1726	0.3300	0.1892
0.6557	0.2623	0.8522	0.6512	0.0457	0.1815	0.6719	0.7211	0.6262	0.7503

Columns 11 through 20

0.0507	0.0576	0.6670	0.0923	0.6497	0.1132	0.6893	0.0108	0.7094	0.0427
0.7378	0.2660	0.1091	0.0971	0.1801	0.6406	0.1642	0.9393	0.0919	0.3048
0.2115	0.6763	0.2239	0.8107	0.1701	0.2463	0.1465	0.0500	0.1987	0.6525

Columns 21 through 30

0.0463	0.0299	0.8531	0.3603	0.0327	0.0592	0.0435	0.0470	0.8602	0.0845
0.7249	0.0572	0.0618	0.3797	0.8855	0.1855	0.3987	0.3511	0.0572	0.7507
0.2288	0.9129	0.0852	0.2600	0.0818	0.7553	0.5578	0.6018	0.0826	0.1648

Columns 31 through 40

0.8104	0.0560	0.1747	0.0183	0.3397	0.6931	0.0611	0.0556	0.2106	0.6219
0.0792	0.8198	0.6184	0.0430	0.3074	0.1693	0.1484	0.1314	0.5727	0.1510
0.1105	0.1242	0.2069	0.9386	0.3530	0.1376	0.7904	0.8130	0.2167	0.2271

Columns 41 through 50

0.0513	0.1739	0.5558	0.8307	0.8829	0.1085	0.0982	0.0307	0.0732	0.0897
0.8324	0.6221	0.2029	0.0788	0.0535	0.2561	0.1566	0.0896	0.4973	0.0934
0.1163	0.2040	0.2413	0.0905	0.0636	0.6354	0.7452	0.8797	0.4295	0.8170

Columns 51 through 60

0.0050	0.1026	0.0297	0.0466	0.0378	0.0673	0.0446	0.0707	0.5855	0.5157
0.9734	0.1039	0.0565	0.2871	0.1144	0.1508	0.3947	0.5757	0.1587	0.2346
0.0216	0.7934	0.9139	0.6663	0.8478	0.7819	0.5607	0.3536	0.2558	0.2497

Columns 61 through 70

0.0825	0.0826	0.0433	0.1208	0.6158	0.0877	0.1063	0.7764	0.0346	0.8865
0.2752	0.2753	0.3423	0.1954	0.1421	0.1292	0.1073	0.0887	0.0653	0.0543
0.6423	0.6421	0.6144	0.6838	0.2421	0.7831	0.7864	0.1349	0.9000	0.0592

Columns 71 through 75

0.8488	0.0548	0.7033	0.6666	0.7703
0.0635	0.4938	0.0943	0.1236	0.1006
0.0877	0.4514	0.2024	0.2097	0.1291

Pada iterasi terakhir (iterasi ke-37), pusat *cluster Vkj* yang dihasilkan dengan k=1,2,3,4 dan j=1,2 adalah:

$$V_{kj} = \begin{pmatrix} 3.2649 & 1.8625 & 2.2484 & 4.2701 \\ 3.1653 & 1.6271 & 5.1550 & 1.1770 \\ 3.2225 & 2.4133 & 3.0759 & 0.7967 \end{pmatrix}$$

ISSN:2442-5567 | E-ISSN:2443-289X

Nilai ini merupakan nilai dari koordinat ketiga titik pusat *cluster* dan memberikan garis besar tiap *cluster* yaitu:

1. Kelompok pertama (*cluster* ke-1), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.2649; memiliki rata-rata penghasilan orang tua sekitar 1.8625; tanggungan orang tua 2.2484 dan memiliki prestasi 4.2701.
2. Kelompok kedua (*cluster* ke-2), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.1653; memiliki rata-rata penghasilan orang tua sekitar 1.6271; tanggungan orang tua 5.1550 dan memiliki prestasi 1.1770.
3. Kelompok ketiga (*cluster* ke-3), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.2225; memiliki rata-rata penghasilan orang tua sekitar 2.4133; tanggungan orang tua 3.0759 dan memiliki prestasi 0.7967.

Derajat keanggotaan penentuan beasiswa untuk sertiap *cluster* pada iterasi terakhir (iterasi ke-37) dapat dilihat pada Tabel 1. di bawah ini:

Tabel 1. Derajat Keanggotaan Penentuan Beasiswa Untuk Setiap *Cluster* Pada Iterasi Terakhir

No	Derajat Keanggotaan (μ) pada Iterasi Terakhir			Kecenderungan data masuk pada *cluster*		
	(μ_1)	(μ_2)	(μ_3)	C1	C2	C3
1	0.1634	0.1810	0.6557			*
2	0.5724	0.1653	0.2623	*		
3	0.0523	0.0954	0.8522			*
4	0.0467	0.3021	0.6512			*
5	0.9198	0.0345	0.0457	*		
6	0.0643	0.7542	0.1815		*	
7	0.0525	0.2756	0.6719			*
8	0.1063	0.1726	0.7211			*
9	0.0438	0.3300	0.6262			*
10	0.0605	0.1892	0.7503			*
11	0.0507	0.7378	0.2115		*	
12	0.0576	0.2660	0.6763			*
13	0.6670	0.1091	0.2239	*		
14	0.0923	0.0971	0.8107			*
15	0.6497	0.1801	0.1701	*		
16	0.1132	0.6406	0.2463		*	
17	0.6893	0.1642	0.1465	*		
18	0.0108	0.9393	0.0500		*	
19	0.7094	0.0919	0.1987	*		
20	0.0427	0.3048	0.6525			*
21	0.0463	0.7249	0.2288		*	
22	0.0299	0.0572	0.9129			*
23	0.8531	0.0618	0.0852	*		
24	0.3603	0.3797	0.2600		*	
25	0.0327	0.8855	0.0818		*	
26	0.0592	0.1855	0.7553			*

ISSN:2442-5567 | E-ISSN:2443-289X

Nilai ini merupakan nilai dari koordinat ketiga titik pusat *cluster* dan memberikan garis besar tiap *cluster* yaitu:

1. Kelompok pertama (*cluster* ke-1), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.2649; memiliki rata-rata penghasilan orang tua sekitar 1.8625; tanggungan orang tua 2.2484 dan memiliki prestasi 4.2701.
2. Kelompok kedua (*cluster* ke-2), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.1653; memiliki rata-rata penghasilan orang tua sekitar 1.6271; tanggungan orang tua 5.1550 dan memiliki prestasi 1.1770.
3. Kelompok ketiga (*cluster* ke-3), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.2225; memiliki rata-rata penghasilan orang tua sekitar 2.4133; tanggungan orang tua 3.0759 dan memiliki prestasi 0.7967.

Derajat keanggotaan penentuan beasiswa untuk sertiap *cluster* pada iterasi terakhir (iterasi ke-37) dapat dilihat pada Tabel 2. di bawah ini:

Tabel 2. Derajat Keanggotaan Penentuan Beasiswa Untuk Setiap *Cluster* Pada Iterasi Terakhir

No	Derajat Keanggotaan (μ) pada Iterasi Terakhir			Kecenderungan data masuk pada *cluster*		
	(μ_1)	(μ_2)	(μ_3)	C1	C2	C3
1	0.1634	0.1810	0.6557			*
2	0.5724	0.1653	0.2623	*		
3	0.0523	0.0954	0.8522			*
4	0.0467	0.3021	0.6512			*
5	0.9198	0.0345	0.0457	*		
6	0.0643	0.7542	0.1815		*	
7	0.0525	0.2756	0.6719			*
8	0.1063	0.1726	0.7211			*
9	0.0438	0.3300	0.6262			*
10	0.0605	0.1892	0.7503			*
11	0.0507	0.7378	0.2115		*	
12	0.0576	0.2660	0.6763			*
13	0.6670	0.1091	0.2239	*		
14	0.0923	0.0971	0.8107			*
15	0.6497	0.1801	0.1701	*		
16	0.1132	0.6406	0.2463		*	
17	0.6893	0.1642	0.1465	*		
18	0.0108	0.9393	0.0500		*	
19	0.7094	0.0919	0.1987	*		
20	0.0427	0.3048	0.6525			*
21	0.0463	0.7249	0.2288		*	
22	0.0299	0.0572	0.9129			*
23	0.8531	0.0618	0.0852	*		
24	0.3603	0.3797	0.2600		*	
25	0.0327	0.8855	0.0818		*	
26	0.0592	0.1855	0.7553			*

Nilai ini merupakan nilai dari koordinat ketiga titik pusat *cluster* dan memberikan garis besar tiap *cluster* yaitu:

1. Kelompok pertama (*cluster* ke-1), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.2649; memiliki rata-rata penghasilan orang tua sekitar 1.8625; tanggungan orang tua 2.2484 dan memiliki prestasi 4.2701.
2. Kelompok kedua (*cluster* ke-2), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.1653; memiliki rata-rata penghasilan orang tua sekitar 1.6271; tanggungan orang tua 5.1550 dan memiliki prestasi 1.1770.
3. Kelompok ketiga (*cluster* ke-3), berisi data mahasiswa yang memiliki rata-rata ipk sekitar 3.2225; memiliki rata-rata penghasilan orang tua sekitar 2.4133; tanggungan orang tua 3.0759 dan memiliki prestasi 0.7967.

Derajat keanggotaan penentuan beasiswa untuk sertiap *cluster* pada iterasi terakhir (iterasi ke-37) dapat dilihat pada Tabel 3. di bawah ini:

Tabel 3. Derajat Keanggotaan Penentuan Beasiswa Untuk Setiap *Cluster* Pada Iterasi Terakhir

No	Derajat Keanggotaan (μ) pada Iterasi Terakhir			Kecenderungan data masuk pada *cluster*		
	(μ_1)	(μ_2)	(μ_3)	C1	C2	C3
1	0.1634	0.1810	0.6557			*
2	0.5724	0.1653	0.2623	*		
3	0.0523	0.0954	0.8522			*
4	0.0467	0.3021	0.6512			*
5	0.9198	0.0345	0.0457	*		
6	0.0643	0.7542	0.1815		*	
7	0.0525	0.2756	0.6719			*
8	0.1063	0.1726	0.7211			*
9	0.0438	0.3300	0.6262			*
10	0.0605	0.1892	0.7503			*
11	0.0507	0.7378	0.2115		*	
12	0.0576	0.2660	0.6763			*
13	0.6670	0.1091	0.2239	*		
14	0.0923	0.0971	0.8107			*
15	0.6497	0.1801	0.1701	*		
16	0.1132	0.6406	0.2463		*	
17	0.6893	0.1642	0.1465	*		
18	0.0108	0.9393	0.0500		*	
19	0.7094	0.0919	0.1987	*		
20	0.0427	0.3048	0.6525			*
21	0.0463	0.7249	0.2288		*	
22	0.0299	0.0572	0.9129			*
23	0.8531	0.0618	0.0852	*		
24	0.3603	0.3797	0.2600		*	
25	0.0327	0.8855	0.0818		*	
26	0.0592	0.1855	0.7553			*

ISSN:2442-5567 | E-ISSN:2443-289X

1. Kelompok pertama (*cluster* ke-1), akan berisi data mahasiswa ke : 2, 5, 13, 15, 17, 19, 23, 29, 31, 36, 40, 43, 44, 45, 59, 60, 65, 68, 70, 71, 73, 74 dan 75.
2. Kelompok kedua (*cluster* ke-2), akan berisi data mahasiswa ke : 6, 11, 16, 18, 21, 24, 25, 30, 32, 33, 39, 41, 42, 49, 51, 58 dan 72.
3. Kelompok ketiga (*cluster* ke-3), akan berisi data mahasiswa ke : 1, 3, 4, 7, 8, 9, 10, 12, 14, 20, 22, 26, 27, 28, 34, 35, 37, 38, 46, 47, 48, 50, 52, 53, 54, 55, 56, 57, 61, 62, 63, 64, 66, 67 dan 69.

3.2 PEMBAHAAN

Proses *Clustering* memerlukan klasifikasi untuk menentukan kelompok (*cluster*) mana yang berhak untuk menerima beasiswa. Dalam penelitian akan mengelompokkan mahasiswa mejadi tiga (3) *cluster* yaitu :

1. *Cluster* yang menerima beasiswa
2. *Cluster* yang dipertimbangkan menerima beasiswa.
3. *Cluster* yang tidak menerima beasiswa.

Kemudian setiap *cluster* dibagi berdasarkan kriteria mana yang lebih diprioritaskan (berdasarkan IPK atau PO (Penghasilan Orang tua), Tanggungan Orang Tua (TO), Prestasi mahasiswa selama menjadi mahasiswa).

Iterasi pada percobaan ini berhenti pada iterasi ke-37. Hasil akhir *clutering* yang di perolah adalah :

1. *Cluster* pertama memiliki pusat *cluster* (3.2649; 1.8625; 2.2484; 4.2701)
2. *Cluster* kedua memiliki pusat *cluster* (3.1653; 1.6271; 5.1550; 1.1770)
3. *Cluster* ketiga memiliki pusat *cluster* (3.2225; 2.4133; 3.0759; 0.7967)

Hasil klasifikasi *clustering* dapat dilihat pada Tabel 4.

Tabel 4. Hasil Kalasifikasi

Proritas IPK		
Menerima	Dipertimbangkan	Tidak Menerima
Cluster 1	*Cluster* 3	*Cluster* 2
3.2649	3.2225	3.1653
Prioritas TK		
Menerima	Dipertimbangkan	Tidak Menerima
Cluster 3	*Cluster* 1	*Cluster* 2
2.4133	1.8625	1.6271
Prioritas TO		
Menerima	Dipertimbangkan	Tidak Menerima
Cluster 2	*Cluster* 3	*Cluster* 1
5.1550	3.0759	2.2484
Prioritas Prestasi		
Menerima	Dipertimbangkan	Tidak Menerima
Cluster 1	*Cluster* 2	*Cluster* 3
4.2701	1.1770	0.7967

ISSN:2442-5567 | E-ISSN:2443-289X

Hasil klasifikasi mahasiswa berdasarkan Tabel 4 untuk 75 data dapat dilihat pada Tabel 5 di bawah ini.

Tabel 5. Hasil klasifikasi prioritas mahasiswa

No	IPK	PO	TO	Prestasi	*C*	IPK	PO	TO	Prestasi
1	2.81	1.2	2	1	3	DP	DP	M	TM
2	2.85	2	3	3	1	M	TM	DP	M
3	2.84	3.2	3	1	3	DP	DP	M	TM
4	2.87	2.5	4	1	3	DP	DP	M	TM
5	2.87	2.3	2	5	1	M	TM	DP	M
6	2.88	0.8	6	0	2	TM	M	TM	DP
7	2.89	2.8	4	1	3	DP	DP	M	TM
8	2.89	1.5	2	0	3	DP	DP	M	TM
9	2.95	2.3	4	1	3	DP	DP	M	TM
10	2.98	1.5	3	0	3	DP	DP	M	TM
11	3	1	5	0	2	TM	M	TM	DP
12	3	3	4	1	3	DP	DP	M	TM
13	3	1.5	1	3	1	M	TM	DP	M
14	3.02	2.3	2	1	3	DP	DP	M	TM
15	3.03	3	4	5	3	DP	DP	M	TM
16	3.03	3.3	7	1	2	TM	M	TM	DP
17	3.03	2.4	4	5	1	M	TM	DP	M
18	3.04	2	5	1	2	TM	M	TM	DP
19	3.05	2	2	3	1	M	TM	DP	M
20	3.06	2.4	4	1	3	DP	DP	M	TM
21	3.08	2.7	5	1	2	TM	M	TM	DP
22	3.08	3	3	1	3	DP	DP	M	TM
23	3.1	3	2	5	1	M	TM	DP	M
24	3.1	1.3	4	3	2	TM	M	TM	DP
25	3.11	1.2	6	1	2	TM	M	TM	DP
26	3.12	1.5	3	0	3	DP	DP	M	TM
27	3.12	2	4	1	3	DP	DP	M	TM
28	3.13	2	4	0	1	M	TM	DP	M
29	3.13	2	1	5	2	TM	M	TM	DP
30	3.15	1	7	1	2	TM	M	TM	DP
31	3.15	3.3	2	5	1	M	TM	DP	M
32	3.15	0.6	6	1	2	TM	M	TM	DP
33	3.16	2.3	6	3	2	TM	M	TM	DP
34	3.17	2	3	1	3	DP	DP	M	TM
35	3.22	3.3	4	3	3	DP	DP	M	TM
36	3.19	1.3	4	5	1	M	TM	DP	M
37	3.21	1.5	3	1	3	DP	DP	M	TM

ISSN:2442-5567 | E-ISSN:2443-289X

38	3.21	3.3	3	0	3	DP	DP	M	TM
39	3.22	1.8	5	3	2	TM	M	TM	DP
40	3.23	0.5	2	3	1	M	TM	DP	M
41	3.24	0.7	6	1	2	TM	M	TM	DP
42	3.25	1.8	7	3	2	TM	M	TM	DP
43	3.25	0.9	3	3	1	M	TM	DP	M
44	3.27	0.5	2	5	1	M	TM	DP	M
45	3.29	0.9	2	5	1	M	TM	DP	M
46	3.3	1	3	1	3	DP	DP	M	TM
47	3.32	3.8	3	1	3	DP	DP	M	TM
48	3.32	2.3	3	0	3	DP	DP	M	TM
49	3.32	0.9	4	0	2	TM	M	TM	DP
50	3.34	2.4	2	1	3	DP	DP	M	TM
51	3.35	1.7	5	1	2	TM	M	TM	DP
52	3.35	3	2	1	3	DP	DP	M	TM
53	3.35	3	3	1	3	DP	DP	M	TM
54	3.37	2.5	4	0	3	DP	DP	M	TM
55	3.39	2	3	0	3	DP	DP	M	TM
56	3.39	3.5	3	0	3	DP	DP	M	TM
57	3.4	2	4	1	3	DP	DP	M	TM
58	3.32	1	4	1	2	TM	M	TM	DP
59	3.23	2	3	3	2	TM	M	TM	DP
60	3.33	0.4	3	3	1	M	TM	DP	M
61	3.42	3.5	4	1	3	DP	DP	M	TM
62	3.43	3.5	4	1	3	DP	DP	M	TM
63	3.42	2.2	4	1	3	DP	DP	M	TM
64	3.45	1.2	2	0	3	DP	DP	M	TM
65	3.45	0.5	1	3	1	M	TM	DP	M
66	3.47	3	2	0	3	DP	DP	M	TM
67	3.53	3	2	1	3	DP	DP	M	TM
68	3.51	3.2	1	5	1	M	TM	DP	M
69	3.53	3	3	1	3	DP	DP	M	TM
70	3.55	2	3	5	1	M	TM	DP	M
71	3.59	3	2	5	1	M	TM	DP	M
72	3.59	1.6	4	1	2	TM	M	TM	DP
73	3.63	2	2	3	1	M	TM	DP	M
74	3.67	1	2	3	1	M	TM	DP	M
75	3.67	0.4	1	5	1	M	TM	DP	M

Keterangan :
M : menerima.
DP : dipertimbangkan.

TM : tidak menerima.

Berdasarkan tabel prioritas di atas terdapat 3 *cluster* yaitu menerima, dipertimbangkan dan tidak menerima yang di hasilkan dari tabel kalsifikasi.

4. KESIMPULAN DAN SARAN

4.1 KESIMPULAN

Berdasarkan pembahasan maka dapat diberikan beberapa simpulan sebagai berikut:

1. Algoritma *Fuzzy C-means* dapat digunakan untuk mengelompokan data mahasiswa lebih halus dengan menerap- kan besarnya derajat keanggotaan setiap elemen untuk masuk ke dalam kelom- pok-kelompok yang ada.
2. Pengujian data dilakukan sebanyak 37 iterasi, diperoleh tiga kelompok berdasarkan nilai rata-rata penentuan beasiswa, yaitu :
   a. Kelompok pertama (*cluster* ke-1), berisi mahasiswa yang memiliki rata-rata ipk sekitar 3.2649; memiliki rata-rata pendapatan orang tua sekitar 1.8625; tanggungan orang tua 2.2484 dan memiliki prestasi 4.2701.
   b. Kelompok kedua (*cluster* ke-2), berisi mahasiswa yang memiliki rata-rata ipk sekitar 3.1653; memiliki rata-rata pendapatan orang tua sekitar 1.6271; tanggungan orang tua 5.1550 dan memiliki prestasi 1.1770.
   c. Kelompok ketiga (*cluster* ke-3), berisi mahasiswa yang memiliki rata-rata ipk sekitar 3.2225; memiliki rata-rata pendapatan orang tua sekitar 2.4133; tanggungan orang tua 3.0759 dan memiliki prestasi 0.7967.
3. Kemudian setiap *cluster* diklasifikasikan berdasarkan kriteria mana yang lebih diprioritaskan dengan nilai terbesar pada jarak akhir merupakan *cluster* yang menerima beasiswa, sedangkan *cluster* dengan nilai terkecil merupakan *cluster* yang tidak berhak menerima beasiswa.

4.2 SARAN

Saran yang dapat digunakan untuk mendapat tingkat akurasi yang lebih baik dalam penerapan algoritma *Fuzzy C-Means* untuk penentuan beasiswa di Univeristas Megou Pak Tulang Bawang, serta untuk memberikan konstribusi yang lebih besar di dunia riset, disarankan agar hasil penelitian ini dikembangkan dengan cara memodifikasi/*updating* algoritma *Fuzzy C-means* yang digunakan saat ini, atau dengan menggabungkan algoritma *Fuzzy C-means* dengan algoritma lain.

REFERENSI

[1] Ahmad Mauliyadi M, dkk. (2013) , " Perbandingan Metode *Fuzzy C-Means* (Fcm) Dan *Fuzzy Gustafson-Kessel* (Fgk) Menggunakan Data Citra Satelit *Quickbird", Matematika 00 (2013) 01–05,*

[2] Ali Keshavarzi, Fereydoon Sarmadian (2012)," Fuzzy Clustering Analysis For Modeling Of Soil Cation Exchange Capacity", AJAE 3(1):27-33 (2012), ISSN:1836-9448.

[3] Apriansyah Putra, Dinna Yunika Hardiyanti (2011) "Penentuan Penerima Beasiswa Dengan Menggunakan *Fuzzy Multiple Atribute Decission Makin ",Vol. 3, No. 1, April 2011, Halaman 286-293.*

ISSN:2442-5567 | E-ISSN:2443-289X

[4] Bahar, Vincent Suhartono, dkk (2011), "Penentuan Jurusan Sekolah Menengah Atas Dengan Algoritma Fuzzy C-Means", Jurnal Teknologi Informasi, Issn 1414-9999.

[5] Dimas Wahyu Wibowo, dkk (2013) " Perhitungan Jumlah dan Jenis Kendaraan Menggunakan Metode Fuzzy C-means dan Segmentasi Deteksi Tepi Canny" , *Vol. 7, No. 2, Desember 2013, Halaman 103-110.*

[6] Eko Sediyono, dkk (2006), " Penentuan Lokasi Fasilitas Gudang Menggunakan *Fuzzy C – Means* (FCM)", Vol.2, No. 2, Desember 2006: 155 – 166.

[7] Ivan Hardiyanto, dkk (2012), " Implementasi Segmentasi Citra Dengan Menggunakan Metode Generalized Fuzzy Cmeans Clustering Algorithm With Improved Fuzzy Partitions", *Vol. 1, No. 1, (2012) 1-5*

[8] Heru Dibyo Laksono, Abraham Arief, (2012)," Perkiraan Kebutuhan Energi Listrik Jangka Panjang Di Provinsi Sumatera Barat Sampai Tahun 2021 Dengan Logika Fuzzy Clustering ",Vol. 5 No. 2 September 2012, Issn : 2086 – 4981.

[9] Kusumadewi, S., Purnomo, H., (2010), *Aplikasi Fuzzy Untuk Pendukung Keputusan*, Graha Ilmu, Jakarta

[10] Kusrini, (2006), Algoritma Data Mining, Penerbit ANDI, Yogyakarta.

[11] Noor Fitriana Hastuti, (2010), Pemanfaatan Metode *K-Means Clustering* dalam Penentuan Penerima Beasiswa, Jurusan Informatika. Fakultas MIPA. Universitas Sebelas Maret.

[12] Prabowo Pudjo Widodo, dkk (2013), " Penerapan *Data Mining* dengan *Matlab",* Penerbit Rekayasa Sains, Bandung.

[13] Turban, Efraim, et al . (2005). Decision Support Systems and Intelligent Systems. Yogyakarta : Penerbit Andi

ISSN 2229-5518

A review and analysis of fuzzy-c means clustering techniques

Kalema Kiweewa Abdu, Bukenya Faiza,.Rose Alinda Alias
Faculty of Computing, Universiti Tekinologi Malaysia

Abstract-In research, change detection result accuracy does not only depend on the image registration accuracy but also depends on the segmentation and classification accuracy since the errors that arise from image registration are minimised during the segmentation and classification phase. Among the techniques that have been proposed to deal with the misalignment problem that arises from image registration is fuzzy c-means clustering (this technique is a combination of c-means used during grouping pixels in to objects and fuzzy logic used for classification of image objects in to changed and un-changed objects) . Fuzzy c-means clustering technique works by partitioning the image pixels into a set of fuzzy clusters through an iterative optimisation of the objective function adopted from Tariq Rashid, with the update of membership and the cluster centre. Many fuzzy c-means clustering techniques have been proposed by researchers to deal with the image registration errors. This paper reviews some of them and Fuzzy-Statistics-Based Affinity Propagation C-means clustering for segmentation-classification phase is proposed.This is donesince the method is fast, it enables incorporation of spatial information, which leads to prevention of noise as well as the intensity variation in images. It also treats clusters equally which prevents poor solutions caused by hard decisions during clustering, hence avoiding misclassification of clusters especially small clusters, leading to quality segmentation and good classification results.

Index Terms— Image Segmentation, Change Detection, classification, Fuzzy c-means clustering

INTRODUCTION

In change detection, some of the crucial steps that determine the change detection result accuracy are the segmentation-classification phase. In this phase, its where the image registration errors (i.e misalignment problem) are reduced, therefore better techniques are required for better change detection results. According to [11][10][9] true changes can be isolated from false changes if Image object information is manipulated. Researchers have tried to come up with techniques that can minimise the image registration error (ie misalignment problem) but research shows there is still much needed. Among the techniques that have been proposed to be good in dealing with the misalignment problem that arise from image registration is fuzzy c-means clustering [12][13][14]. Fuzzy c-means clustering technique works by partitioning the image pixels into a set of fuzzy clusters through an iterative optimisation of the objective function adopted from Tariq Rashid, with the update of membership and the cluster centre. In this algorithm, centroids are defined for each of the clusters and the quality of results depend on the initial set of clusters, the value of centroid and the number of iteration[8][7]. Many fuzzy c-means clustering techniques have been proposed by researchers to deal with the image registration errors. Section 2.0 of this paper reviews some of the fuzzy c-means clustering and threshold level based fuzzy c means clustering is proposed. Section 3 discusses the findings and also concludes the paper.

2 FUZZY C-MEAN CLUSTERING TECHNIQUES

Among the presented fuzzy clustering techniques, threshold level-based fuzzy clustering is the only fuzzy clustering technique that caters for intensity variation in the images, besides being robust to noise and intra-class dependencies. Threshold level- based fuzzy clustering was applied on the biological image during segmentation of cell, membership connectedness; size-weighted fuzzy clustering and fuzzy statistics-based affinity propagation were applied on multispectral images. Spatially weighted fuzzy c-means (SWFCM) clustering algorithm was applied on synthetic images. The five recently presented fuzzy clustering techniques are described in the subsections that follow.

(I) SPATIALLY WEIGHTED FUZZY C-MEANS (SWFCM) CLUSTERING ALGORITHM

The technique was presented by [3] incorporating the spatial neighbourhood information into the standard FCM clustering algorithm. In this algorithm, k-nearest neighbour (k-NN) algorithm was used to calculate the weight in the SWFCM algorithm in order to enhance the performance of image thresholding. Iterations were carried out with gray level histogram thresholding of the image for a faster segmentation process. According to the author SWFCM is fast and the algorithm is less sensitive to noise since it incorporates spatial information.

But however better segmentation results would be achieved if multi-thresholding was applied instead of single level thresholding.

IJSER

(ii) Threshold level-based algorithm Based on 3 Class Fuzzy C-mean Clustering

The technique was proposed by [5], inorder to segment the cells. In this technique, segments the images (sensed image and reference image) into three classes using FCM clustering thresholding. The threshold is obtained by averaging the maximum in the class with the smallest centre and the minimum in the class with the middle centre. According to the author, the method is robust to noise, prevents weak-edge leakage and has the ability to handle interior intensity variation, also it can used for further analysis but however the method was not tried on othercell types or rather on other applications.

(iii) Membership Connectedness Method

[4] presented this technique in order to utilize spatial information, in this algorithm, a combination of fuzzy connectedness and fuzzy clustering were applied during the functional form of membership connectedness in order to automatically select the required seeds. According to the author the membership connectedness segmentation method considers the local and global spatial relations which leads to reduced sensitivity to noise and false segmentation results, the method also utilizes the expert knowledge in the form of selected seeds and segments the images as desired by the expert which increases the chance of its success.

However the method is faced with the inherent intraclass redundancies in multispectral images caused by difference in the cluster size (the big cluster influence the learning stage which lead to small clusters from being misclassified or even missed during segmentation) which seriously decrease the accuracy and efficiency of the method, it also leads to spatial dependency caused by similarity of the spectral information of neighbouring pixels in the image.

(iv) Size-WeightedFuzzy Clustering

The technique was presented in order to deal with the spatial and the intraclass redundancy that exists in the image faced by the membership connected method, in this unsupervised algorithm, watershed transform was used to deal with the spatial redundancy., Size-weighted fuzzy clustering was also used to deal with intra-class redundancy. During watershed phase, the spatial redundancy is omitted, and in the Membership connectedness construction phase, the spatial relation among image pixels is taken into account. According to the [2], the method can not only detect small regions (that often appear in remote-sensing images) but also it can detect the overlapped regions in the satellite images, although the method does not provide detailed information about the clusters.

(v) Fuzzy Statistics-Based Affinity Propagation (FS-AP)

The technique was presented in order to extract land-cover information in multispectral images, in this algorithm, affinity propagation (a clustering algorithm) was used to find clusters with small error in large datasets and fuzzy statistical similarity measure was used to determine the clusters to which the pixels belong to through obtaining objective estimates of how closely two pixel vectors resemble each other. According to the [6], the method is fast ,robust to noise, and works well in the case of mixed pixels, since it treats all pixels equally which reduces on dependence and avoids poor solutions caused by hard decisions during clustering. However fuzzy statistics-based affinity propagation works well with application of multispectral remote sensing images.

(vi) ANT COLONY OPTIMIZED FUZZY C-MEANS CLUSTERING

The techniques was presented by [14] to obtain obvious edge and segmentation result, in this algorithm the centres and number of clusters are determined by ant colony optimization algorithm, then later initialization fuzzy C-means algorithm is used for remote sensing image classification. According to[14] the method is fast and the visual interpretation of segmentation results is improved but the method is good on land cover/use detection, i.e. Impervious area, water, etc.

2.1 Comparison of Fuzzy C-mean clustering techniques

Table 2 shows comparison of each of the fuzzy clustering methods discussed previously emphasising their advantages, disadvantages, the types of their input they can handle. Among the new presented fuzzy clustering techniques, the threshold level based fuzzy clustering is the only fuzzy clustering technique that has the ability to handle the intensity variations in the image.

Table 2
Comparison of the Fuzzy Clustering Methods

Author	Technique	Objective	Advantage	Disadvantage	input

Yong, Y., Z. Chongxun.2004	Spartially weighted fuzzy c-means clustering	To incorporate spartial neighbourhood information in to standard FCM	-it is fast -it less sensitive to noise	Single level thresholding	Synthetic image
Guanglei Xiong, 2006	Threshold level based algorithm based of fuzzy c-means algorithm	To segment cell	-to prevent weak edge leakage -it is robust to noise -interior intensity variation, incorporates spatial data	-It was not tried on other different cell types	- Drosophila RNAi Fluorescence Cellular Images
Maryam Hasanzadeh and Shohreh Kasaei,2008	Membership Connectedness	To utilise global spatial relations besides the fuzzy clustering stage in order to improve segmentation accuracy.	-reduced sensitivity to noise. -reduced false segmentation results. - can utilize the expert knowledge in the form of selected seeds and segments	- faced with intraclass redundancies. -it also leads to spatial dependency	- multispectral image
M. Hasanzadeh and S. Kasaei, Senior Member, IEEE,2010	Size-weighted fuzzy clustering method.	to deal with the spatial and the intra-class redundancy that exist in the image faced by the membership connected method	-to detect small regions that often appear in remote sensed images -It can detect the overlapped regions in the satellite images.	-the method does not provide detailed information about the clusters	- multispectral image

IJSER

Chen Yang, Bruzzone, L,2010	A Fuzzy-Statistics-Based Affinity Propagation Technique for Clustering	- To extract land-cover information in multispectral images	-it is fast -it is roburst to noise - Reduces on dependence among pixels. - Avoids poor solutions caused by hard decisions during clustering	-works well with application of multispectral remote sensing images. - Does not handle variation in the intensity	-multispectral image
Jingfeng Yan,2014	ant colony optimized fuzzy C-means clustering	to obtain obvious edge and segmentation result.	-improved visual interpretation of segmentation results. -Reduces on the computation time.	- good for land cover/use detection, i.e. impervious area, water, etc.	Middle spatial resolution remote sensing images.

3 DISCUSSION AND CONCLUSION

During segmentation, incorporation of spatial information is an important practice in order to avoid noise, and also to treat clusters equally. In order to avoid poor solutions caused by hard decisions (pixels finding it hard to decide cluster to belong to) during clustering hence avoiding misclassification of clusters especially small clusters leading to quality segmentation and classification results. Subsequently intensity variation (which is as result of image being captured by sensor at different intensity) is an important issue as it may lead to misclassification of the pixels, (water pixel may be treated as non-water pixel) which leads to false segmentation results hence leading to misclassification of object which leads to false changes in the changed image.

According to the table 1, In 2004,[3] incorporated spartial neighbourhood information in to standardize, FCM in order to deal with the reduced sensitivity to noise, the method is fast and less sensitive to noise however, its based on single level thresholding which minimises the chances of getting a better scale for better results, and yet according to [5] a small scale yields better segmentation classification results. The multilevel thresholding technique introduced by [5]enable segmentation of image at many scales which increases the chances of getting a better result. The method also prevents weak edge leakage, interior intensity variation and incorporates spatial data which reduces sensitivity to noise but the method was is good on biological images.

In 2008, [4] reduced sensitivitity to noise by utilising global spatial relations using the

membership connectedness in order to utilize the expert knowledge in the form of selected seeds and segments which reduced sensitivity to noise, reduced false segmentation results, however the method has a problem of intraclass redundancies and spatial dependency. [2], dealt with the spatial and the intra-class redundancy that exist in the image faced by the membership connected method with the use of a Size-weighted fuzzy clustering method. This method is capable of detecting small regions that often appear in remote sensed images and can also detect the overlapped regions in the satellite images but the method does not provide detailed information about the clusters. In the same year, [6],introduced A Fuzzy-Statistics-Based Affinity Propagation Technique for Clustering in order to extract land-cover information in multispectral images, the method avoids poor solutions caused by hard decisions during clustering, reduces dependence among pixels, fast and robust to noise but works well with application of multispectral remote sensing images. And does not handle variation in the intensity which increases sensitivity to noise..Recently,[14] optimized fuzzy C-means clustering in order to extract obvious edge and segmentation result from a SAR image, but the method is good on land cover/use detection.

According to the above comparison, a combination of a Fuzzy-Statistics-Based Affinity Propagation Technique and threshold level based fuzzy C-means clustering would yield better segmentation and classification results, since threshold level based fuzzy c-means clustering enable incorporation of spatial information which leads to prevention of noise and deals with intensity variation in images. Moreover, Fuzzy-Statistics-Based Affinity Propagation technique treats clusters equally which prevents poor solutions caused by hard decisions during clustering, hence avoiding misclassification of clusters especially small clusters leading to quality segmentation and classification results. The novelity of this paper is it e proposes a Fuzzy-Statistics-Based Affinity Propagation C-means clustering for segmentation-classification phase, and that makes up the next stage of this research..

REFERENCES

[1] Richards, J. A. and Jia, X, "Remote Sensing Digital Image Analysis: An Introduction.", 2006. Springer-Verlag, Heidelberg, New York, pp. 67-68; pp. 128-130; pp. 342-352; pp. 243.

[2] Hasanzadeh, M. and S. Kasaei, "A multispectral image segmentation method using size-weighted fuzzy clustering and membership connectedness",2010. IEEE Geoscience and Remote Sensing Letters 7(3): 520-524.

[3] Yong, Y., Z. Chongxun. "A Novel Fuzzy C-Means Clustering Algorithm for Image Thresholding",2004. Measurement Science Review 4(1): 11-19.

[4] Maryam Hasanzadeh and Shohreh Kasaei, "Fuzzy image segmentation using membership connectedness",2008.

[5] Guanglei Xiong, X. Zhou,L. ji,. Bradley.S.Wong, "Automated segmentation of Drosophila RNAi fluorescence cellullar Images using deformable models", IEEE Transactions on circuits andsystems.vol 53. No.11,pp.2415-2424, 2006.

[6] Chen Yang, Bruzzone, L, "A Fuzzy-Statistics-Based Affinity Propagation Technique for clustering in multispectral images", 2010.

[7] Andrew Moore,"A Tutorial on Clustering Algorithms",website: http://www.autonlab.org/tutorials/kmeans.html accessed on24th /09/2011.

[8] HongLei Yang,JunHuan Peng,BaiRu Xia,DingXuan Zhang, "An improved algorithm for remote sensing classification",2013.

[9] Bovolo, F, "A multilevel parcel-based approach to change detection in very high

resolution multitemporal images", 2009. IEEE Geoscience and Remote Sensing Letters 6(1): 33-37

[10] F.Bovolo, CMarin,L. Bruzzone, "A Hierarchical Approach to change detection in very high resolution multitemporal SAR mages", 2013

[11] Pacifici F.;Del Frate,F.;Solimini,C.;Emery,W.j, "An innovative neural net method to detect temporal changes in high resolution optical satellite imagery",2007

[12] Usman seljud,Rashid Hussein, "SAR Image clustering by fuzzy c-means clustering technique with thresholding for icebags",2014.

IJSER

[13] B. Faiza,S. S.Yuhaniz,S. Z. Mohd Hashim,A. K. Kalema, "Detecting Floods Using an Object Based Change Detection Approach",2012

[14] Jingfeng Yan, "Remote sensing image segmentation based on ant colony optimized fuzzy C-means clustering", 2014

IJSER

Kumpulan jurnaL Ilmu Komputer (KLIK)
Volume 01, No.01 September 2014
ISSN: 2406-7857

IMPLEMENTASI METODE FUZZY C-MEANS PADA SISTEM CLUSTERINGDATA VARIETAS PADI

Nurjanah[1], Andi Farmadi[2], Fatma Indriani[2]
[1,2,3] Prodi Ilmu Komputer Fakultas MIPA Universitas Lambung Mangkurat

Jl. A. Yani Km 36 Banjarbaru, Kalimantan Selatan
Email: nurjanah_abs@yahoo.co.id

Abstract

Mutations with gamma rays conducted on five local rice varieties tidal South Kalimantan produce a lot of data availability. In order for these data not only become a graveyard of useless data required a method that could be used to probe the hidden information from the data. The method known as data mining. Data mining is a technique to gain knowledge from the data by looking for certain patterns or rules of a number of large amounts of data. One method of data mining is clustering, where clustering is usually used to group objects that are similar in the same class or segment. By utilizing the data of local rice varieties tidal South Kalimantan mutated by gamma rays, data mining process is done by grouping the data based on the harvest age, productive tillers, and weight of 1000 seeds into 4 groups using fuzzy c-means algorithm. From that cluster information, carried ranking using the Simple Additive Weighting method and acquired knowledge about improved varieties by harvest age, productive tillers, and a weight of 1000 is kuatek with a dose of 30 krad.

Keywords *: Data Mining, Cluster, Fuzzy C-Means, Local Rice Varieties*

Abstrak

Mutasi dengan sinar gamma yang dilakukan terhadap lima varietas padi lokal pasang surut kalimantan selatan menghasilkan tersedianya banyak data. Agar data-data tersebut tidak hanya menjadi kuburan data yang tidak berguna dibutuhkan sebuah metode yang bisa digunakan untuk menggali informasi–informasi tersembunyi dari data tersebut. Metode tersebut dikenal dengan data mining. Data mining merupakan suatu teknik untuk menggali pengetahuan dari data dengan mencari pola atau aturan tertentu dari sejumlah data dalam jumlah besar. Salah satu metode data mining adalah klastering, dimana klastering biasanya digunakan untuk mengelompokan objek-objek yang memiliki kemiripan dalam kelas atau segmen yang sama. Dengan memanfaatkan data varietas padi hasil mutasi dengan sinar gamma dilakukan proses penggalian data dengan cara mengelompokkan data umur panen, anakan produktif, dan bobot 1000 menjadi 4 klaster menggunakan algoritma fuzzy c-means. Dari informasi 4 klaster tersebut, dilakukan perangkingan klaster menggunakan metode Simple Additive Weighting dan diperoleh pengetahuan tentang varietas unggul berdasarkan umur panen, anakan produktif, dan bobot 1000 yaitu kuatek dengan dosis 30 krad.

Kata kunci *: Data Mining, Klaster, Fuzzy C-Means, Varietas Padi Lokal*

1. PENDAHULUAN

Varietas padi lokal Kalimantan Selatan masih menjadi pilihan para petani untuk ditanam, yang dalam penanamannya tidak terlalu bergantung pada input pertanian. Selain itu, varietas padi lokal juga bersifat adaptif dengan lingkungan Kalimantan Selatan yang didominasi oleh lahan rawa, meskipun berumur panjang dan produktivitas relatif rendah daripada varietas padi unggul. Akan tetapi sifat "pera" nasi yang dihasilkan membuat harga jual varietas padi lokal lebih tinggi dibandingkan dengan varietas padi unggul [3].

Perjuangan dalam peningkatan produksi pangan untuk memenuhi kebutuhan masyarakat Kalimantan Selatan merupakan tantangan bagi masyarakat pertanian yang dapat melahirkan peluang berupa pengembangan teknologi alternatif untuk menerobos kebuntuan dan stagnasi produksi beras lokal. Alternatif itu dapat diarahkan pada pengembangan varietas lokal yang dapat beradaptasi pada lingkungan produksi yang buruk dengan produktivitas yang tinggi dan umur panen yang pendek. Peningkatan keragaman populasi dapat dilakukan salah satunya melalui induksi mutasi secara fisik dengan iradiasi sinar gamma.

Salah satu upaya untuk meningkatkan produksi pangan varietas padi lokal pasang surut kalimantan selatan yaitu dengan cara meningkatkan keragaman populasi, dilakukan radiasi dengan sinar gamma terhadap lima varietas padi. Dari upaya tersebut dihasilkan banyak data salah satunya yaitu data varietas padi hasil mutasi dengan sinar gammamaka dibutuhkan sebuah metode yang bisa digunakan untuk menggali informasi–informasi tersembunyi dari data tersebut. Metode tersebut dikenal dengan *data mining*.

Data mining adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar. Salah satu tugas dalam data mining adalah klastering. Tujuan utama dari klastering adalah pengelompokan sejumlah data/obyek ke dalam klaster sehingga dalam setiap klaster akan berisi data yang semirip mungkin [1].

Penggalian suatu data menjadi pengetahuan adalah kegiatan yang sangat berguna, karena pengetahuan tersebut bisa digunakan untuk memecahkan suatu masalah yang ada. *Data mining* merupakan salah satu teknik untuk menggali pengetahuan dari data. Agar data-data tersebut tidak menjadi kuburan data yang tidak bergunamaka dibuat sistem yang mampu menggali data varietas padi yang dimutasi dengan sinar gamma yang dapat mengorganisir data varietas padi yang dimutasi dengan sinar gamma dalam suatu database serta dapat menghasilkan pengelompokan sejumlah data/obyek ke dalam klaster yang memberikan pengetahuan mengenai varietas unggulmenggunakan metode *Fuzzy C-Means* (FCM).

2. METODOLOGI PENELITIAN

2.1. Analisa Data

Pada tahap ini akan dilakukan analisis data dengan KDD (*knowledge Discovery in Databases*). Keseluruhan proses KDD dari mulai memasukan data mengalami 3 tahap secara global sebelum menjadi hasil yang berupa informasi, yaitu:

(a) *Preprocessing*

Pada tahap ini struktur basis data akan dipersiapkan sehingga mempermudah proses mining. Proses *preprocessing* ini mencakup tiga hal utama yaitu :

1) seleksi data
   untuk memilih data yang akan digunakan dalam proses data mining. Dalam penelitian ini dipilih 3 data varietas padi yaitu umur panen, anakan produktif, dan bobot 1000 biji;
2) *Cleaning*
   Memastikan kualitas data yang telah dipilih pada tahap seleksi data, pada tahap ini masalah yang harus dihadapi adalah noisy data dan missing values. Proses cleaning data perlu dilakukan agar data bersih dari duplikasi data, data yang tidak konsisten, atau kesalahan cetak. Sehingga data yang telah melewati proses ini siap untuk di proses di data mining. Pada penelitian ini proses pembersihan data hanya dilakukan pada data yang missing value;
3) integrasi data
   Setelah melakukan pembersihan data yang menyeluruh, maka akan dilakukan tahapan selanjutnya yaitu integrasi data, karena data lebih dari satu tabel maka dilakukan tahap penggabungan data dari berbagai tabel. Kemudian data padi diberi kode agar tidak ada data yang memiliki nama yang sama. Kode padi terdiri dari nomor padi, baris padi, dosis radiasi, dan varietas padi.

(b) Data Mining

Proses data mining yang dilakukan pada penelitian ini adalah menerapkan teknik *clustering* dengan algoritma *Fuzzy C-Means* untuk mendapatkan karakteristik dari data. *Cluster* yang digunakan pada penelitian ini sebanyak empat klaster.

Konsep dasar FCM (*Fuzzy C-Means*) pertama kali adalah menentukan pusat *cluster* yang akan menandai lokasi rata-rata untuk tiap *cluster*. Pada kondisi awal, pusat *cluster* ini masih belum akurat. Tiap-tiap titik data memiliki derajat keanggotan untuk tiap-tiap *cluster*. Dengan cara memperbaiki pusat *cluster* dan derajat keanggotaan tiap-tiap titik data secara berulang, maka akan dapat dilihat bahwa pusat *cluster* akan bergerak menuju lokasi yang tepat. Perulangan ini didasarkan pada derajat keanggotaan yang menggambarkan jarak dari titik data yang diberikan ke pusat *cluster* yang terbobot oleh derajat keanggotaan titik data tersebut [6].

Pada perhitungan menggunakan FCM dengan menggunakan data yang sama tetapi diolah dengan jumlah *cluster* yang berbeda, maka hasil pengelompokannya akan sedikit berbeda, karena data tidak diolah dengan satu variabel saja tetapi dengan semua variabel. Perbedaan hasil pengelompokan itu dikarenakan data pada kelompok tertentu kemungkinan akan berpindah pada kelompok lain jika diolah dengan jumlah *cluster* yang berbeda, ini menunjukkan bahwa sistem aplikasi sudah berjalan dengan benar [6]

Output dari FCM bukan merupakan fuzzy inference sistem, namun merupakan derajat pusat *cluster* dan beberapa derajat keanggotaan untuk

tiap-tiap titik data. Algoritma *Fuzzy C-Means* disusun dengan langkah sebagai [6]:

(1) Input data
input data yang akan di*cluster*yaitu data varietas padi (X) berupa matrik berukuran n × m (n = jumlah data, m = atribut setiap data). X_{ij} = data ke- i (i = 1, 2, …, n), atribut ke- j (j = 1, 2, …, m).

(2) Batasan
(a) Jumlah *cluster* = c = 4
(b) Pangkat = w = 2
(c) Maksimum iterasi = Maxit = 100
(d) Error terkecil yang diharapkan = ξ = 10^{-3}
(e) Fungsi obyektif awal = Po = 0
(f) Iterasi awal = t = 1

(3) Membangkitkan bilangan random μ_{ik}, i = 1, 2, …, n; k = 1, 2, …,c; sebagai elemen-elemen matrik partisi awal U, dengan jumlah setiap nilai elemen kolom dalam satu baris adalah 1 (satu)

$$\sum_{i=1}^{c} \mu_{ci} = 1 \qquad \dots (1)$$

(4) Menghitung pusat *cluster* ke- k : V_{kj} dengan k = 1, 2, …, c; dan j = 1, 2, …,m.

$$V_{kj} = \frac{\sum_{i=1}^{n}\left((\mu_{ik})^{w} * X_{ij}\right)}{\sum_{i-1}^{n}(\mu_{ik})^{w}} \qquad \dots (2)$$

(5) Menghitung fungsi obyektif pada iterasi ke- t, P_t :

$$P_t = \sum_{i=1}^{n}\sum_{k=1}^{c}\left(\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right](\mu_{ik})^{w}\right) \qquad \dots (3)$$

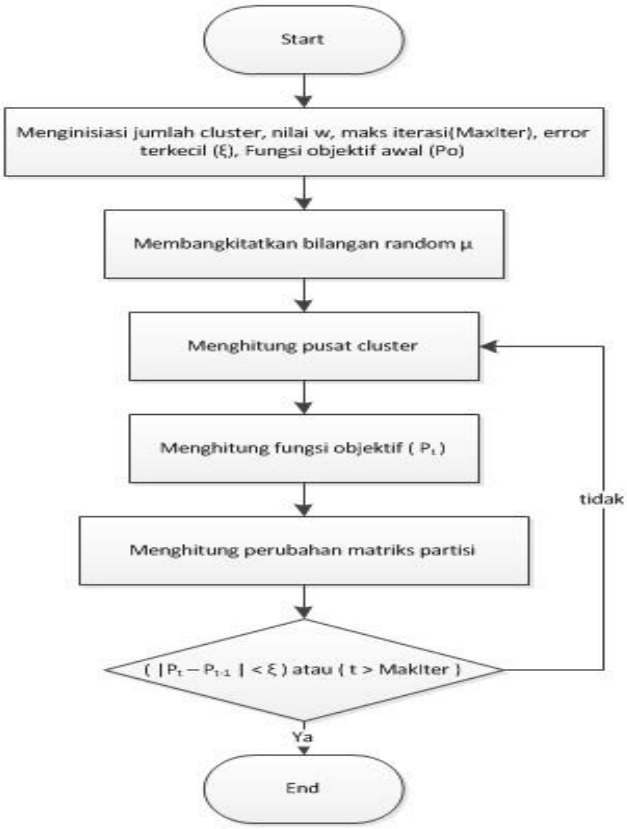
(6) Menghitung perubahan matrik partisi:

$$\mu_{ik} = \frac{\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right]^{\frac{-1}{w-1}}} \dots (4)$$

dengan i = 1, 2, …, n; dan k = 1, 2, …, c

(7) Mengecek kondisi berhenti:
1. Jika : ($|P_t - P_{t-1}| < \xi$) atau (t > MakIter) maka berhenti;
2. Jika tidak : t = t + 1, ulangi langkah ke- d (menghitung V_{kj}).

Bagan alir untuk metode *Fuzzy C-Means* yaitu sebagai berikut :



Gambar 1. Bagan Alir Metode *Fuzzy C-Means*

(c) ***Postprocessing***

Postprocessing bertujuan untuk membantu pengguna dalam memahami informasi, Karena pusat *cluster* masih tidak bisa dienterpretasikan secara langsung, maka diperlukan metode *Simple Additive Weighting* yang bisa melakukan perangkingan nilai pusat *cluster* terakhir tersebut.

Metode SAW (*Simple Additive Weighting*) sering juga dikenal istilah metode penjumlahan terbobot. Konsep dasar metode SAW adalah mencari penjumlahan terbobot dari rating kinerja pada setiap alternatif pada semua atribut. Langkah-langkah metode dalam metode SAW adalah [7] :

1. Membuat matriks keputusan Z berukuran m x n, dimana m = alternatif yang akan dipilih dan n = kriteria .
2. Memberikan nilai x setiap alternatif (i) pada setiap kriteria (j) yang sudah ditentukan, dimana, i=1,2,…m dan j=1,2,…n pada matriks keputusan Z,

$$\mathbf{Z} = \begin{vmatrix} x_{11} & x_{y12} & \dots & x_{1j} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ x_{i1} & x_{i2} & & x_{ij} \end{vmatrix} \dots (5)$$

3. Memberikan nilai bobot preferensi (W) oleh pengambil keputusan untuk masing-masing kriteria yang sudah ditentukan.

$$W = [W_1\ W_2 W_3 \dots\ W_j] \dots(6)$$

4. Melakukan normalisasi matriks keputusan Z dengan cara menghitung nilai rating kinerja ternormalisasi (r_{ij}) dari alternatif A_i pada atribut C_j

$$r_{ij} = \begin{cases} \frac{x_{ij}}{\underset{i}{MAX}(x_{ij})} & \text{jika j adalah atribut keuntungan} \\ \frac{\underset{i}{MIN}(x_{ij})}{x_{ij}} & \text{jika j adalah atribut biaya} \end{cases} \dots(7)$$

Dengan ketentuan :

a. Dikatakan atribut keuntungan apabila atribut banyak memberikan keuntungan bagi pengambil keputusan, sedangkan atribut biaya merupakan atribut yang banyak memberikan pengeluaran jika nilainya semakin besar bagi pengambil keputusan.
b. Apabila berupa atribut keuntungan maka nilai (x_{ij}) dari setiap kolom atribut dibagi dengan nilai (MAX x_{ij}) dari tiap kolom, sedangkan untuk atribut biaya, nilai (MIN x_{ij}) dari tiap kolom atribut dibagi dengan nilai (x_{ij}) setiap kolom.

5. Hasil dari nilai rating kinerja ternormalisasi (r_{ij}) membentuk matriks ternormalisasi (N)

$$\mathbf{N} = \begin{vmatrix} r_{11} & r_{y12} & \dots & r_{1j} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ r_{i1} & r_{i2} & & r_{ij} \end{vmatrix} \dots (8)$$

6. Melakukan proses perankingan dengan cara mengalikan matriks ternormalisasi (N) dengan nilai bobot preferensi (W).
7. Menentukan nilai preferensi untuk setiap alternatif (V_i) dengan cara menjumlahkan hasil kali antara matriks ternormalisasi (N) dengan nilai bobot preferensi (W).

$$V_i = \sum_{j=1}^{n} w_j r_{ij} \dots (9)$$

Nilai Vi yang lebih besar mengindikasikan bahwa alternatif Ai merupakan alternatif terbaik.

2.3. Implementasi

Implementasi dilakukan dengan metode Fuzzy C-Means (FCM) pada sistem clustering data varietas padilokal pasang surut kalimantan selatan yang dimutasi dengan sinar gamma menggunakan delphi XE2

2.4. Pengujian

Untuk memastikan apakah perhitungan dengan metode fuzzy c-means sudah benar dilakukan uji coba perhitungan secara manual meggunakan MS. Excel dan Matlab.Pengujian dilakukan untuk menghindari kesalahan dari sistem yang dibuat. Apabila terjadi kesalahan, maka sistem akan diperbaiki kembali sampai hasil proses sesuai dengan apa yang diharapkan.

3. HASIL DAN PEMBAHASAN

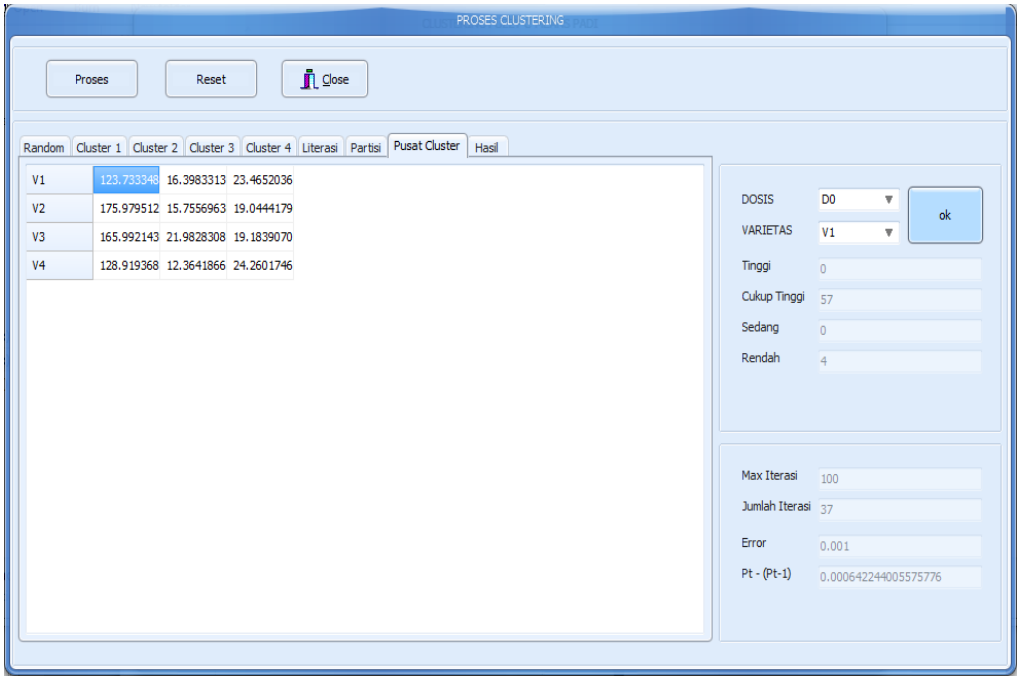
Dariproses clustering Fuzzy C-Means dengan batasan jumlah cluster sebanyak 4 cluster, pangkat 2, masimal iterasi 100, error 0,001 proses berhenti pada iterasi ke-37 dengan error 0,000642 dan diperoleh pusat cluster iterasi terakhir yaitu pada Tabel 1 atau pada Gambar 2 dengan menggunakan rumus dibawah ini (persamaan 2):

$$V_{kj} = \frac{\sum_{i=1}^{2420}\left((\mu_{ik})^2 * X_{ij}\right)}{\sum_{i=1}^{2420}(\mu_{ik})^2}$$

Keterangan :
V_{kj}: pusat cluster dengan k=1,2,...,c dan j = 1,2,...,n
X_{ij} : data ke-i (i=1,2,...m), atribut ke-j (j=1,2,...n)
μ_{ik} : bilangan acak dengan i=1,2,...m dan k=1,2,...c

Tabel 1. Pusat Cluster Delphi

	Umur Panen	Anakan Produktif	Bobot 1000
Klaster 1	165.992143	21.9828308	19.1839070
Klaster 2	123.733348	16.3983313	23.4652036
Klaster 3	128.919368	12.3641866	24.2601746
Klaster 4	175.979512	15.7556963	19.0444179



Gambar 2. Form Proses Fuzzy C-Means

Dari Tabel 1 dapat dilihat bahwa klaster 1 memiliki umur panen terendah ketiga, anakan produktif terbanyak pertama, dan bobot 1000 terberat ketiga. Klaster 2 memiliki umur panen terendah pertama, anakan produktif terbanyak kedua, dan bobot 1000 terberat kedua. Klaster 3 memiliki umur panen terendah kedua, anakan produktif terbanyak keempat, dan bobot 1000 terberat pertama.

Klaster 4 memiliki umur panen terendah keempat, anakan produktif terbanyak ketiga, dan bobot 1000 terberat keempat.

Dilihat dari umur panen klaster 2 lebih unggul karena memiliki umur panen terendah dibandingkan klaster 1, klaster 3, dan klaster 4. Untuk anakan produktif klaster 1 lebih unggul karena memiliki anakan produktif terbanyak. Dan untuk bobot 1000 klaster 3 lebih unggul karena memiliki bobot seribu terberat apabila dibandingkan dengan klaster lainnya.

Dari hasil pusat klaster (Tabel 1) tersebut dilakukan perangkingan dengan metode *Simple Additive Weighting* (SAW) untuk melihat klaster mana yang memiliki umur panen, anakan produktif dan bobot 1000 terbaik. Perhitungan nilai prefrensi dilakukan menggunakan rumus dibawah ini (persamaam 9) :

$$V_i = \sum_{j=1}^{n} w_j r_{ij}$$

Keterangan :
V_i : nilaipreferensiuntuksetiapalternatif
w_j : nilai bobot preferensi
r_{ij} : matriks ternormalisasi

Berikut tabel nilai prefrensi dan tabel hasil pengelompokan per varietas :

Tabel 2. Nilai Prefensi Klaster

Klaster	Nila Prefensi	Kategori
Klaster 1	2.535898966	Cukup Tinggi
Klaster 2	2.712760587	Tinggi
Klaster 3	2.53314209	Sedang
Klaster 4	2.209208036	Rendah

Tabel 3. Hasil Pengelompokan Varietas Siam Harli

Kategori	Varietas Siam Harli (V1)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	-	-	-	46
Cukup Tinggi	57	77	32	-
Sedang	-	-	-	44
Rendah	4	2	3	-

Tabel 4. Hasil Pengelompokan Varietas Siam Unus

Kategori	Varietas Siam Unus (V2)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	-	1	-	-
Cukup Tinggi	48	205	48	86
Sedang	-	1	4	-
Rendah	4	-	5	8

Tabel 5. Hasil Pengelompokan Varietas Kuatek

Kategori	Varietas Kuatek (V3)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	-	-	-	371
Cukup Tinggi	7	152	18	176
Sedang	-	-	-	-
Rendah	41	5	132	-

Tabel 6. Hasil Pengelompokan Varietas Siam Sebelas

Kategori	Varietas Siam Sebelas (V4)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	5	-	-	25
Cukup Tinggi	1	28	9	-
Sedang	1	-	-	22
Rendah	25	158	78	-

Tabel 7. Hasil Pengelompokan Varietas Siam Gumpal

Kategori	Varietas Siam Gumpal (V5)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	-	-	-	13
Cukup Tinggi	-	-	-	-
Sedang	-	-	-	116
Rendah	54	173	105	-

Pengetahuan yang didapat dari hasil penambangan data/klastering ini berupa varietas unggul yaitu varietas yang masuk kategori tinggi. Varietas yang masuk kategori tinggi terdiri atas rata-rata umur panen sebesar 123,7665737; rata-rata anakan produktif sebesar 16,2834365; dan rata-rata bobot 1000 sebesar 23,46984343 dengan nilai preferensi sebesar 2,712760587. Diantara kelima varietas tersebut yang paling banyak masuk kategori tinggi yaitu varietas kuatek dengan dosis 30 krad yakni sebanyak 371 data padi atau sebesar 67,82 % data padi masuk kategori tinggi

Tabel 8. Presentase Data Padi Yang Termasuk Kategori Tinggi

Varietas	Dosis			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Siam Harli (V1)	-	-	-	51,11%
Siam Unus (V2)	-	0,93%	-	-
Kuatek (V3)	-	-	-	67,82%
Siam Sebelas (V4)	15,61%	-	-	53,19%
Siam Gumpal (V5)	-	-	-	10,08%

4. KESIMPULAN

Dari hasil penelitian dan pengamatan dari sistem yang telah dibuat, maka dapat diperoleh kesimpulan sebagai berikut :

1. Metode Fuzzy C-Means berhasil diimplemantasikan pada clustering data varietas padi lokal pasang surut kalimantan selatan yang dimutasi dengan sinar gamma.
2. Dilihat dari umur panen klaster 2 lebih unggul karena memiliki umur panen terendah, untuk anakan produktif klaster 1 lebih unggul karena memiliki anakan produktif terbanyak, dan untuk bobot 1000 klaster 3 lebih unggul karena memiliki bobot seribu terberat apabila dibandingkan dengan klaster lainnya.
3. Pengetahuan yang didapat mengenai varietas unggul berdasarkan umur panen, anakan produktif dan bobot 1000 terbaik menggunakan SAW memiliki nilai preferensi sebesar 23,46984343 yaitu kategori tinggi.
4. Diantara kelima varietas yang paling banyak masuk kategori tinggi yaitu varietas kuatek dengan dosis 30 krad yakni sebanyak 371 data padi atau sebesar 67,82 % data padi masuk kategori tinggi.

DAFTAR PUSTAKA

[1]. Budi Santosa, "**Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis***",* Graha Ilmu, Yogyakarta, 2007.

[2]. Iko Pramudiono. "***Pengantar Data Mining : Menambang Permata Pengetahuan di Gunung Data***" 2003.

[3]. Noor, M., Y. Rina dan Noorginayuwati, "**Persepsi petani tentang lahan gambut dan pengelolaannya",** Laporan Penelitian, Balai Penelitian Pertanian Lahan Rawa, Banjarbaru, 2007.

[4]. Nurjanah, "**Clustering Varietas Padi Lokal Pasang Surut Kalimantan Selatan Yang Dimutasi Dengan Sinar Gamma Menggunakan Metode Fuzzy C-Means**" Naskah Skripsi Universitas Lambung Mangkurat, Banjarbaru, 2014.

[5]. Nuqson Masykur Huda, "***Aplikasi Data Mining Untuk Menampilkan Informasi Tingkat Kelulusan Mahasiswa (Studi Kasus Di Fakultas MIPA Universitas Diponegoro)",*** Naskah Skripsi Universitas Diponegoro, Semarang, 2010.

[6]. Sri Kusumadewi and Hari Purnomo, "**Aplikasi Logika Fuzzy Untuk Pendukung Keputusan***",* Yogyakarta, Graha Ilmu, 2010.

[7]. Wibowo S, Henry., Amalia, Riska., Fadlun M, Andi., Arivanty, Kurnia, "**Sistem Pendukung Keputusan Untuk MenentukanPenerima Beasiswa Bank BRI Menggunakan FMADM (Studi Kasus: Mahasiswa Fakultas Teknologi Industri Unversitas Islam Indonesia)***",* Seminar Nasional Aplikasi Teknologi Informasi, Yogyakarta, 2008.

International Journal of Database Theory and Application
Vol.6, No.6 (2013), pp.1-18
http://dx.doi.org/10.14257/ijdta.2013.6.6.01

Implementation of the Fuzzy C-Means Clustering Algorithm in Meteorological Data

Yinghua Lu[1], Tinghuai Ma[1,*], Changhong Yin[2], Xiaoyu Xie[2], Wei Tian[1] and ShuiMing Zhong[1]

[1]*School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing 210044;*
[2]*Wuhan Meteorological Service, Wuhan 430040*

cloveryunyan@163.com

Abstract

An improved fuzzy c-means algorithm is put forward and applied to deal with meteorological data on top of the traditional fuzzy c-means algorithm. The proposed algorithm improves the classical fuzzy c-means algorithm (FCM) by adopting a novel strategy for selecting the initial cluster centers, to solve the problem that the traditional fuzzy c-means (FCM) clustering algorithm has difficulty in selecting the initial cluster centers. Furthermore, this paper introduces the features and the mining process of the open source data mining platform WEKA, while it doesn't implement the FCM algorithm. Considering this shortcoming of WEKA, we successfully implement the FCM algorithm and the advanced FCM algorithm taking advantage of the basic classes in WEKA. Finally, the experimental clustering results of meteorological data are given, which can exactly prove that our proposed algorithm will generate better clustering results than those of the K-Means algorithm and the traditional FCM algorithm.

Keywords: *Fuzzy C-Means algorithm; WEKA; Meteorological Data*

1. Introduction

Clustering analysis plays an important role in the data mining field, it is a method of clustering objects or patterns into several groups. It attempts to organize unlabeled input objects into clusters or "natural groups" such that data points within a cluster are more similar to each other than those belonging to different clusters, *i.e.*, to maximize the intra-cluster similarity while minimizing the inter-cluster similarity. In the field of clustering analysis, a number of methods have been put forward and many successful applications have been reported.

Clustering algorithms can be loosely categorized into the following categories: hierarchical, partition-based, density-based, grid-based and model-based clustering algorithms [1-4]. Among them, partition-based algorithms which partition objects with some membership matrices are most widely studied. Traditional partition-based clustering methods usually are deterministic clustering methods which usually obtain the specific group which objects belong to, *i.e.*, membership functions of these methods take on a value of 0 or 1. We can accurately know which group that the observation object pertains to. This characteristic brings about these clustering methods' common drawback, that we can not clearly know the

* Corresponding Author, email: thma@nuist.edu.cn

ISSN: 2005-4270 IJDTA
Copyright ⓒ 2013 SERSC

probability of the observation object being a part of different groups, which reduces the effectiveness of hard clustering methods in many real situations. For this purpose, fuzzy clustering methods which incorporate fuzzy set theory [5] have emerged. Fuzzy clustering methods [6-8] quantitatively determine the affinities of different objects with mathematical methods, described by a member function, to divide types objectively.

Among the fuzzy clustering method, the fuzzy c-means (FCM) algorithm [9] is the most well-known method because it has the advantage of robustness for ambiguity and maintains much more information than any hard clustering methods. The algorithm is an extension of the classical and the crisp k-means clustering method in fuzzy set domain. It is widely studied and applied in pattern recognition, image segmentation and image clustering [10-12], data mining [13], wireless sensor network [14] and so on.

WEKA (Waikato Environment for Knowledge Analysis) based on JAVA environment is a free, non-commercial and open-source platform aiming at machine learning and data mining. In WEKA, it implements several famous data mining algorithms. Users can call the appropriate algorithms according to their various purposes. However, the FCM algorithm is not integrated into WEKA.

In this paper, we implement the FCM algorithm and successfully integrate it into WEKA to expand the system functions of the open-source platform, so that users can directly call the FCM algorithm to do fuzzy clustering analysis. Besides, considering the shortcoming of the classical FCM algorithm in selecting the initial cluster centers, we represent an improved FCM algorithm which adopts a new strategy to optimize the selection of original cluster centers.

The structure of this paper is as follows. In the next section, we start a brief review of WEKA and the FCM algorithm. Section 3 describes the main ideas of the traditional FCM algorithm. In Section 4, we present our proposed algorithm based on the traditional FCM algorithm. Experiments results on meteorological data will be shown in Section 5. Finally, conclusions and future work are summarized.

2. Related Work

2.1. WEKA

The full name of WEKA is Waikato Environment for Knowledge Analysis and WEKA is also the name of a kind of birds which come from New Zealand. The package of WEKA can be downloaded from the website of Waikato University in New Zealand (http://www.cs.waikato.ac.nz/ml/weka/), the latest version number is 3.7.2.

2.1.1. The Data Mining Process in WEKA:

(1) Data input

According to different users' demands, WEKA provides three types of methods of inputting data sets: files, the web sites and databases. Note that the format of files is limited when we input data from existing files. There are three main formats of files: Arff data files (*.arff), C4.5 data files (*.names or *.data), CSV data files (*.csv). When we do clustering analysis or association analysis, Arff data files are the best choice. Meanwhile, the recognized attribute types are also restricted including numeric type, nominal type, string type, date type and relational type.

(2) Data preprocess

Data preprocess is also called data filter. Objects doing data preprocessing can be

 Copyright ⓒ 2013 SERSC

divided into two categories, attributes and instances. There are also two methods for data preprocessing, supervised methods and unsupervised methods.

Among so many methods of preprocessing different types of data, what we frequently use in this stage is: missing values processing, standardization processing, normalization processing and discrete processing for attributes.

1) missing values processing: the corresponding class is weak.filters.unsupervised.attribute.ReplaceMissingValues. For numeric attributes, WEKA replaces the missing values with the average value. For nominal attributes, WEKA replaces the missing values with the mode which occurs most frequently.

2) standardization processing: the corresponding class is weka.filters.unsupervised.attribute.Standardize. It is just applicable to numeric attributes. After the standardization processing, all values of numeric attributes form a normal distribution.

3) normalization processing: the corresponding class is weka.filters.unsupervised.attribute.Normalize. It is also limited to numeric attributes. We can normalize numeric values into any interval taking advantage of zoom and pan parameters. In default the resulting values are restricted in the interval [0, 1].

4) discrete processing: the corresponding class is weka.filters.unsupervised.attribute.Discretize and weka.filters.supervised.attribute.Discretize. These two categories respectively discretize numeric attributes in supervised and unsupervised ways.

(3) Data mining

The data mining process consists of classification, clustering analysis, association rules and other pattern analysis. WEKA almost implements all frequently used algorithms for mining different patterns. Here we illustrate the process of classification and clustering analysis.

The process of classification in WEKA is as follows:

1) Input training and test samples;

2) Initialize the classifiers;

3) Use training samples to train classifiers;

4) Test the performance of classifiers making use of test samples;

5) Output the classification results.

The process of clustering analysis in WEKA is as follows:

1) Read samples which need to predict;

2) Initialize the clustering algorithms and set parameters;

3) Cluster the samples using algorithms;

4) Output the clustering results.

(4) Visualization

Generally simple data results may not satisfy users' demands, sometimes we need to vividly describe the changing trend of data results. Hence visualization appears.

Copyright ⓒ 2013 SERSC

Visualization makes the data mining process and results of data mining visualize. It is a useful tool for the mining process and improves the efficiency of the mining process.

2.1.2. The Comparison between WEKA and Other Data Mining Platforms: Nowadays data mining is still an emerging field, and is closely associated with other fields like statistics, machine learning and artificial intelligence. Recently, more and more data mining platforms are appearing. Here we describe some frequently used data mining platforms and discuss their advantages and disadvantages.

(1) Intelligent Miner

Intelligent Miner developed by IBM is the data mining software, consisting of Intelligent Miner for Data and Intelligent Miner for Text, that can extract useful information from both databases and texts.

Intelligent Miner for Data can extract implicit information from databases, data warehouses and data centre, and can find patterns from traditional databases or ordinary files taking advantage of structural data. It is widely used in market analysis, fraud monitoring and customer contact management.

Intelligent Miner for Text allows enterprises to execute the data mining process from texts, here texts can be text files, web pages, e-mails and so on.

Intelligent Miner has the capability of linguistic analysis and the ability of aggregation and filtering. It can deal with the huge amount of data and support parallel processing. Nevertheless, its GUI may not be friendly for users, and users need to be familiar with UNIX.

(2) Enterprise Miner

Enterprise Miner from SAS is a common data mining tool. The data mining process in Enterprise Miner is usually in the following order: sampling-exploration-conversion-modeling-assessment. It can be integrated with SAS data warehouse and OLAP, and can implement the end to end knowledge discovery process of inputting data, analyzing data and obtaining results.

Compared with Intelligent Miner, Enterprise Miner provides graphical interfaces and visual operations for beginners. However, it desires more space requirements to store temporary files and is difficult to output the decision tree.

(3) SPSS Clementine

As an open data mining tool, SPSS Clementine not only supports the overall data mining process, from data acquisition, transformation, modeling, evaluation to final deployment, but it also supports the industry standard (CRISP-DM) in the field of data mining. Generally speaking, SPSS Clementine pays more attention to solving the problem itself instead of doing some technical work.

(4) Darwin

Darwin developed by Oracle Corporation supports several data mining algorithms, like neural networks, classification, regression, K-nearest neighbor algorithm (KNN), genetic algorithms and so on.

Darwin has the following three important advantages. Firstly, the parallel processing of data mining algorithms accelerates Darwin while dealing with huge amounts of data. Secondly, its simplicity of extracting patterns is beneficial for incorporating itself with other applications. Finally, GUI with the windows style is friendly for clients to use.

 Copyright ⓒ 2013 SERSC

(5)WEKA

WEKA is the most well known open-source machine learning and data mining software. WEKA provides a friendly interactive interface for ordinary users who need simple data analysis, while for the researchers who research the theory of data mining, WEKA provides open-source methods for learning and implementation. For instance, when researchers want to compare the performance between the existing algorithms and their own algorithms, what they need to do is to achieve their own algorithms without considering the realization of all comparison algorithms that have been implemented in WEKA [15].

Nevertheless, comparing to its high performance in machine learning, WEKA is weak in statistical analysis. The most important drawbacks of WEKA are the inability to perform multi-relational process and the lack of a merge tool for interconnected tables [16].

2.2. The Fuzzy c-means Algorithm

The fuzzy c-means algorithm was introduced by Ruspini [17] and later extended by Dunn [18] and Bezdek [9, 19, 20] and has been widely used in cluster analysis, pattern recognition and image processing *etc*. The fuzzy c-means clustering algorithm (FCM) introduces the fuzziness for the belongingness of each object and can retain more information of the data set than the hard k-means clustering algorithm (HCM). Although the FCM algorithm has considerable advantages compared to the k-means clustering algorithm, there are also some shortcomings when using the FCM algorithm in practice.

The main limitation of the FCM algorithm is its sensitivity to noises. The FCM algorithm implements the clustering task for a data set by minimizing an objective-function subject to the probabilistic constraint that the summation of all the membership degrees of every data point to all clusters must be one. This constraint results in the problem of this membership assignment, that noises are treated the same as points which are close to the cluster centers. However, in reality, these points should be assigned very low or even zero membership in either cluster. In order to further enhance its robustness to noise and outliers, many researches have been conducted. The possibilistic clustering algorithm (PCA) was put forward by Krishnapuram and Keller [7, 21] to improve this drawback of FCM. PCA relaxes the column sum constraint so that the sum of each column satisfies the looser constraint. However, PCA is heavily dependent on parameters used and may obtains coincident clusters. So several variants of PCA have been brought about to solve the above two restricts of PCA [22, 25]. Nevertheless, although PCA can achieve an overall higher accuracy, FCM is proved to be more consistent and stable experimental results. Another method of solving the FCM's sensitivity to noises is incorporating the FCM algorithm with kernel methods, which has been proved to be robust to outliers or noises of the dataset [26]. M. Gong *et al.*, introduced an improved fuzzy c-means algorithm by applying a kernel distance measure to the objective function [10]. The main idea of kernel methods is to transform complex nonlinear problems in original low-dimensional feature space to the easily solved problems in the high-dimensional transformed space. FLICM [11] was proposed by S. Krinidis and V. Chatzis taking advantage of a fuzzy local similarity measure, which achieves the goal of ensuring noise insensitiveness.

Another shortcoming of the FCM algorithm is the difficulty in selecting appropriate parameters. One of the important parameters is the fuzziness index m which influences the performance of the FCM algorithm when clusters in the data set have

Copyright ⓒ 2013 SERSC

different densities. When $m=1$, the FCM algorithm degenerates into the HCM algorithm. A good choice of m should take the data distribution of the given data set into accout [10]. L. Zhu *et al.*, [27] presented a generalized algorithm called GIFP-FCM, which allows the fuzziness index m not to be fixed at the usual value $m=2$ and improves the robustness and convergence. The method that GIFP-FCM utilizes is conducting a new objective function making use of a novel membership constraint function. The other way to deal with the parameter m is realizing the management of uncertainty on the basis of the fuzziness index. I. Ozkan and I. Turksen [28] introduced a approach that evaluate m according to entropies after removing uncertainties from all other parameters. C. Hwang *et al.*, [29] incorporated the interval type-2 fuzzy set into the FCM algorithm to manage the uncertainty for fuzziness index m.

The last drawback we have to indicate is that the FCM algorithm is easy to get struck in the local minima, while what we want to find is the global extrema. To increase the probability of finding global extrema, various alternative methods for the optimization of clustering algorithm were suggested in many literatures. The most commonly adopted way to ameliorate this problem is integrating genetic algorithms into the FCM algorithm. Ant colony optimization (ACO) [30] has been successfully applied to fuzzy clustering [31]. Runkler introduced an ACO algorithm that explicitly minimizes the HCM and FCM cluster models [32]. Also particle swarm optimization (PSO) [33] has been applied to clustering [34, 35]. ACO and PSO both belong to swarm intelligent algorithms [36]. Swarm intelligent algorithms have the advantages of convenient implementation, parallel capability and ability to avoid local minima.

In addition to the disadvantages mentioned above, the FCM algorithm still has other restricts, for example, influenced by equal partition trend of the data set and sensitivity to initial conditions like the cluster number and the cluster centers. These issues have also been studied in many literatures [37, 38].

Based on the above discussions about limitations of the FCM algorithm, this paper brings about an improved FCM algorithm which realizes the optimization of selecting the initial cluster centers.

3. FCM in WEKA

In this section, we first describe the theory about the FCM algorithm in detail. Secondly, we introduce the basic core classes used in the FCM algorithm at the platform WEKA. Finally, the implement classes of the FCM algorithm are shown.

3.1. Theory on the FCM Algorithm

Contrary to traditional clustering analysis methods, which distribute each object to a unique group, fuzzy clustering algorithms gain the membership values between 0 and 1 that indicate the degree of membership for each objects to each group. Obviously, the sum of the membership values for each object to all the groups is definitely equal to 1. Different membership values show the probability of each object to different groups.

The FCM algorithm is one of the most popular clustering methods based on minimization of a generalized least-squared errors function. Given a data set $X=\{x_1, x_2, ..., x_N\} \subseteq R^{N\times q}$, n is the number of samples, q is the dimension of the sample $x_j (j=1,2,...,N)$. The FCM algorithm is based on minimizing the criterion with respect to the membership value u_{ij} and the distance d_{ij}.

 Copyright ⓒ 2013 SERSC

$$J_m = \sum_{j=1}^{N}\sum_{i=1}^{C}(u_{ij})^m d_{ij}{}^2(x_j, \omega_i) \qquad (1)$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\| x_i - \omega_j \|}{\| x_i - \omega_k \|}\right)^{\frac{2}{m-1}}} \qquad (2)$$

$$\omega_j = \frac{\sum_{i=1}^{N} u_{ij}{}^m x_i}{\sum_{i=1}^{N} u_{ij}{}^m} \qquad (3)$$

Subject to

$$\begin{cases} \sum_{i=1}^{C} u_{ij} = 1,\ j = 1, 2, ..., N \\ 1 \geq u_{ij} \geq 0,\ i = 1, 2, ..., C,\ j = 1, 2, ..., N \\ n > \sum_{j=1}^{N} u_{ij} > 0,\ i = 1, 2, ..., C \end{cases} \qquad (4)$$

Here N is the number of objects and C is the number of clusters, Where u_{ij} is the degree of membership that the object x_j pertains to the cluster center ω_i, $U = \{u_{ij},\ i = 1, 2, ..., C,\ j = 1, 2, ..., N\}$ which is the membership matrix has to satisfy the constraints in (2). $V = \{v_i,\ i = 1, 2, ..., C\}$ is the cluster prototype matrix and v_i is the prototype of the center of cluster i. $m \in [1, \infty)$ is the fuzzy factor. According to many studies, $m \in [2, 2.5]$ is practical [39].

The FCM algorithm can be summarized by the following steps:

Step1: Initialize matrix $U = [u_{ij}]$ with the initial value $U^{(0)}$;

Step2: At k-step: calculate the cluster prototype matrix $V^{(k)} = [v_i]$ with $U^{(k)}$;

Step3: Update $U^{(k)}$, $U^{(k+1)}$;

Step4: if $\| U^{(k+1)} - U^{(k)} \| < \varepsilon$ then stop, or to step2.

To sum up, the basic idea of the FCM algorithm is that use iterative method for solving equation (2) and (3), until a termination condition is met. Here, ε is the threshold of the termination condition.

3.2. Basic Core WEKA Classes used in Clustering

(1) ArffLoader

The ArffLoader class is used to read the standard file type ARFF (Attribute-Relation File Format) used in algorithms.

In WEKA, limited file types are allowed to be handled with. When we do clustering analysis, the standard file type is the ARFF (Attribute-Relation File Format) file. Before we

deal with files storing data sets, we must transfer them into the standard type. The conversion process is complicated before WEKA v6.0 appears. WEKA whose version is higher than 6.0 provides the tool used to transfer file types into ARFF. The dataset representation in ARFF is shown below with some numeric data:

@relation name

@attribute a1 numeric

@attribute a2 numeric

…

@attribute an numeric

@data

As is shown above, the ARFF file consists of two distinct sections, the header section and the data section. The header section contains the name of the relation, while the data section includes the name and the type of attributes.

(2) Instances

The Instances class shows the way to store all the records in data sets. It is responsible for inputting all the records from data source files which are always the ARFF files. In this class, we should pay more attention to the following functions.

Instances(data): the constructor function which return instances from the given data;

numAttributes(): return the number of attributes in the data set;

numInstances(); return the number of instances in the data set;

instance(i): return the ith instance of the instances;

(3) Instance

The instance class store just one record of the data set, it just can be seen as the format of a single record in the algorithm.

(4) DecisionTableHashKey

The function of this class is exploiting decision tables to randomly generate the cluster prototype matrix. The FCM algorithm should initialize the cluster prototype matrix or the membership matrix after setting the parameters specified. In this paper, initializing the cluster prototype matrix is adopted, and then computing the membership matrix according the formula (2). So this class is applied to confirm the original cluster prototype matrix.

(5) Matrix

According to the theory of the FCM algorithm, matrix is the most manifestation of the data in FCM. For instance, cluster centers are stored in the format of the cluster prototype matrix; the possibilities of each object belongs to different clusters are also in the form of the membership matrix.

3.3. The Implement Classes of the FCM Algorithm

In traditional WEKA, the FCM algorithm is not integrated in WEKA. Now, we implement the FCM algorithm making use of the existing classes in WEKA. The class implementing the

 Copyright ⓒ 2013 SERSC

FCM algorithm in WEKA is called FCMWeka. In this class, several interesting functions are used:

(1) setNumClusters() and getNumClusters()

This function is used to set the number of cluster centers. If you don't call this function, the FCM algorithm also set the default value of the number of cluster centers which is 2.

(2) setFuzzifier() and getFuzzifier()

In the FCM algorithm, a significant parameter is the fuzzy factor which is 2 in default. The value of the fuzzy factor plays an important role in the performance of the FCM algorithm. Although repeated studies have shown that 2 is the most appropriate value of the fuzzy factor, the fuzzy factor is closely related to the different experimental data. So we still give users more choices to choose the most favorable value according to their own experimental data. For this intent, the function setFuzzifier() appears.

(3) getClusterCentriods()

The getClusterCentriods() function returns the specific value of the cluster centroids.

(4) getClusterSizes()

The getClusterSizes() function returns the number of objects in each cluster.

(5) buildClusterer()

The buildCluster() function is the most important function to generate clusters. Updating the membership matrix and the cluster matrix and computing the objective function to close to the threshold are all implemented in this function.

(6) ClusterProcessedInstance()

This function can tell which cluster every instance belongs to and return the corresponding cluster number.

4. The Improved FCM Algorithm for Meteorological Data

According to the discussion about the traditional FCM algorithm in Section 2, the initial condition of cluster centers influences the performance of the algorithm. The best choice of the original cluster centers needs to consider the features of the data set. In this paper, meteorological data is chosen as our experimental data. Meteorological data is different from other experiment data. If we just use the traditional FCM algorithm to deal with the meteorological data, there will be a large error when clustering a certain object. To solve the initialization problem, we put forward an improved FCM algorithm in term of selecting the initial cluster centers.

Nowadays, there are several methods to select the original cluster centers. In the following section we will go through some commonly used methods.

(1) Randomly

The traditional FCM algorithm determines initial cluster centers randomly. This method is simple and generally applicable to all data but usually causes local minima.

(2) user-specified

Copyright ⓒ 2013 SERSC

Normally, users decide original cluster centers by some priori knowledge. According to the understanding of the data, users always can obtain logical cluster centers to achieve the purpose of the global optimum.

(3) Randomly classify objects into several clusters, compute the center of each cluster and determine them as cluster centers

More time consumption is spent when randomly classifying objects in this method. When the number of objects in data sets is very small, the cost of time can be ignored. Nevertheless, as the number of objects increases, the speed of the increasing cost of time can be largely rapid.

(4) Select the farthest points as cluster centers

Generally speaking, this method selects initial cluster centers following to the maximum distance principle. It can achieve high efficiency if there are no outliers or noisy points in data sets. But if the data sets contain some outliers, outliers are easier to be chosen as the cluster centers.

(5) Select points with the maximum density

The number of objects whose distance is less than the given radius r from the observed object is defined as the density of the observed object. After computing the density of each object, the object whose density is the largest is chosen as the cluster center. Then compute densities of objects whose distances are larger than the given distance d from the selected center centers, also choose the object whose density is the largest as the second cluster center. And so forth until the number of cluster centers reaches the given number. This method ensures that cluster centers are far away from each other to avoid the objective function into local minima.

In our paper, we adopt a new method to determine cluster centers which is based on the fifth method as mentioned above. In our method, we first randomly select the observed object and compute the density of the observed object. If the density of the observed object is not less than the given density parameter, the observed object can be seen as the cluster center. Secondly we keep selecting the second cluster center satisfying the above constraints in the data set which excludes the objects which are cluster centers or objects whose distances are less than the given distance parameter. Finally we obtain the given number of cluster centers after repeating the above process. The distance parameter and the density parameter are decided by users according to the characteristics of the data sets and the priori knowledge. This selection strategy spends less time than the fifth method because time of computing densities of all objects in the data set is saved, while this method maintains the advantage of avoiding the object function into local minima.

5. Experiment and Discussion

In this section, we describe the detail information of our experiments. The traditional K-means algorithm, the traditional FCM algorithm and above mentioned improved FCM algorithm are all implemented on the platform WEKA with our experimental data. Meanwhile, our experimental results are visualized.

Currently, there are no standard metrics to measure the performance of clustering algorithms in the meteorological application field. But there are many comparison methods to compare different clustering algorithms, like iterative counts, MSE (Mean Square Error), partition coefficient and partition entropy and so on [37, 38]. In our

 Copyright ⓒ 2013 SERSC

experiment, what we use to compare the strengths and weakness of different algorithms depends on the basic metrics in the clustering field: number of iterations and squared errors, and some priori knowledge of meteorology in addition.

5.1. Data Description

Our experimental data come from the drought data in different counties of Anhui province. There are three attributes in the data set, respectively latitude, longitude and the drought area.

Table I. Experimental Datasets Summarize

Dataset	#Objects	#Dimensions	Missing
Drought Areas	463	3	Null

In Table I, the information about the dataset can be concluded that there are 463 objects with 3 attributes in the dataset. Missing represents whether the data set includes the missing values, Null means there is no missing value in the current data set.

5.2. Data Preprocessing

Generally, the experimental data set usually has some objects which have uncertain value or missing value. In our paper, we just ignore these objects to simplify the experiment. To better visualize our outcome of the experiment and effectively describe the characteristics of objects, we firstly normalize the experimental data sets after eliminating the missing objects before the clustering analysis.

Here, we adopt the min-max normalization to preprocess our experiment data.

$$v' = \frac{v - \min_A}{\max_A - \min_A} \qquad (5)$$

Where v is the original value of the attribute A, v' is the value after data normalization, $\min_A$ is the minimum value of the attribute A and $\max_A$ is the maximum value of the attribute A.

After this kind of data normalization, the value of all attributes can be limited in the interval [0, 1]. But, there is a problem about min-max normalization. It will encounter an “out-of-bounds” error if a future inputting value for normalization falls outside of the original data range of the attribute A [40].

5.3. Clustering Analysis

The main result of our experiments is that the traditional FCM algorithm has the advantage of robustness for ambiguity and maintains much more information than the traditional k-means algorithm. Besides, our proposed algorithm is more applicable to the meteorological data than the traditional FCM algorithm.

For the traditional FCM algorithm and the improved FCM algorithm, several parameters must to be specified. The fuzziness index $m = 2$, the number of clusters

$C = 3$. After setting these parameters, we can start our FCM algorithm. For the reason that we have successfully integrate the FCM algorithm and improved FCM algorithm into WEKA, what we need to do is just calling the corresponding class to satisfy our experimental purpose.

In Table II, number of iteration represents the count of the objective function needs to satisfy the constraints; cluster centers means the cluster centers of the data set; squared errors hold the squared errors for all clusters; Clustered Instances represent the distribution of all instances in the data set.

Comparing with the results of the K-means algorithm and the traditional FCM algorithm, what we can discover is that the traditional FCM algorithm needs much fewer iterative counts than the K-Means algorithm, while squared errors of the traditional FCM algorithm is larger than the K-Means algorithm. Now we consider the results of the traditional FCM algorithm and the improved FCM algorithm. These two algorithms share the same number of iterations. However the squared error of the improved FCM algorithm is smaller than that of the traditional FCM algorithm but still larger than the K-means algorithm. This phenomenon appears because the squared error is not appropriate to measure the performance of fuzzy clustering algorithms. In this paper, we determine the cluster which each object belongs to according to the membership values of each object in different clusters. This method may be not appropriate because the membership just describes the probability of belongingness of different clusters. The conclusion can be made that both the traditional FCM algorithm and the improved FCM algorithm converge faster than the K-means algorithm.

Table II. Comparison of Different Algorithms

Name of algorithm	Number of iterations	Cluster centers	squared errors	Clustered Instances	
K-Means	14	0.761179 0.401403 0.109383	4.30497234165914	0 1 2	134(29%) 143(31%) 186(40%)
FCM	2	0.705194 0.300523 0.069379	83.60259931111237	0 1 2	175(37%) 154(33%) 134(28%)
Improved FCM	2	0.717584 0.085929 0.349827	80.76605579692948	0 1 2	158(34%) 164(35%) 141(30%)

Figure 1, Figure 2 and Figure 3 respectively describe the distribution of the data set with the K-means algorithm, the traditional FCM algorithm and the improved FCM algorithm. In these figures, X coordinate represents the latitude, Y coordinate represents the longitude and Z coordinate represents the drought area in the given latitude and longitude.

In terms of specific meteorological characteristics, we carefully compare the difference between Figure 1 and Figure 3. The percentage of the drought area in Si town, Suzhou city of Anhui province is 0.23873 after normalization which is obviously belong to the second cluster while K-means groups it into the third cluster, while this point is correctly classified into the second cluster in the improved FCM algorithm. This point is represented as a pentagram in Figure 4 on the basis of Figure 1. Apart from this point, there are other points misclassified into inappropriate clusters according to the

 Copyright ⓒ 2013 SERSC

priori knowledge. So the conclusion is that the improved FCM algorithm is better than the traditional K-means algorithm from the perspective of the priori meteorological knowledge.

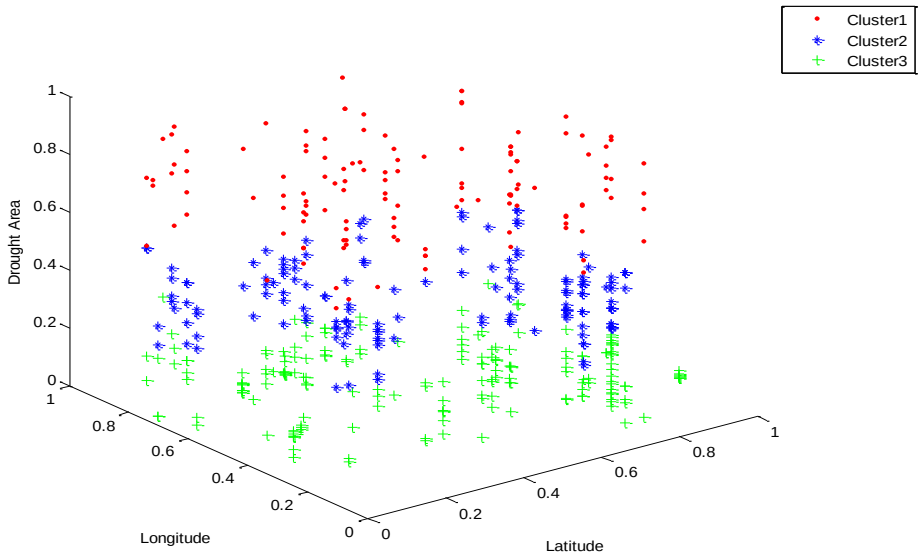


Figure 1. The Distribution of the Data Set with K-Means

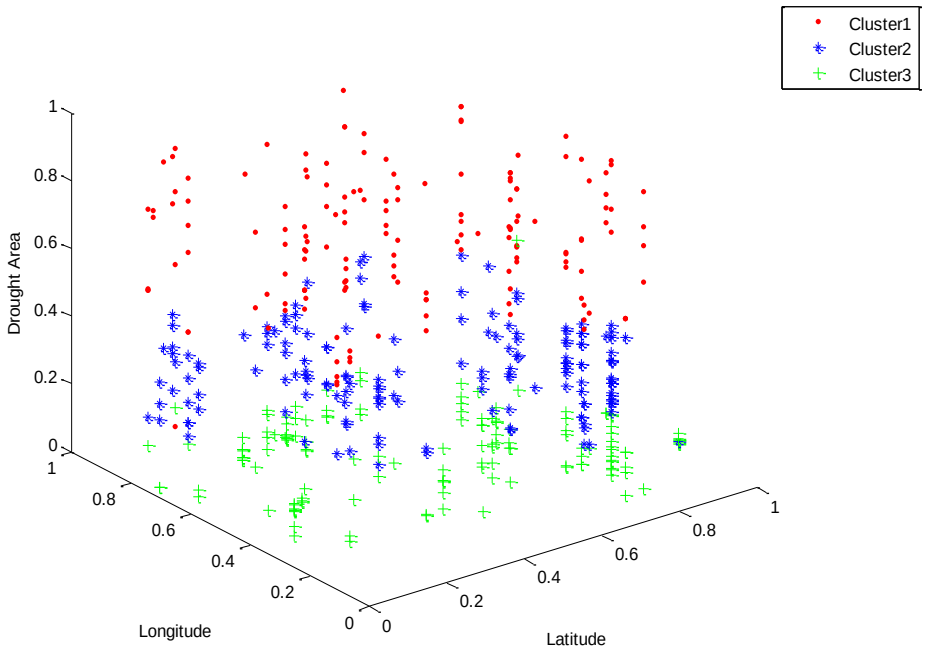


Figure 2. The Distribution of the Data Set with Traditional FCM

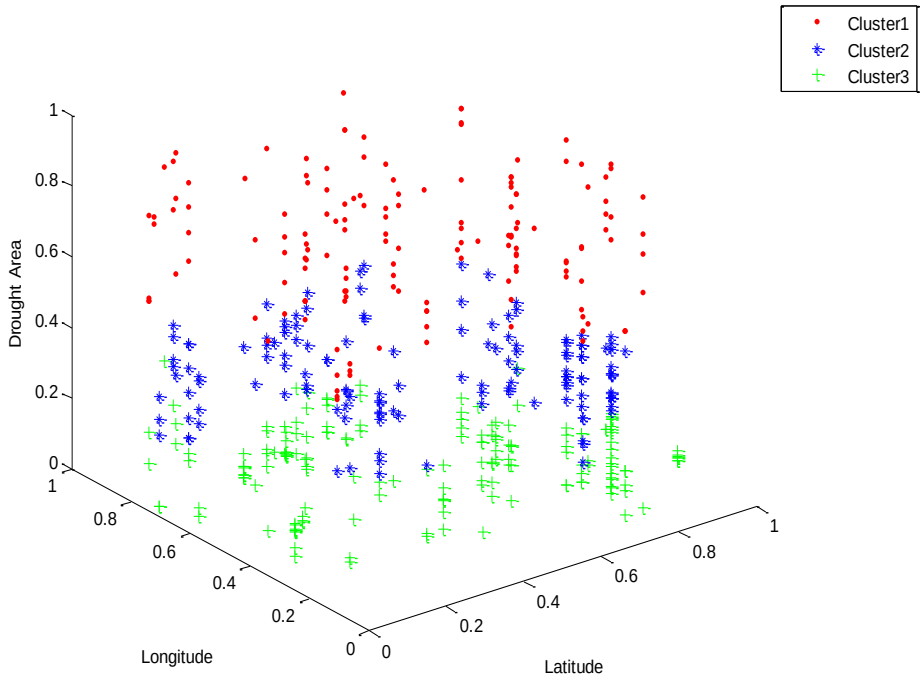


Figure 3. The Distribution of the Data Set with Improved FCM

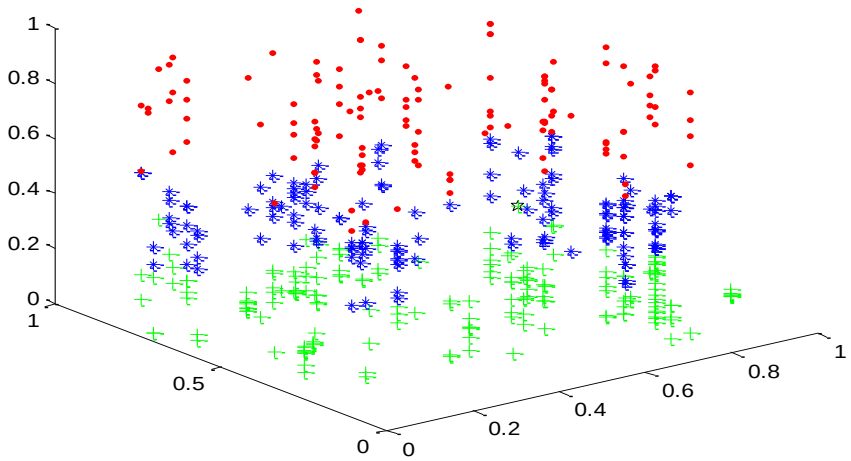


Figure 4. The Wrong Distribution of the Data Set with K-Means

6. Conclusions and Future Research

As the well-known open-source machine learning and data mining software, WEKA includes many java packages such as associations, classifiers, core, clusters and so on.

 Copyright ⓒ 2013 SERSC

However, it doesn't implement the traditional fuzzy clustering algorithm-FCM. In our paper, we successfully integrate the FCM algorithm into WEKA. Compared with the K-Means algorithm existing in WEKA, the FCM algorithm has few iterative counts to faster converge to the global minima than the K-Means algorithm. Furthermore, we improve the traditional FCM algorithm in term of the selection strategy of initial cluster centers to fit the characteristics of meteorological data. The improved FCM algorithm has smaller squared errors than the traditional FCM algorithm while maintaining the rapid speed of convergence. Besides, the performance of the improved FCM algorithm is better than K-means algorithm in terms of the priori meteorological knowledge. Nevertheless, this paper just improves the FCM algorithm in selecting the better cluster centers but does not consider other shortcomings of the FCM algorithm. In the future research, we will improve the performance of the FCM algorithm in the field of meteorology from other aspects.

Acknowledgements

This work was supported in part by National Science Foundation of China (No. 61173143), China Postdoctoral Science Foundation (No.2012M511783), also was supported by Qing Lan Project of Jiangsu Province and was also supported by PAPD.

References

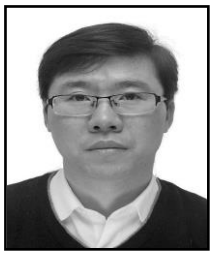
[1] G. Karypis, E. H. Han and V. Kumar, J. Computer, vol. 32, no. 8, **(1999)**.
[2] T. Zhang, R. Ramakrishnan and M. Livny, J. Data Mining Knowledge Discovery, vol. 1, no. 2, **(1997)**.
[3] G. Sheikholeslami, S. Chatterjee and A. Zhang, Editors, A. Gupta, O. Shmueli and J. Widom. Proceedings of the 24th International Conference on Very Large Data Bases, New York, USA, **(1998)** August 24-27.
[4] W. Wang, J. Yang and R. Muntz, Editors. M. Jarke, M. J. Carey, K. R. Dittrich, F. H. Lochovsky, P. Loucopoulos and M. A. Jeusfeld, Proceedings of the 23rd International Conference on Very Large Data Bases, Athens, Greece, **(1997)** August 25-29.
[5] L. A. Zadeh, J. Infection Control, vol. 8, **(1965)**.
[6] J. Valente de Oliveira and W. Pedrycz, Editor, "Advances in Fuzzy Clustering and its Applications", Wiley, Hoboken, **(2007)**.
[7] R. Krishnapuram and J. M. Keller, J. IEEE Transactions on Fuzzy System, vol. 1, no. 2, **(1993)**.
[8] F. L. Chung and T. Lee, J. Neural Networks, vol. 7, no. 3, **(1994)**.
[9] J. C. Bezdek, Editor, "Pattern Recognition with Fuzzy Objective Function Algorithms", Springer Publishers, New York, **(1981)**.
[10] M. Gong, Y. Liang, W. Ma and J. Ma, J. IEEE Transactions on Image Processing, vol. 22, no. 2, **(2013)**.
[11] S. Krinidis and V. Chatzis, J. IEEE Transactions on Image Processing, vol. 19, no. 5, **(2010)**.
[12] C. C. Hung, S. Kulkarni and B. C. Kuo, J. Selected Topics in Signal Processing, vol. 5, no. 3, **(2011)**.
[13] X. Yang, J. Lu and J. Ma, J. IEEE Transactions on Fuzzy Systems, vol. 19, no. 1, **(2011)**.
[14] D. C. Hoang, R. Kumar and S. K. Panda, J. IET Wireless Sensor Systems, vol. 3, no. 3, **(2012)**.
[15] Y. Shen, J. Liu and J. Shen, Editors. IEEE Computer Society Washington, DC, USA. Proceedings of International Conference on Intelligent Computation Technology and Automation, Changsha, China, **(2010)** May 11-12.
[16] I. Charalampopoulos and I. Anagnostopoulos, Editors, IEEE Computer Society Washington, DC, USA, Proceedings of the 15th Panhellenic Conference on Informatics, Kastoria, Greece, **(2011)** September 30-October 2.
[17] E. H. Ruspini, J. Information Sciences, vol. 2, no. 3, **(1970)**.
[18] J. C. Dunn, J. Cybernetics, vol. 3, no. 3, **(1974)**.
[19] J. C. Bezdek, R. Ehrlich and W. Full, J. Computers & Geosciences, vol. 10, **(1984)**, pp. 2-3.
[20] R. L. Cannon, J. V. Dave and J. C. Bezdek, J. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 8, no. 2, **(1986)**.
[21] R. Krishnapuram and J. M. Keller, J. IEEE Transactions on Fuzzy Systems, vol. 4, no. 3, **(1996)**.
[22] M. S. Yang and K. L. Wu, J. Pattern Recognition, vol. 39, no. 1, **(2006)**.

Copyright ⓒ 2013 SERSC

[23] H. Timm, C. Borgelt and R. Kruse, Editors. P. Sincak, J. Strackeljan, M. Kolcun, J. Bojtos, A. Toth, P. Szathmary, M. Hrehus and D. Novotny, Proceedings of Europe Symp. Intelligent Technologies, Tenerife, Spain, **(2001)** November 20-26.
[24] H. Timm and R. Kruse, Editors, IEEE, Proceedings of the 2002 IEEE International Conference on Fuzzy Systems, Honolulu, Hawaii, **(2002)** May 12-17.
[25] H. Timm, C. Borgelt, C. Doring and R. Kruse, J. Fuzzy Sets and Systems, vol. 147, no. 1, **(2004)**.
[26] K. L. Wu and M. S. Yang, J. Pattern Recognition, vol. 35, no. 10, **(2002)**.
[27] L. Zhu, F. L. Chung and S. Wang, J. IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, vol. 39, no. 3, **(2009)**.
[28] I. Ozkan and I. B. Turksen, Editors, IEEE. Proceedings of the 13th International Conference on Fuzzy Systems, Budapest, Hungary, **(2004)** July 25-29.
[29] C. Hwang and F. C. H. Rhee, J. IEEE Transactions on Fuzzy Systems, vol. 15, no. 1, **(2007)**.
[30] M. Dorigo, M. Birattari and T. Stiitzle, J. IEEE Computational Intelligence Magazine, vol. 1, no. 4, **(2006)**.
[31] P. M. Kanade and O. H. Lawrence, J. IEEE Transactions on Systems, Man and Cybernetics-Part A: Systems and Humans, vol. 37, no. 5, **(2007)**.
[32] T. A. Runkler, J. Intelligent Systems, vol. 20, **(2005)**.
[33] J. Kennedy and R. Eberhart, Editors, IEEE, Proceedings of IEEE International Conference on Neural Network, Perth, Australia, **(1995)** November 27-December 1.
[34] T. A. Runkler and C. Katz, IEEE, Proceedings of the 15th International Conference on Fuzzy Systems, Sheraton Vancouver Wall Centre, Vancouver, BC, Canada, **(2006)** July 16-21.
[35] F. Q. Yang, T. L. Sun and C. H. Zhang, J. Expert Systems with Applications, vol. 36, no. 6, **(2009)**.
[36] R. Xu, J. Xu and D. C. Wunsch, J. IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics, vol. 42, no. 4, **(2012)**.
[37] H. Z. Zhang, H. Chen and L. X. Bao, Editors, L. Wang, L. Jiao, G. Shi and J. Liu, Proceedings of the 3rd International Conference on Fuzzy Systems and Knowledge Discovery, Xi'an, China, **(2006)** September 24-28.
[38] X. Liu and C. Yang, Editors, "Circuits and System Society", Proceedings of the 6th International Conference on Natural Computation, Yantai, China, **(2010)** August 10-12.
[39] N. R. Pal and I. C. Bezdek, J. IEEE Transactions on Fuzzy Systems, vol. 3, no. 4, **(1995)**.
[40] S. N. Deepa, N. Malathi and C. V. Subbulakshmi, Editors, Madras and Women in Engineering, Proceedings of IEEE International Conference on Advanced Communication Control and Computing Technoligies, Ramanathapuram, India, **(2012)** August 23-25.

Authors

Yinhua Lu, received her Bachelor degree in Computer Science and Engineering from Nanjing University of Information Science & Technology, China in 2009. Currently, she is a candidate for the degree of Master of Computer Science and Engineering in Nanjing University of Information Science & Technology. Her research interests include GPU programming, image processing etc.



Tinghuai Ma, he received his Bachelor (HUST, China, 1997), Master (HUST, China, 2000), PhD (Chinese Academy of Science, 2003) and was Post-doctoral associate (AJOU University, 2004) and a visiting Professor in Kyung Hee University, Korea (KHU, 2009). Now, he is a professor in Computer Sciences at Nanjing University of Information Science & Technology, China. His research interests are data mining, grid computing, ubiquitous computing, privacy preserving, *etc*.

 Copyright ⓒ 2013 SERSC

Wei Tian, he is an assistant professor of Computer Sciences at Nanjing University of Information Science & Technology, China. He received his master degree from Nanjing University of Information Science & Technology, China, 2006. Now, he is a doctoral candidate of Applied Meteorology, Nanjing University of Information and Science Technology. His main research interests are in the areas of Cloud Copmputing and Meteorological Data Processing.

Shuiming Zhong, he received the M.S. and Ph.D. degrees from the Hohai University, China, in 2007 and 2011 respectively. He currently is a lecturer in the School of Computer and Software, Nanjing University of Information Science and Technology, China. His research interests involve artificial neural networks, machine learning, and pattern recognition.

Copyright ⓒ 2013 SERSC

Copyright ⓒ 2013 SERSC

Fuzzy c-Means Clustering of Incomplete Data

Richard J. Hathaway, *Member, IEEE*, and James C. Bezdek, *Fellow, IEEE*

***Abstract*—The problem of clustering a real s-dimensional data set $X = \{x_1, \ldots, x_n\} \subset R^s$ is considered. Usually, each observation (or datum) consists of numerical values for all s features (such as height, length, etc.), but sometimes data sets can contain vectors that are missing one or more of the feature values. For example, a particular datum x_k might be incomplete, having the form x_k = (254.3, ?, 333.2, 47.44, ?)^T, where the second and fifth feature values are missing. The fuzzy c-means (FCM) algorithm is a useful tool for clustering real s-dimensional data, but it is not directly applicable to the case of incomplete data. Four strategies for doing FCM clustering of incomplete data sets are given, three of which involve modified versions of the FCM algorithm. Numerical convergence properties of the new algorithms are discussed, and all approaches are tested using real and artificially generated incomplete data sets.**

***Index Terms*—Clustering, fuzzy c-means (FCM), incomplete data, missing data.**

I. Introduction

We are interested in clustering a set of objects $O = \{o_1, \ldots, o_n\}$ represented by a numerical object data set $X = \{x_1, \ldots, x_n\} \subset R^s$ into c clusters, $1 < c < n$. The numerical data describes the objects by specifying values for s particular features. For example, if the jth component of each datum corresponds to the feature of weight, then the jth component of x_k, denoted by x_{kj}, gives the weight of object o_k. In some situations, feature vectors in X can have missing components [1], [2]. Any datum with some (but not all) missing feature values is referred to as an *incomplete datum*. An example of an incomplete datum is $x_5 = (1.23, 3.14, ?, 4.66, ?)^T$, where x_{53} and x_{55} are missing. A data set X with at least one incomplete datum is referred to as an *incomplete data set*; otherwise, it is called complete.

A. Previous Work

The problem of doing pattern recognition with incomplete data is certainly not new. Theoretically-oriented approaches based on probabilistic assumptions date back to at least 1962 [3]. An important empirically-oriented study was done in 1979 [1], and those results are summarized in [2]. It may be the case that predicted values of the missing data are desired [4], [5] (or even predicted values of new data, as in [6]), but often the goal is to simply use the available data to perform some analysis [7], [8]. The goal of the new clustering approaches introduced later is to partition the data set into fuzzy clusters and provide estimates of their cluster centers. Two of our approaches produce predicted values of the missing data, as a byproduct of the calculations.

There is an important, general statistical method for handling incomplete data based on the expectation–maximization (EM) algorithm [9], [10]. In essence, this approach iteratively produces maximum likelihood estimates using a two-part iteration. The first part (E-step) uses current distributional parameter estimates to calculate **e**xpected values for the missing features. The second part (M-step) uses the current expected values to complete the available data and then calculates improved (complete data) **m**aximum-likelihood parameter estimates. The two-part EM iteration continues using the improved parameter estimates, until convergence is achieved. An important connection between EM and pattern recognition arises through its use to calculate parameter estimates for mixture distributions, which are used in one approach to probabilistic clustering [11]. Two other useful sources of information for statistical approaches to analysis of incomplete data are [12] and[13].

B. Geometry of Missing Data

We begin with a discussion of visualization of incomplete data sets. In the special case that $X \subset R^2$, it is possible to accurately represent both complete and incomplete data sets via scatter diagrams. To do this, we represent an incomplete datum $x_k = (x_{k1}, ?)^T$ as a vertical line with horizontal component x_{k1} and $x_k = (?, x_{k2})^T$ as a horizontal line with vertical component x_{k2}. (Of course the null datum (?,?) can not be usefully represented in such a plot.) In higher dimensions, an incomplete datum could correspond to a line, plane, etc., depending on the number of missing feature values. The left scatter diagram [Fig. 1(a)] corresponds to a complete data set. The right plot [Fig. 1(b)] depicts an incomplete version of the data, where 25% of the (20) feature values are missing. In this example, three values of the first feature and two of the second feature are missing. Notice the increased difficulty in visually identifying the presence (and location) of the two clusters when a substantial percentage of feature values is missing.

There is an interesting extreme case of the visual difficulty demonstrated in Fig. 1 that involves what we will call "virtual clusters." To make this phenomenon clear, we consider the most extreme case of incompleteness. We will refer to a data set X as *maximally incomplete* if each datum in X has exactly one known feature value. In the case of a maximally incomplete data set $X \subset R^2$, exactly 50% of the feature values are missing. For this case a plot such as the right side of

Manuscript received March 12, 2000; revised April 24, 2001. The work of R. J. Hathaway was supported in part by the Faculty Research Committee of Georgia Southern University. The work of J. C. Bezdek was supported by Office of Naval Research (ONR) Grant 00014-96-1-0642. This paper was recommended by Associate Editor A. Bouzerdoum.

R. J. Hathaway is with the Mathematics and Computer Science Department, Georgia Southern University, Statesboro, GA 30460 USA (e-mail: hathaway@ieee.org).

J. C. Bezdek is with the Department of Computer Science, University of West Florida, Pensacola, FL 32504 USA.

Publisher Item Identifier S 1083-4419(01)08549-1.

1083–4419/01$10.00 © 2001 IEEE

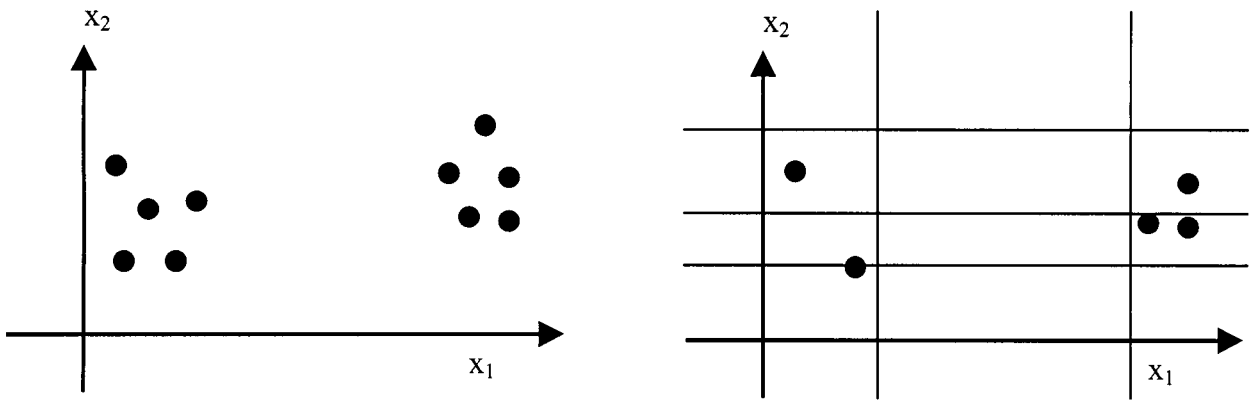


Fig. 1. Complete data on left. Incomplete data with 25% of feature values missing on right.

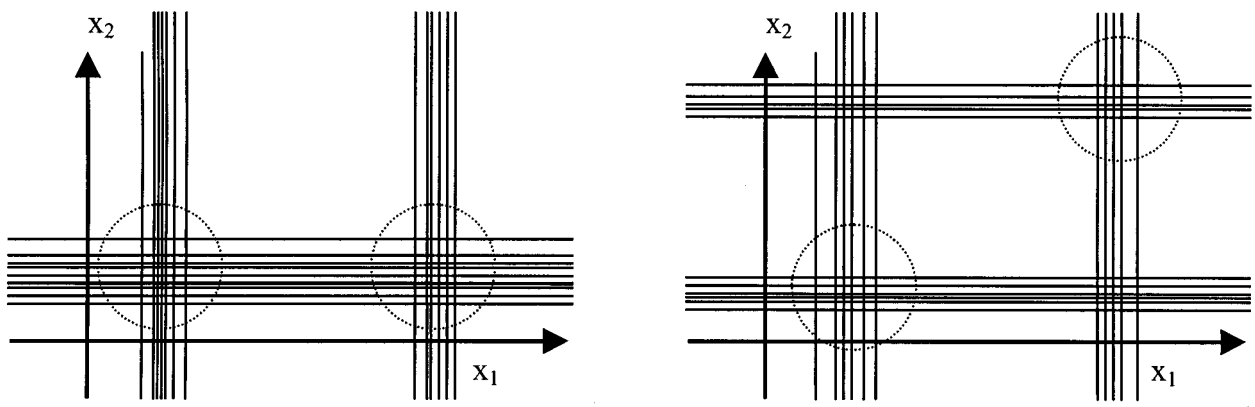


Fig. 2. Two maximally incomplete two-dimensional (2-D) data sets.

Fig. 1 will contain no points; instead, it will consist entirely of a lattice of n horizontal/vertical lines such as in Fig. 2. More generally, a maximally incomplete data set $X \subset R^s$ is missing exactly $100(s-1)/s\%$ of its feature values. It is generally impossible to find good estimates for the cluster centers if X is maximally incomplete; we demonstrate this using the examples in Fig. 2.

Each of the two plots in Fig. 2 depicts a maximally incomplete two-dimensional (2-D) data set, which, if complete, would have two visual clusters. These maximally incomplete data sets were obtained by starting with complete data, whose general positions are indicated by the circles; and then for each datum, designating one of its two feature values as "missing." If we know that $c = 2$, then is it possible to recover reasonable estimates of the original (complete) data *cluster centers* from the maximally incomplete data set? This is one goal of cluster analysis in an incomplete data set. The answer is yes for the data underlying the left view in Fig. 2, but is no for the right view, where two additional "virtual clusters" are indicated. (In fact, each of the two plots in Fig. 2 is consistent with certain other complete data sets having more than two clusters.) Clustering is difficult, and this example illustrates an additional difficulty inherent in clustering incomplete data.

C. Organization of Paper

The purpose of this paper is to introduce four strategies for clustering incomplete data sets. Three of these consist of new adaptations of the *fuzzy c-means* (FCM) algorithm [14], and all four provide estimates of the locations of cluster centers and fuzzy partitions of the data. The next section fixes our notation, gives a quick review of FCM, gives the new algorithms, and discusses their theoretical convergence properties. In Section III, the four approaches are compared using several artificially generated data sets and the IRIS data [15]. We examine the accuracy of the computed cluster centers and the correctness of the clusterings produced by each of the methods. The final section summarizes the results of this paper, adds some relevant discussion, and identifies several problems that are deserving of future consideration.

II. Four FCM-Based Strategies for Clustering Incomplete Data

The notation that will be used throughout follows. Let

$$x_k = k^{th}\ s\text{-dimensional data vector, for } 1 \le k \le n; \tag{1a}$$

$$x_{kj} = j^{th} \text{ feature value of the } k^{th} \text{ data vector, for } 1 \le j \le s,\ 1 \le k \le n; \tag{1b}$$

$$X = \{x_1, \ldots, x_n\}; \tag{1c}$$

$$X_W = \{x_k \in X | x_k \text{ is a complete datum}\} \text{ (the } whole\text{-}data \text{ subset of } X); \tag{1d}$$

$$X_P = \{x_{kj} \text{ for } 1 \le j \le s,\ 1 \le k \le n \mid \text{the value for } x_{kj} \text{ is present in } X\}; \tag{1e}$$

$$X_M = \{x_{kj} = ? \text{ for } 1 \le j \le s,\ 1 \le k \le n \mid \text{the value for } x_{kj} \text{ is missing from } X\}. \tag{1f}$$

For example, let $s = 3$, $n = 4$, and

$$X = \left\{ \begin{bmatrix} 2 \\ ? \\ ? \end{bmatrix} \quad \begin{bmatrix} 3 \\ 2 \\ 4 \end{bmatrix} \quad \begin{bmatrix} 2 \\ 1 \\ ? \end{bmatrix} \quad \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} \right\}.$$

Then

$$X_W = \left\{ \begin{bmatrix} 3 \\ 2 \\ 4 \end{bmatrix} \quad \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} \right\}$$

$X_M = \{x_{12}, x_{13}, x_{33}\}$ and $X_P = \{x_{11} = 2, x_{21} = 3, x_{22} = 2, x_{23} = 4, x_{31} = 2, x_{32} = 1, x_{41} = 1, x_{42} = 1, x_{43} = 2\}$. Notice that the available feature values in X_P and X are the same, and that only a subset of these feature values is available in X_W. In this sense, X_P and X contain more "information" than X_W.

The FCM clustering algorithm is widely used to cluster (whole) object data sets X, and is used in a pure or modified form for all four approaches given in this section. FCM attempts to simultaneously produce a description of the (fuzzy) cluster substructure of X along with exemplars (or prototypes) for each of the clusters via an iterative scheme for solving

$$\min_{(U,v)} \left\{ J_m(U, v) = \sum_{i=1}^{c} \sum_{k=1}^{n} U_{ik}^m \|x_k - v_i\|_A^2 \right\} \tag{2}$$

where $m > 1$ is the fuzzification parameter, $\|x\|_A^2 = x^T A x$ is the vector A-norm

$$U \in M_{fcn} = \left\{ U \in R^{c \times n}\colon U_{ik} \in [0, 1] \quad \forall i, k; \quad \sum_{i=1}^{c} U_{ik} = 1 \quad \forall k; \sum_{k=1}^{n} U_{ik} > 0 \quad \forall i \right\} \tag{3}$$

and

$$v = \{v_1, \ldots, v_c\} \subset R^s. \tag{4}$$

When $A = I_{s \times s}$, $\|x\|_A = \|x\|_2$, the Euclidean norm. The elements of the membership matrix U represent the (relative) degrees of membership of each datum in each cluster; i.e., $U_{ik} =$ the degree to which x_k belongs to the ith cluster. Minimizing U values of (2) are less fuzzy for values of m near 1 and fuzzier for large values of m. The choice of $m = 2$ is widely accepted as a good choice of fuzzification parameter. The set M_{fcn} in (3) is referred to as the set of nondegenerate (meaning each cluster is nonempty) fuzzy $c \times n$ partition matrices. The vector v_i is the prototype or center of the ith cluster.

The FCM approach has been widely used, adapted, and generalized [16]–[19]. The FCM algorithm for solving (2) alternates optimizations of J_m over the U and v variables. This method of alternating optimization has been well studied [20], [21] and typically provides iterate sequences that converge q-linearly to local (and sometimes global) minimizers. The details of one version of FCM are now given. Other versions arise if we initialize with $U^{(0)}$ rather than $v^{(0)}$; or if we check $\|U^{(r+1)} - U^{(r)}\|$ rather than $\|v^{(r+1)} - v^{(r)}\|$ for convergence.

Fuzzy c-Means Algorithm:

FCM-1 Fix m, c, and ε satisfying: $m > 1$, $1 < c < n$, and $\varepsilon > 0$. Pick $v^{(0)} \subset R^s$, an $\| \ \|_A$ norm for J_m, and a termination norm $\| \ \|$ for FCM. Then at step r, $r = 0, 1, 2, \ldots$

FCM-2 Calculate

$$U^{(r+1)} = \underset{U \in M_{fcn}}{\arg\min} \left\{ J_m\left(U, v^{(r)}\right) \right\}$$

using the following for $1 \le i \le c$ and $1 \le k \le n$:

$$U_{ik}^{(r+1)} = \left(D_{ik}^{1/(1-m)}\right) \Big/ \left(\sum_{j=1}^{c} D_{jk}^{1/(1-m)} \right) \quad \text{where } D_{ik} = \left\| x_k - v_i^{(r)} \right\|_A^2. \tag{5}$$

FCM-3 Calculate

$$v^{(r+1)} = \underset{v}{\arg\min} \left\{ J_m(U^{(r+1)}, v) \right\}$$

using the following for $1 \le i \le c$ and $1 \le j \le s$:

$$v_{ij}^{(r+1)} = \left(\sum_{k=1}^{n} \left(U_{ik}^{(r+1)}\right)^m x_{kj} \right) \Big/ \left(\sum_{k=1}^{n} \left(U_{ik}^{(r+1)}\right)^m \right). \tag{6}$$

FCM-4 Compare $v^{(r+1)}$ to $v^{(r)}$ using $\|v^{(r+1)} - v^{(r)}\| < \varepsilon$. If true, then stop. Otherwise, set $r = r + 1$ and return to FCM-2.

Note that the FCM algorithm references each component of each datum in $X = \{x_1, \ldots, x_n\} \subset R^s$. Therefore, FCM cannot be directly applied to data sets containing missing feature values.

A. Whole Data Strategy (WDS)

If the proportion of incomplete data is small, then it may be useful to simply delete all incomplete data and apply FCM to the remaining complete data. We will refer to this as the *whole data strategy (WDS).* Although we have not studied this issue carefully, it seems reasonable to set a fairly stringent limit on the number of vectors that are cast out of X because they have one or more missing values. Our supposition is that the WDS should be used only if $n_p/n_s \ge 0.75$, where $n_p = |X_P|$ and $n_s = |X| \cdot s$. In the notation introduced at the beginning of the section, this simply amounts to applying FCM to the data subset X_W.

The WDS approach may provide good estimates of the prototypes $v = \{v_1, \ldots, v_c\}$, but it will not directly provide cluster membership information for feature vectors in $X - X_W$, since data in this set are not represented by columns in the corresponding reduced version of U. Nonetheless, cluster membership can be estimated for the incomplete data using a nearest-prototype classification scheme based on the partial distances (described in the next subsection) from each incomplete datum

to each of the computed cluster centers. In the following we will represent this approach as WDSFCM. This approach enjoys all the standard numerical convergence properties of FCM since it is exactly FCM.

B. Partial Distance Strategy (PDS)

The top strategy recommended by Dixon [1] for the case that $|X_M|$ is sufficiently large that the WDS cannot be justified consists of calculating partial (squared Euclidean) distances using all available (i.e., nonmissing) feature values, and then scaling this quantity by the reciprocal of the proportion of components used. We will call this the *partial distance strategy (PDS)* and illustrate the PDS form of the D_{ik} calculation in (5) using the example

$$\begin{aligned} D_{ik} &= \left\| x_k - v_i^{(r)} \right\|_2^2 \\ &= \left\| (1 \quad ? \quad ? \quad 4 \quad ?)^T - (5 \quad 6 \quad 7 \quad 8 \quad 9)^T \right\|_2^2 \\ &= \frac{5}{(5-3)} \left((1-5)^2 + (4-8)^2 \right). \end{aligned} \tag{7}$$

The general formula for the partial distance calculation of D_{ik} is given by

$$D_{ik} = \frac{s}{I_k} \sum_{j=1}^{s} (x_{kj} - v_{ij})^2 I_{kj} \tag{8a}$$

where

$$I_{kj} = \begin{cases} 0, & \text{if } x_{kj} \in X_M \\ 1, & \text{if } x_{kj} \in X_P \end{cases} \qquad \text{for } 1 \le j \le s \text{ and } 1 \le k \le n$$

$$\text{and } I_k = \sum_{j=1}^{s} I_{kj}. \tag{8b}$$

The PDS version of the FCM algorithm, referred to here as PDSFCM, is obtained by making two modifications of the FCM algorithm given earlier for the special case of $\|x\|_A = \|x\|_2$. They are 1) calculate D_{ik} in (5) for incomplete data according to (8) and 2) replace the calculation of $v^{(r+1)}$ in (6) with

$$v_{ij}^{(r+1)} = \left(\sum_{k=1}^{n} \left(U_{ik}^{(r+1)} \right)^m I_{kj} x_{kj} \right) \Bigg/ \left(\sum_{k=1}^{n} \left(U_{ik}^{(r+1)} \right)^m I_{kj} \right). \tag{9}$$

It is interesting to notice that the scaling factor s/I_k in D_{ik} is completely irrelevant, since it has no direct or indirect effect on $v^{(r+1)}$ in (9). The scaling factor also has no effect on U because it appears in both the top and bottom, so will cancel in

$$U_{ik}^{(r+1)} = \left(D_{ik}^{1/(1-m)} \right) \Bigg/ \left(\sum_{j=1}^{c} D_{jk}^{1/(1-m)} \right)$$

from (5). This approach enjoys all the standard convergence properties of FCM because it is an instance of alternating optimization [20]. Specifically, using the techniques originally applied for FCM in [14], it is straightforward to establish that $U^{(r+1)}$ and $v^{(r+1)}$ of PDSFCM satisfy

$$U^{(r+1)} = \underset{U \in M_{fcn}}{\arg\min} \left\{ K_m \left(U, v^{(r)} \right) \right\} \tag{10a}$$

and

$$v^{(r+1)} = \underset{v}{\arg\min} \left\{ K_m \left(U^{(r+1)}, v \right) \right\} \tag{10b}$$

for

$$K_m(U, v) = \sum_{i=1}^{c} \sum_{k=1}^{n} U_{ik}^m \sum_{j=1}^{s} (x_{kj} - v_{ij})^2 I_{kj}. \tag{10c}$$

The global convergence theory of Zangwill [22] is applicable, as is the local alternating optimization convergence theory [20]. In a nutshell, the global result states that any limit point of a PDSFCM iteration sequence will either be a local minimizer, or at worst a saddle point of K_m. The local result states that if we initialize the iteration sequence sufficiently near a local minimizer of K_m satisfying the second-order sufficiency conditions for a minimizer, then the iteration sequence will converge q-linearly to that minimizer. (This is the standard FCM theory.)

We mention that another seemingly reasonable modification of FCM based on the distance in (8) is to replace the FCM alternating optimization of $\sum_{i=1}^{c} \sum_{k=1}^{n} U_{ik}^m \|x_k - v_i\|_A^2$ with a similar optimization of $\sum_{i=1}^{c} \sum_{k=1}^{n} U_{ik}^m (s/I_k) \sum_{j=1}^{s} (x_{kj} - v_{ij})^2 I_{kj}$. [Note that this differs with (10c) in that the scaling factor s/I_k is included.] However the scaling factor has the effect of giving extra weight to components of missing data, compared with components from $x_k \in X_P$, in the calculation of cluster means, and we see no justification for such an inequity.

C. Optimal Completion Strategy (OCS)

The third approach to FCM clustering of incomplete data is based on what we will call the *optimal completion strategy (OCS)*. In this approach we view the missing components (X_M) as additional variables over which we optimize in order to get the smallest possible value of the FCM functional J_m. That is, our strategy is to complete the missing part of the data set in the way that leads to the smallest possible value of J_m given the available data X_P. This modification of FCM, referred to here as OCSFCM, is obtained by adding an additional optimization step over the X_M "variables" during each iteration. We state the new algorithm next.

Optimal Completion Strategy Fuzzy c-Means Algorithm:

OCSFCM-1 Fix m, c and ε satisfying: $m > 1$, $1 < c < n$ and $\varepsilon > 0$. Pick $v^{(0)} \in R^{s \times c}$, an $\| \ \|_A$ norm for J_m, and a termination norm $\| \ \|$ for FCM. Pick initial values for cluster means $v^{(0)} \subset R^s$ and missing data $X_M^{(0)}$. Then at step r, $r = 0, 1, 2, \ldots$

OCSFCM-2 Calculate

$$U^{(r+1)} = \underset{U \in M_{fcn}}{\arg\min} \left\{ J_m \left(U, v^{(r)}, X_M^{(r)} \right) \right\}$$

using

$$U_{ik}^{(r+1)} = \left(D_{ik}^{1/(1-m)}\right) \Bigg/ \left(\sum_{j=1}^{c} D_{jk}^{1/(1-m)}\right)$$
$$\text{where} \quad D_{ik} = \left\|x_k - v_i^{(r)}\right\|_A^2 \tag{11}$$

for $1 \le i \le c$ and $1 \le k \le n$.

OCSFCM-3 Calculate

$$v^{(r+1)} = \arg\min_{v} \left\{J_m\left(U^{(r+1)}, v, X_M^{(r)}\right)\right\}$$

using

$$v_{ij}^{(r+1)} = \left(\sum_{k=1}^{n} \left(U_{ik}^{(r+1)}\right)^m x_{kj}\right) \Bigg/ \left(\sum_{k=1}^{n} \left(U_{ik}^{(r+1)}\right)^m\right)$$
$$\text{for } 1 \le i \le c \text{ and } 1 \le j \le s. \tag{12}$$

OCSFCM-4 Compare $v^{(r+1)}$ to $v^{(r)}$ using $\|v^{(r+1)} - v^{(r)}\| < \varepsilon$. If true, then stop. Otherwise, go to OCSFCM-5.

OCSFCM-5 Calculate

$$X_M^{(r+1)} = \arg\min_{X_M} \left\{J_m\left(U^{(r+1)}, v^{(r+1)}, X_M\right)\right\}$$

for all $x_{kj} \in X_M$ using

$$x_{kj}^{(r+1)} = \left(\sum_{i=1}^{c} \left(U_{ik}^{(r+1)}\right)^m v_{ij}^{(r+1)}\right) \Bigg/ \left(\sum_{i=1}^{c} \left(U_{ik}^{(r+1)}\right)^m\right). \tag{13}$$

Now set $r = r + 1$ and return to step OCSFCM-2.

Equation (13) is the first-order necessary condition for unconstrained minimization of the function $J(x_k) = \sum_{i=1}^{c} U_{ik}^m \|x_k - v_i\|_A^2$. Zeroing the gradient of J with respect to the missing features values of x_k leads to (13).

Examination of steps OCSFCM-2, 3, and 5 demonstrates that OCSFCM is a trilevel alternating optimization of $J_m(U, v, X_M)$. As such, it has the numerical convergence properties of *trilevel* alternating optimization which are given in [23] and which are similar to those for the original FCM algorithm. Any limit point of an OCSFCM iteration sequence will be a minimizer of $J_m(U, v, X_M)$, or at worst will be a special type of saddle point that is optimal in each of the three variables while holding the other two fixed. This global result holds for virtually any initialization of the missing part of the data $X_M^{(0)}$. In our implementation, we obtained $X_M^{(0)}$ from $v^{(0)}$ using step NPSFCM-5 given later in (14). The iteration sequence will always converge q-linearly to any minimizer of $J_m(U, v, X_M)$ satisfying the second-order sufficiency conditions if initialized sufficiently near it.

D. Nearest Prototype Strategy (NPS)

The final approach uses the *nearest prototype strategy (NPS)* and can be described as a simple modification of OCSFCM. We call this approach NPSFCM and define it as OCSFCM with Step OCSFCM-5 replaced by

NPSFCM-5 Calculate $X_M^{(r+1)}$ for all $x_{kj} \in X_M$ using partial distances D_{ik} in (8) by

$$x_{kj}^{(r+1)} = v_{ij}^{(r+1)} \quad \text{where } D_{ik} = \min\{D_{1k}, D_{2k}, \ldots, D_{ck}\}. \tag{14}$$

Now set $r = r + 1$ and return to step OCSFCM-2.

In the rare case that an incomplete datum x_k is equidistant to two or more prototypes, a tiebreaking rule must be used in defining $x_{kj}^{(r+1)}$. While NPSFCM terminated in all numerical tests, we have not established theoretically that this procedure must converge.

III. Numerical Examples

In this section we compare the WDS, PDS, OCS, and NPS versions of FCM using artificially generated incomplete data sets. The scheme for artificially generating an incomplete data set X is now described. First a complete data set X is chosen. Then X is modified to obtain an incomplete data set by randomly selecting a specified percentage of its components $\{x_{kj}\}$ to be designated as missing. The random selection of missing feature values is constrained so that

C1) each original feature vector x_k retains at least one component;

C2) each feature has at least one value present in the incomplete data set X.

The initialization $v^{(0)}$ used in all cases is the set $v*$ of (globally) optimal FCM prototypes obtained using the original (complete) data. The additional initialization $X_M^{(0)}$ needed for the OCS and NPS versions is obtained using $v*$ in NPSFCM-5 (14). The stopping criterion $\|v^{(r+1)} - v^{(r)}\| < \varepsilon$ that we used is

$$\max\left(\left|v_{ji}^{(r+1)} - v_{ji}^{(r)}\right|, \text{ for } 1 \le j \le s \text{ and } 1 \le i \le c\right) \le 0.000\,01. \tag{15}$$

A. IRIS Data Example

The IRIS data [15] consists of 150 four-dimensional feature vectors (so 600 feature values), with 50 vectors for each of three physically labeled classes. The maximum possible percentage of missing feature values (for $s = 4$) subject to constraints C1) and C2) is 75%. In this case the whole-data subset X_W is empty and X is maximally incomplete. (X_W can also be empty with proportions of missing values less than 75%.) We noticed significant variation in the results from trial to trial, so Table I represents the averages obtained over 50 trials. In particular, 50 trials are generated for each row in the table; and the same sample is used in each trial for each of the four approaches so that the results can be correctly compared.

The second through fifth columns of Table I give the average number of iterations required for (15) to terminate each of the four schemes. Note that for 0% missing data, all approaches reduce to regular FCM, and they therefore terminate after one iteration since the initialization is a set $v*$ of optimal FCM prototypes. The middle columns of Table I give the average number of misclassifications (resubstitution or training error) obtained when the terminal $(U^{(t)}, v^{(t)})$ is used to reclassify each of the

TABLE I
AVERAGED RESULTS OF 50 TRIALS USING INCOMPLETE IRIS DATA SETS

% Missing	Mean Number of Iterations to Termination				Mean Number of Misclassifications (Training Error)				Mean Prototype Error			
	WDS	PDS	OCS	NPS	WDS	PDS	OCS	NPS	WDS	PDS	OCS	NPS
0	1	1	1	1	16	16	16	16	0.00	0.00	0.00	0.00
10	18.18	15.64	22.20	16.00	16.04	16.24	16.28	16.28	0.17	0.07	0.07	0.07
20	20.74	17.24	22.64	18.12	16.42	16.62	16.48	16.54	0.33	0.12	0.11	0.12
30	22.68	17.84	29.90	20.42	19.00	17.04	16.72	16.98	0.49	0.15	0.14	0.15
40	23.62	18.48	29.02	22.40	22.46	18.76	18.82	18.84	0.78	0.18	0.17	0.18
50	14.80	20.14	33.06	29.08	42.64	21.44	21.14	21.36	2.11	0.25	0.21	0.24
75	N/A	47.20	68.92	70.78	N/A	33.44	29.12	30.54	N/A	0.74	0.39	0.52

150 feature vectors. For WDSFCM, each vector in X is classified according to its nearest prototype from $v^{(t)}$ where the distance calculation used all nonmissing feature values. For all other approaches, the classification was according to the maximum membership value in each column of $U^{(t)}$. (The different approach for WDSFCM was necessary since its terminal $U^{(t)}$ does not have a column for each datum in X). The 0% row shows that regular FCM on the full IRIS data produces prototypes that incorrectly reclassify 16 of the 150 data. The last four columns of the table give the average Frobenius norm distance between the particular terminal prototypes $v^{(t)}$ and the full IRIS data prototypes $v*$ which is computed for each terminal prototype set $v^{(t)}$ as

$$\left\|v^{(t)} - v*\right\|_F = \left(\sum_{j=1}^{s}\sum_{i=1}^{c}\left|v_{ji}^{(t)} - v_{ji}^{*}\right|^2\right)^{1/2}.$$

The smallest entries in each row (i.e., the "best" values in this study) are highlighted by shading.

The accuracy of prototype estimation by OCS is always at least as good as the other approaches, based on the errors in the last four columns. The simplest strategy, WDS, provides good testing errors for percentages up to 20%. While the general accuracy and misclassification errors of PDS, OCS, and NPS are quite similar in all but the worst (75%) case, the PDS approach almost always requires fewer iterations than the other three methods. Testing errors for the WDS are reasonable for $(n_p/n_s =)$ 0.9 and 0.8, but at 0.7 (30% missing), the error rate jumps to 19.00, and increases thereafter. For IRIS then, we can throw away about 20% of the data and maintain decent resubstitution errors for the prototype estimates.

B. Artificial Two-Dimensional (2-D) Clusters

We now report experiments involving four artificially labeled two-dimensional (2-D) data sets. These experiments are structured in exactly the same way as those for the IRIS data. The complete data sets are shown in Figs. 3–6, respectively. Incomplete data sets DS, DL, HS, and HL are generated from each of these four sets using the same procedure that we described earlier for the IRIS experiment. Since our data are now from R^2, maximally incomplete data sets occur when the percentage of missing data is 50%. As with IRIS, the initial set of prototypes $v^{(0)}$ is always taken to be $v*$, the globally optimal FCM prototypes for the original (complete) data. The last four columns of each table use $v*$ in determining the Frobenius norm error in the terminal prototypes $v^{(t)}$. We chose four cases involving small and large clusters, and horizontal and diagonal cluster placement. The problem of virtual clusters (e.g., Fig. 2) never occurs for diagonal placement of clusters because we use good initialization $(v*)$. The full data FCM prototypes $v*$ are represented by in the following scatter diagrams for each of the four complete data sets (Figs. 3–6). The number of data in each of the four (complete) artificial data sets is $n = 100$, evenly divided $(n_1 = n_2 = 50)$ between the two clusters. For the purpose of evaluating resubstitution errors, we labeled points in each data set as 1 or 2, using the obvious visual clusters.

The visually apparent structure of the two small clusters in DS (Fig. 3) is well characterized by all four methods. This is evident in the middle set of columns in Table II, which show perfect (zero) resubstitution errors for all methods and all incomplete data sets. Although the mean prototype error in Table II increases with the percentage of data values missing, there is little to distinguish any of the four methods, by this criterion—they are essentially equal (row by row). Thus, when the data are clearly well structured, all four methods are fairly equal, and the best choice in this case might be to pick WDS or PDS, which both terminate more rapidly than OCS and NPS.

We offer several observations regarding the two experiments involving diagonally placed clusters. The OCS version produces the most accurate terminal prototypes $v^{(t)}$, seen by comparisons among the last four columns of Tables II and III. The simple WDS approach consistently requires the fewest number of iterations to converge, and OCS and NPS are the most expensive in this respect. For DL, the lowest misclassification errors are consistently obtained by NPS.

For the two experiments with horizontally-placed clusters, WDS continues to converge most quickly. NPS and OCS are the overall best performers in terms of misclassification errors. PDS and WDS did well in terms of terminal prototype accuracy, as measured by the last four columns of Tables IV and V. It is interesting that throughout all the experiments the approach

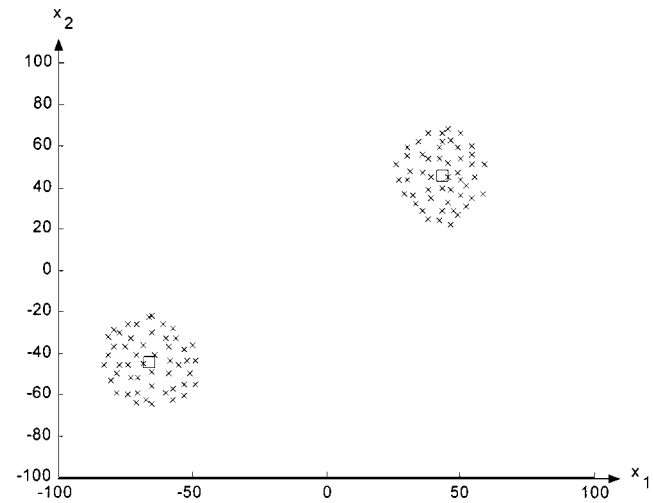


Fig. 3. Data set DS: Two Diagonally-positioned Small clusters. Full data $v*$ represented by □.

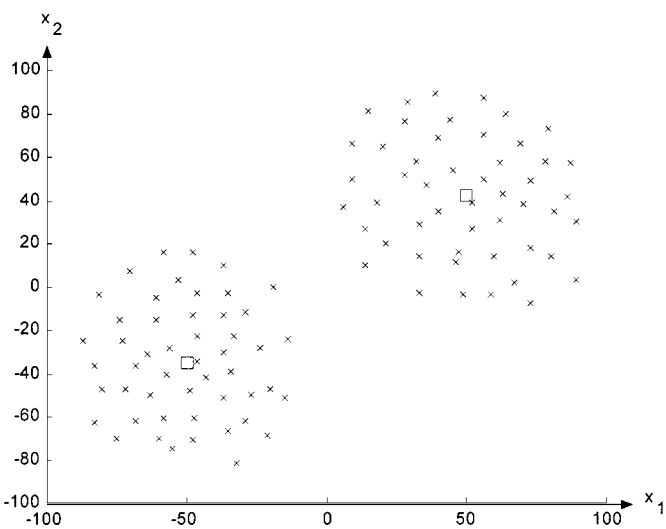


Fig. 4. Data set DL: Two Diagonally-positioned Large clusters. Full data $v*$ represented by □.

with the smallest mean prototype error does not necessarily produce the smallest number of misclassifications. Another point worth mentioning is that the training error rates for the horizontal clusters (Tables IV and V) are much higher with all four methods than the corresponding error rates for the two sets of diagonally placed clusters. For example, the *maximum* error rate from Tables II and III is 5.28%, committed by OCS on data set DL with half of its values missing. On the other hand, this error rate (5.28%) is exceeded by all four methods for both HS and HL when only 20% of their values are missing. And the maximum error rate for the horizontally aligned clusters in Tables IV and V is 21.26% (OCS, 50% row of Table V); this is more than four times the maximum for Tables II and III. So, while "diagonal" clusters can be "harder" than "horizontal" clusters in the sense illustrated by Fig. 2, the numerical evidence is (expectedly) to the contrary when good initializations are used.

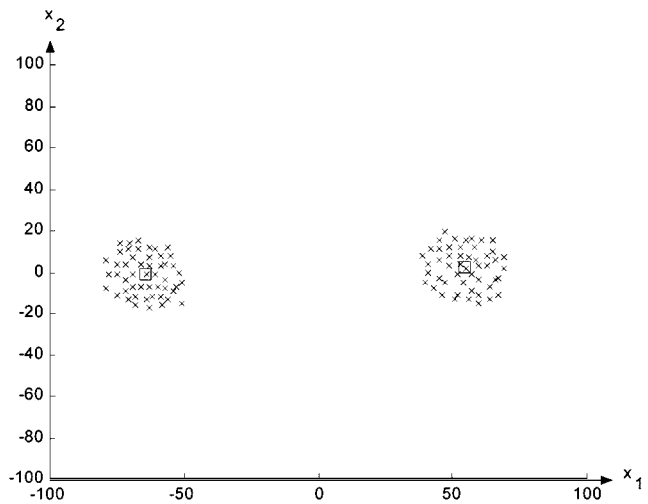


Fig. 5. Data set HS: Two Horizontally-positioned Small clusters. Full data $v*$ represented by □.

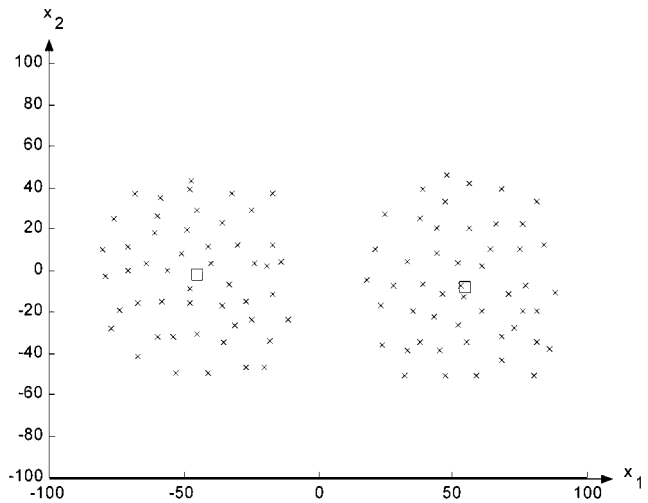


Fig. 6. Data set HL: Two Horizontally-positioned Large clusters. Full data $v*$ represented by □.

C. Artificial Five-Dimensional Clusters

Our final experiment uses a data set consisting of two clusters, each consisting of 500 points in R^5. The first cluster's points are randomly distributed according to a Gaussian (normal) distribution with mean $\mu_1 = (-1, -1, -1, -1, -1)^T$ and identity covariance; the second cluster's points are also Gaussian with identity covariance, but with mean $\mu_2 = (1, 1, 1, 1, 1)^T$. The MATLAB function mvnrnd was used to generate the data. We include this larger example to be sure that the observations made earlier about the various approaches "scale up" for more realistically sized problems.

These experiments are structured in a similar way as those for the earlier tests except that only 25 trials are done. For this data set we include a comparison with a general data completion strategy called imputation which is based on the regularized EM approach described in [24]. The interested reader is referred to [24] for details on that particular EM approach and to the web site: http://www.math.nyu.edu/~tapio/imputation/ for the

TABLE II
AVERAGED RESULTS OF 50 TRIALS USING INCOMPLETE DS DATA SETS

% Missing	Mean Number of Iterations to Termination				Mean Number of Misclassifications (Training Error)				Mean Prototype Error			
	WDS	PDS	OCS	NPS	WDS	PDS	OCS	NPS	WDS	PDS	OCS	NPS
0	1	1	1	1	0	0	0	0	0.00	0.00	0.00	0.00
10	4.84	4.94	7.36	7.36	0	0	0	0	1.42	0.97	0.96	0.96
20	4.98	5.04	9.88	9.86	0	0	0	0	2.55	1.54	1.53	1.53
30	4.98	5.40	12.88	12.90	0	0	0	0	3.78	1.96	1.92	1.92
40	5.02	5.78	17.34	17.32	0	0	0	0	6.05	2.59	2.53	2.53
50	N/A	6.44	22.92	22.88	N/A	0	0	0	N/A	3.12	3.02	3.02

TABLE III
AVERAGED RESULTS OF 50 TRIALS USING INCOMPLETE DL DATA SETS

% Missing	Mean Number of Iterations to Termination				Mean Number of Misclassifications (Training Error)				Mean Prototype Error			
	WDS	PDS	OCS	NPS	WDS	PDS	OCS	NPS	WDS	PDS	OCS	NPS
0	1	1	1	1	0	0	0	0	0.00	0.00	0.00	0.00
10	8.82	8.98	10.80	10.72	0.92	0.90	0.90	0.90	3.85	2.53	2.59	2.59
20	9.06	10.36	13.68	13.62	2.06	2.06	2.06	2.06	5.71	4.02	3.82	3.92
30	9.74	12.28	17.90	17.80	3.04	3.00	3.00	2.94	8.54	5.60	5.05	5.27
40	10.22	15.58	21.58	21.28	3.98	3.82	3.88	3.80	14.53	6.52	5.54	5.82
50	N/A	22.58	27.74	27.96	N/A	5.10	5.28	5.06	N/A	9.96	6.79	7.92

TABLE IV
AVERAGED RESULTS OF 50 TRIALS USING INCOMPLETE HS DATA SETS

% Missing	Mean Number of Iterations to Termination				Mean Number of Misclassifications (Training Error)				Mean Prototype Error			
	WDS	PDS	OCS	NPS	WDS	PDS	OCS	NPS	WDS	PDS	OCS	NPS
0	1	1	1	1	0	0	0	0	0.00	0.00	0.00	0.00
10	4.66	7.18	7.42	7.42	3.98	4.00	4.02	4.00	1.23	1.04	1.26	1.27
20	4.90	11.32	10.16	10.14	8.68	8.64	8.62	8.64	2.13	1.90	2.53	2.54
30	4.84	14.58	13.16	13.22	13.46	13.38	13.32	13.30	2.82	3.31	4.37	4.40
40	4.96	15.32	17.56	17.68	18.78	17.00	16.68	16.76	4.77	6.12	6.65	6.68
50	N/A	16.56	23.04	23.08	N/A	21.04	20.96	21.02	N/A	9.70	9.14	9.16

MATLAB routine (called regem) that we used to do this completion. For convenience and consistency in the table and discussion below, we will use EMS (EM strategy) to refer to this additional approach of completing data using regem (using default program parameter choices) and then clustering the completed data using the (regular) FCM algorithm. We do not assert that regem is the most appropriate EM-based scheme for use as a clustering tool. We include EMS here as means for comparison of our approaches with an existing alternative for data completion on a larger problem such as this.

Incomplete data sets from the mixed Gaussian data are generated using the same procedure described for earlier experiments. Since our data are now from R^5, maximally incomplete data sets occur when the percentage of missing data is 80%, but we economized somewhat in the study by only considering missing percentages up to 50%. As done earlier, the initial set of prototypes $v^{(0)}$ is always taken to be $v*$, the globally optimal FCM prototypes for the original (complete) data. The first group of five columns of Table VI gives the mean number of misclassifications, using the same reclassification scheme as the earlier experiments. For the purpose of evaluating resubstitution errors, we labeled points in each data set as 1 or 2, depending on whether they were generated from the Gaussian with mean μ_1 or μ_2. The last group of

TABLE V
AVERAGED RESULTS OF 50 TRIALS USING INCOMPLETE HL DATA SETS

% Missing	Mean Number of Iterations to Termination				Mean Number of Misclassifications (Training Error)				Mean Prototype Error			
	WDS	PDS	OCS	NPS	WDS	PDS	OCS	NPS	WDS	PDS	OCS	NPS
0	1	1	1	1	0	0	0	0	0.00	0.00	0.00	0.00
10	10.46	15.72	13.14	13.06	4.50	4.46	4.34	4.34	3.67	3.30	4.58	4.59
20	11.20	25.96	16.76	16.70	9.46	9.42	8.58	8.58	6.08	7.19	9.55	9.61
30	11.10	25.58	19.92	19.96	14.08	13.20	13.08	12.88	7.87	12.75	15.59	15.75
40	13.10	18.02	23.98	24.08	20.08	18.10	17.92	17.82	14.78	21.21	21.82	22.05
50	N/A	21.32	26.70	26.98	N/A	21.18	21.26	20.96	N/A	30.40	28.37	28.68

TABLE VI
AVERAGED RESULTS OF 25 TRIALS USING INCOMPLETE MIXED GAUSSIAN DATA SETS

% Missing	Mean Number of Misclassifications (Training Error)					Mean Prototype Error				
	WDS	PDS	OCS	NPS	EMS	WDS	PDS	OCS	NPS	EMS
0	15	15	15	15	15	0.00	0.00	0.00	0.00	0.00
10	21.84	21.32	21.44	21.48	21.44	0.12	0.05	0.05	0.04	0.06
20	30.24	31.08	31.12	31.08	30.76	0.21	0.09	0.07	0.08	0.11
30	39.48	39.20	39.28	39.24	39.64	0.31	0.13	0.09	0.11	0.17
40	56.48	54.52	54.08	54.36	54.60	0.57	0.17	0.10	0.13	0.25
50	86.84	74.60	74.16	74.44	74.44	1.12	0.24	0.14	0.19	0.36

five columns gives the Frobenius norm error in the terminal prototypes $v^{(t)}$ as compared to $v*$.

Notice that even with 0% missing data, there are 15 misclassifications, which means 15 of the 1000 generated points are probably slightly nearer the other distributional mean than their own. We see from Table VI that EMS is virtually indistinguishable in misclassification performance from all other approaches except WDS. In terms of mean prototype error, EMS is consistently fourth and significantly inferior to PDS, NPS, and especially OCS. While columns for iteration numbers are not given in this table, the overall mean number of iterations required for the five approaches with the mixed Gaussian data are:

2) WDS(8.5);
3) PDS(7.8);
4) OCS(20.6);
5) NPS(12.5);
6) EMS(7.8).

For the first four strategies, this is fairly consistent with iteration requirements on earlier, smaller examples. The low value of 7.8 for EMS does not accurately reflect its total computational cost because considerable time is required to do the data completion using the MATLAB function regem. For example, timings collected for one trial at the level of 10% missing data showed that 300.88 s were required for data completion and only 0.17 s required for the subsequent application of FCM (seven iterations) to the completed data.

IV. CONCLUSION

We considered four different approaches for doing FCM clustering of incomplete data sets. The PDS FCM uses an approach recommended by Dixon to alter the D_{ik} calculations in a way that uses all available information. The OCS FCM treats the missing data X_M as variables that are to be optimized in order to obtain the smallest possible value of the FCM functional J_m. The NPS FCM dynamically estimates the missing data values based on component values of the nearest current prototype. The convergence properties of PDSFCM and OCSFCM exactly parallel those for standard FCM, and they follow from existing alternating optimization [20], [23] and Zangwill's [22] general theory. In practice, all techniques terminated in all cases, but theoretical convergence results for NPSFCM are missing. An important implication of our numerical experiments is that although the simple approach of deleting incomplete data (WDSFCM) works fine for small percentages of missing data (less than, say 15–20%), the other approaches are generally superior if a larger proportion of data is incomplete. Without exception, OCS produced the most accurate terminal prototypes in the case of maximally incomplete data sets.

Several interesting questions and possible areas of generalization are apparent. There are incomplete forms of data other than the case of missing feature values discussed here. For example, it may be that the exact value x_{kj} is unknown but that it is known

to satisfy $x_{kj} \geq 10$. This type of data could arise in an aborted experiment that had attempted to record the time required for some event to occur for object o_k. In this case, the PDS and NPS adaptations are not obvious, but a reasonable OCS approach is to simply add the constraint $x_{kj} \geq 10$ to the optimization over X_M. (These constraints are easily implemented and add very little computational overhead to what is already required in the missing feature value case.) Because of examples such as this, we view the generalization potential of OCS to be greater than that of the other strategies. In addition to studying generalizations involving different types of incomplete data, we are interested in examining various PDS, OCS, and NPS versions of other pattern recognition techniques such as possibilistic clustering [16], noise clustering [17], and relational data clustering [18].

Finally, we plan future work aimed at understanding the relationship between OCS and the EM algorithm for computing statistical maximum-likelihood estimates in the case of incomplete data. Both EM and the obvious extensions of OCS are general approaches that essentially replace some missing part of the data with estimates. The OCS as applied to FCM can be interpreted in two ways. Here we have emphasized that it estimates the missing data X_M in a way that leads to the smallest possible value for J_m. Another interpretation, based on examination of Step OCSFCM-5, is that we are replacing unknown data with an estimate corresponding to a fuzzy mean. (A somewhat similar estimation technique was employed in [6].) This second interpretation is provocatively similar to the EM approach of replacing unknown data with expected values. Is there a real connection here? Is there a general computational approach for fuzzy models that is analogous to the EM approach for statistical models? If yes, is such an approach amenable to very general analysis and application in the same way that EM is? These questions are the subject of our current work.

REFERENCES

[1] J. K. Dixon, "Pattern recognition with partly missing data," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, pp. 617–621, Oct. 1979.

[2] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[3] G. S. Sebestyen, *Decision-Making Processes in Pattern Recognition*. New York: Macmillan, 1962, p. 74.

[4] N. G. Zagoruiko and V. N. Yolkina, "Inference and data tables with missing values," in *Handbook of Statistics*, P. R. Krishnaiah and L. N. Kanal, Eds. Amsterdam, The Netherlands: North-Holland, 1982, vol. 2, pp. 493–500.

[5] R. C. T. Lee *et al.*, "Application of clustering to estimate missing data and improve data integrity," in *Proc. 2nd ISEC*, San Francisco, CA, 1976, pp. 539–544.

[6] J. C. Bezdek and K. Solomon, "Simulation of implicit numerical characteristics using small samples," in *Proc. ICASRC*, vol. VI, G. E. Lasker, Ed., 1981, pp. 2773–2784.

[7] J. Kittler, "Classification and incomplete pattern vectors using modified discriminant functions," *IEEE Trans. Comput.*, vol. C-27, pp. 367–375, Apr. 1978.

[8] C. Z. Janikow, "Fuzzy decision trees: Issues and methods," *IEEE Trans. Syst., Man, Cybern.*, vol. B28, pp. 1–14, Feb. 1998.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc.*, vol. B39, pp. 1–38, 1977.

[10] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. New York: Wiley, 1997.

[11] G. J. McLachlan and K. E. Basford, *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker, 1988.

[12] J. L. Schafer, *Analysis of Incomplete Multivariate Data*. London, U.K.: Chapman & Hall, 1997.

[13] R. J. A. Little and D. B. Rubin, *Statistical Analysis With Missing Data*. New York: Wiley, 1987.

[14] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum, 1981.

[15] J. C. Bezdek, J. M. Keller, R. Krishnapuram, L. I. Kuncheva, and N. R. Pal, "Will the real IRIS data please stand up?," *IEEE Trans. Fuzzy Syst.*, vol. 7, pp. 368–369, June 1999.

[16] R. Krishnapuram and J. M. Keller, "A possibilistic approach to clustering," *IEEE Trans. Fuzzy Syst.*, vol. 1, pp. 98–110, May 1993.

[17] R. N. Dave, "Characterization and detection of noise in clustering," *Pattern Recognit. Lett.*, vol. 12, pp. 657–664, Nov. 1991.

[18] R. J. Hathaway and J. C. Bezdek, "NERF c-means: Non-Euclidean relational fuzzy clustering," *Pattern Recognit.*, vol. 27, pp. 429–437, Mar. 1994.

[19] J. C. Bezdek, J. M. Keller, R. Krishnapuram, and N. R. Pal, *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*. Norwell, MA: Kluwer, 1999.

[20] J. C. Bezdek, R. J. Hathaway, R. E. Howard, C. A. Wilson, and M. P. Windham, "Local convergence analysis of a grouped variable version of coordinate descent," *J. Optim. Theory Appl.*, vol. 54, pp. 471–477, Sept. 1987.

[21] J. C. Bezdek, R. J. Hathaway, M. J. Sabin, and W. T. Tucker, "Convergence theory for fuzzy c-means: Counterexamples and repairs," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-17, pp. 873–877, Oct. 1987.

[22] W. Zangwill, *Nonlinear Programming: A Unified Approach*. Englewood Cliffs, NJ: Prentice-Hall, 1969.

[23] R. J. Hathaway, Y. Hu, and J. C. Bezdek, "Local convergence of tri-level alternating optimization," *Neural, Parallel, Sci. Computat.*, vol. 9, pp. 19–28, Mar., 2001.

[24] T. Schneider, "Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values," *J. Clim.*, vol. 14, pp. 853–871, Mar. 2001.

Richard J. Hathaway (M'93) received the B.S. degree in applied math from the University of Georgia, Athens, in 1979 and the Ph.D. degree in mathematical sciences from Rice University, Houston, TX, in 1983.

He is currently a Professor in the Mathematics and Computer Science Department at Georgia Southern University, Statesboro. His research interests include pattern recognition and numerical optimization.

James C. Bezdek (M'80–SM'90–F'92) received the B.S.C.E. degree from the University of Nevada, Reno, in 1969, and the Ph.D. degree in applied math from Cornell University, Ithaca, NY, in 1973.

He is currently a Professor in the Computer Science Department at the University of West Florida, Pensacola. His interests include woodworking, optimization, motorcycles, pattern recognition, gardening, fishing, image processing, computational neural networks, blues music, and computational medicine.

Dr. Bezdek is the Founding Editor of IEEE TRANSACTIONS ON FUZZY SYSTEMS and the *International Journal of Approximate Reasoning*. He is a Fellow of the IFSA.

Kumpulan jurnaL Ilmu Komputer (KLIK)
Volume 01, No.01 September 2014
ISSN: 2406-7857

IMPLEMENTASI METODE FUZZY C-MEANS PADA SISTEM CLUSTERINGDATA VARIETAS PADI

Nurjanah[1], Andi Farmadi[2], Fatma Indriani[2]
[1,2,3] Prodi Ilmu Komputer Fakultas MIPA Universitas Lambung Mangkurat

Jl. A. Yani Km 36 Banjarbaru, Kalimantan Selatan
Email: nurjanah_abs@yahoo.co.id

Abstract

Mutations with gamma rays conducted on five local rice varieties tidal South Kalimantan produce a lot of data availability. In order for these data not only become a graveyard of useless data required a method that could be used to probe the hidden information from the data. The method known as data mining. Data mining is a technique to gain knowledge from the data by looking for certain patterns or rules of a number of large amounts of data. One method of data mining is clustering, where clustering is usually used to group objects that are similar in the same class or segment. By utilizing the data of local rice varieties tidal South Kalimantan mutated by gamma rays, data mining process is done by grouping the data based on the harvest age, productive tillers, and weight of 1000 seeds into 4 groups using fuzzy c-means algorithm. From that cluster information, carried ranking using the Simple Additive Weighting method and acquired knowledge about improved varieties by harvest age, productive tillers, and a weight of 1000 is kuatek with a dose of 30 krad.

Keywords *: Data Mining, Cluster, Fuzzy C-Means, Local Rice Varieties*

Abstrak

Mutasi dengan sinar gamma yang dilakukan terhadap lima varietas padi lokal pasang surut kalimantan selatan menghasilkan tersedianya banyak data. Agar data-data tersebut tidak hanya menjadi kuburan data yang tidak berguna dibutuhkan sebuah metode yang bisa digunakan untuk menggali informasi–informasi tersembunyi dari data tersebut. Metode tersebut dikenal dengan data mining. Data mining merupakan suatu teknik untuk menggali pengetahuan dari data dengan mencari pola atau aturan tertentu dari sejumlah data dalam jumlah besar. Salah satu metode data mining adalah klastering, dimana klastering biasanya digunakan untuk mengelompokan objek-objek yang memiliki kemiripan dalam kelas atau segmen yang sama. Dengan memanfaatkan data varietas padi hasil mutasi dengan sinar gamma dilakukan proses penggalian data dengan cara mengelompokkan data umur panen, anakan produktif, dan bobot 1000 menjadi 4 klaster menggunakan algoritma fuzzy c-means. Dari informasi 4 klaster tersebut, dilakukan perangkingan klaster menggunakan metode Simple Additive Weighting dan diperoleh pengetahuan tentang varietas unggul berdasarkan umur panen, anakan produktif, dan bobot 1000 yaitu kuatek dengan dosis 30 krad.

Kata kunci *: Data Mining, Klaster, Fuzzy C-Means, Varietas Padi Lokal*

1. PENDAHULUAN

Varietas padi lokal Kalimantan Selatan masih menjadi pilihan para petani untuk ditanam, yang dalam penanamannya tidak terlalu bergantung pada input pertanian. Selain itu, varietas padi lokal juga bersifat adaptif dengan lingkungan Kalimantan Selatan yang didominasi oleh lahan rawa, meskipun berumur panjang dan produktivitas relatif rendah daripada varietas padi unggul. Akan tetapi sifat "pera" nasi yang dihasilkan membuat harga jual varietas padi lokal lebih tinggi dibandingkan dengan varietas padi unggul [3].

Perjuangan dalam peningkatan produksi pangan untuk memenuhi kebutuhan masyarakat Kalimantan Selatan merupakan tantangan bagi masyarakat pertanian yang dapat melahirkan peluang berupa pengembangan teknologi alternatif untuk menerobos kebuntuan dan stagnasi produksi beras lokal. Alternatif itu dapat diarahkan pada pengembangan varietas lokal yang dapat beradaptasi pada lingkungan produksi yang buruk dengan produktivitas yang tinggi dan umur panen yang pendek. Peningkatan keragaman populasi dapat dilakukan salah satunya melalui induksi mutasi secara fisik dengan iradiasi sinar gamma.

Salah satu upaya untuk meningkatkan produksi pangan varietas padi lokal pasang surut kalimantan selatan yaitu dengan cara meningkatkan keragaman populasi, dilakukan radiasi dengan sinar gamma terhadap lima varietas padi. Dari upaya tersebut dihasilkan banyak data salah satunya yaitu data varietas padi hasil mutasi dengan sinar gammamaka dibutuhkan sebuah metode yang bisa digunakan untuk menggali informasi–informasi tersembunyi dari data tersebut. Metode tersebut dikenal dengan *data mining*.

Data mining adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar. Salah satu tugas dalam data mining adalah klastering. Tujuan utama dari klastering adalah pengelompokan sejumlah data/obyek ke dalam klaster sehingga dalam setiap klaster akan berisi data yang semirip mungkin [1].

Penggalian suatu data menjadi pengetahuan adalah kegiatan yang sangat berguna, karena pengetahuan tersebut bisa digunakan untuk memecahkan suatu masalah yang ada. *Data mining* merupakan salah satu teknik untuk menggali pengetahuan dari data. Agar data-data tersebut tidak menjadi kuburan data yang tidak bergunamaka dibuat sistem yang mampu menggali data varietas padi yang dimutasi dengan sinar gamma yang dapat mengorganisir data varietas padi yang dimutasi dengan sinar gamma dalam suatu database serta dapat menghasilkan pengelompokan sejumlah data/obyek ke dalam klaster yang memberikan pengetahuan mengenai varietas unggulmenggunakan metode *Fuzzy C-Means* (FCM).

2. METODOLOGI PENELITIAN

2.1. Analisa Data

Pada tahap ini akan dilakukan analisis data dengan KDD (*knowledge Discovery in Databases*). Keseluruhan proses KDD dari mulai memasukan data mengalami 3 tahap secara global sebelum menjadi hasil yang berupa informasi, yaitu:

(a) *Preprocessing*

Pada tahap ini struktur basis data akan dipersiapkan sehingga mempermudah proses mining. Proses *preprocessing* ini mencakup tiga hal utama yaitu :

1) seleksi data
   untuk memilih data yang akan digunakan dalam proses data mining. Dalam penelitian ini dipilih 3 data varietas padi yaitu umur panen, anakan produktif, dan bobot 1000 biji;
2) *Cleaning*
   Memastikan kualitas data yang telah dipilih pada tahap seleksi data, pada tahap ini masalah yang harus dihadapi adalah noisy data dan missing values. Proses cleaning data perlu dilakukan agar data bersih dari duplikasi data, data yang tidak konsisten, atau kesalahan cetak. Sehingga data yang telah melewati proses ini siap untuk di proses di data mining. Pada penelitian ini proses pembersihan data hanya dilakukan pada data yang missing value;
3) integrasi data
   Setelah melakukan pembersihan data yang menyeluruh, maka akan dilakukan tahapan selanjutnya yaitu integrasi data, karena data lebih dari satu tabel maka dilakukan tahap penggabungan data dari berbagai tabel. Kemudian data padi diberi kode agar tidak ada data yang memiliki nama yang sama. Kode padi terdiri dari nomor padi, baris padi, dosis radiasi, dan varietas padi.

(b) Data Mining

Proses data mining yang dilakukan pada penelitian ini adalah menerapkan teknik *clustering* dengan algoritma *Fuzzy C-Means* untuk mendapatkan karakteristik dari data. *Cluster* yang digunakan pada penelitian ini sebanyak empat klaster.

Konsep dasar FCM (*Fuzzy C-Means*) pertama kali adalah menentukan pusat *cluster* yang akan menandai lokasi rata-rata untuk tiap *cluster*. Pada kondisi awal, pusat *cluster* ini masih belum akurat. Tiap-tiap titik data memiliki derajat keanggotan untuk tiap-tiap *cluster*. Dengan cara memperbaiki pusat *cluster* dan derajat keanggotaan tiap-tiap titik data secara berulang, maka akan dapat dilihat bahwa pusat *cluster* akan bergerak menuju lokasi yang tepat. Perulangan ini didasarkan pada derajat keanggotaan yang menggambarkan jarak dari titik data yang diberikan ke pusat *cluster* yang terbobot oleh derajat keanggotaan titik data tersebut [6].

Pada perhitungan menggunakan FCM dengan menggunakan data yang sama tetapi diolah dengan jumlah *cluster* yang berbeda, maka hasil pengelompokannya akan sedikit berbeda, karena data tidak diolah dengan satu variabel saja tetapi dengan semua variabel. Perbedaan hasil pengelompokan itu dikarenakan data pada kelompok tertentu kemungkinan akan berpindah pada kelompok lain jika diolah dengan jumlah *cluster* yang berbeda, ini menunjukkan bahwa sistem aplikasi sudah berjalan dengan benar [6]

Output dari FCM bukan merupakan fuzzy inference sistem, namun merupakan derajat pusat *cluster* dan beberapa derajat keanggotaan untuk

tiap-tiap titik data. Algoritma *Fuzzy C-Means* disusun dengan langkah sebagai [6]:

(1) Input data
input data yang akan di*cluster*yaitu data varietas padi (X) berupa matrik berukuran n × m (n = jumlah data, m = atribut setiap data). X_{ij} = data ke- i (i = 1, 2, …, n), atribut ke- j (j = 1, 2, …, m).

(2) Batasan
(a) Jumlah *cluster* = c = 4
(b) Pangkat = w = 2
(c) Maksimum iterasi = Maxit = 100
(d) Error terkecil yang diharapkan = ξ = 10^{-3}
(e) Fungsi obyektif awal = Po = 0
(f) Iterasi awal = t = 1

(3) Membangkitkan bilangan random μ_{ik}, i = 1, 2, …, n; k = 1, 2, …,c; sebagai elemen-elemen matrik partisi awal U, dengan jumlah setiap nilai elemen kolom dalam satu baris adalah 1 (satu)

$$\sum_{i=1}^{c} \mu_{ci} = 1 \qquad \dots (1)$$

(4) Menghitung pusat *cluster* ke- k : V_{kj} dengan k = 1, 2, …, c; dan j = 1, 2, …,m.

$$V_{kj} = \frac{\sum_{i=1}^{n}\left((\mu_{ik})^{w} * X_{ij}\right)}{\sum_{i-1}^{n}(\mu_{ik})^{w}} \qquad \dots (2)$$

(5) Menghitung fungsi obyektif pada iterasi ke- t, P_t :

$$P_t = \sum_{i=1}^{n}\sum_{k=1}^{c}\left(\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right](\mu_{ik})^{w}\right) \qquad \dots (3)$$

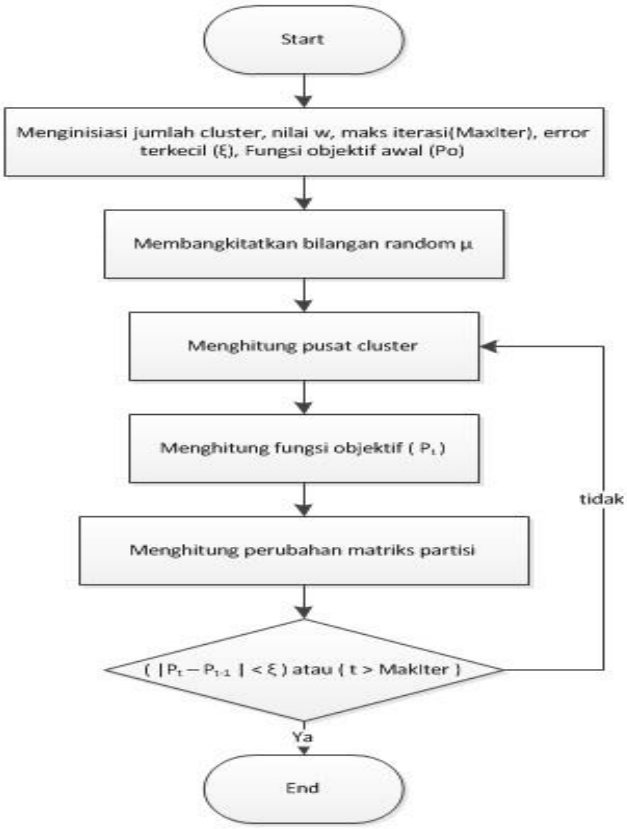
(6) Menghitung perubahan matrik partisi:

$$\mu_{ik} = \frac{\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right]^{\frac{-1}{w-1}}} \dots (4)$$

dengan i = 1, 2, …, n; dan k = 1, 2, …, c

(7) Mengecek kondisi berhenti:
1. Jika : ($|P_t - P_{t-1}| < \xi$) atau (t > MakIter) maka berhenti;
2. Jika tidak : t = t + 1, ulangi langkah ke- d (menghitung V_{kj}).

Bagan alir untuk metode *Fuzzy C-Means* yaitu sebagai berikut :



Gambar 1. Bagan Alir Metode *Fuzzy C-Means*

(c) ***Postprocessing***

Postprocessing bertujuan untuk membantu pengguna dalam memahami informasi, Karena pusat *cluster* masih tidak bisa dienterpretasikan secara langsung, maka diperlukan metode *Simple Additive Weighting* yang bisa melakukan perangkingan nilai pusat *cluster* terakhir tersebut.

Metode SAW (*Simple Additive Weighting*) sering juga dikenal istilah metode penjumlahan terbobot. Konsep dasar metode SAW adalah mencari penjumlahan terbobot dari rating kinerja pada setiap alternatif pada semua atribut. Langkah-langkah metode dalam metode SAW adalah [7] :

1. Membuat matriks keputusan Z berukuran m x n, dimana m = alternatif yang akan dipilih dan n = kriteria .
2. Memberikan nilai x setiap alternatif (i) pada setiap kriteria (j) yang sudah ditentukan, dimana, i=1,2,…m dan j=1,2,…n pada matriks keputusan Z,

$$\mathbf{Z} = \begin{vmatrix} x_{11} & x_{y12} & \dots & x_{1j} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ x_{i1} & x_{i2} & & x_{ij} \end{vmatrix} \dots (5)$$

3. Memberikan nilai bobot preferensi (W) oleh pengambil keputusan untuk masing-masing kriteria yang sudah ditentukan.

$$W = [W_1\ W_2 W_3 \dots\ W_j] \dots(6)$$

4. Melakukan normalisasi matriks keputusan Z dengan cara menghitung nilai rating kinerja ternormalisasi (r_{ij}) dari alternatif A_i pada atribut C_j

$$r_{ij} = \begin{cases} \frac{x_{ij}}{\underset{i}{MAX}(x_{ij})} & \text{jika j adalah atribut keuntungan} \\ \frac{\underset{i}{MIN}(x_{ij})}{x_{ij}} & \text{jika j adalah atribut biaya} \end{cases} \dots(7)$$

Dengan ketentuan :

a. Dikatakan atribut keuntungan apabila atribut banyak memberikan keuntungan bagi pengambil keputusan, sedangkan atribut biaya merupakan atribut yang banyak memberikan pengeluaran jika nilainya semakin besar bagi pengambil keputusan.
b. Apabila berupa atribut keuntungan maka nilai (x_{ij}) dari setiap kolom atribut dibagi dengan nilai (MAX x_{ij}) dari tiap kolom, sedangkan untuk atribut biaya, nilai (MIN x_{ij}) dari tiap kolom atribut dibagi dengan nilai (x_{ij}) setiap kolom.

5. Hasil dari nilai rating kinerja ternormalisasi (r_{ij}) membentuk matriks ternormalisasi (N)

$$\mathbf{N} = \begin{vmatrix} r_{11} & r_{y12} & \dots & r_{1j} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ r_{i1} & r_{i2} & & r_{ij} \end{vmatrix} \dots (8)$$

6. Melakukan proses perankingan dengan cara mengalikan matriks ternormalisasi (N) dengan nilai bobot preferensi (W).
7. Menentukan nilai preferensi untuk setiap alternatif (V_i) dengan cara menjumlahkan hasil kali antara matriks ternormalisasi (N) dengan nilai bobot preferensi (W).

$$V_i = \sum_{j=1}^{n} w_j r_{ij} \dots (9)$$

Nilai Vi yang lebih besar mengindikasikan bahwa alternatif Ai merupakan alternatif terbaik.

2.3. Implementasi

Implementasi dilakukan dengan metode Fuzzy C-Means (FCM) pada sistem clustering data varietas padilokal pasang surut kalimantan selatan yang dimutasi dengan sinar gamma menggunakan delphi XE2

2.4. Pengujian

Untuk memastikan apakah perhitungan dengan metode fuzzy c-means sudah benar dilakukan uji coba perhitungan secara manual meggunakan MS. Excel dan Matlab.Pengujian dilakukan untuk menghindari kesalahan dari sistem yang dibuat. Apabila terjadi kesalahan, maka sistem akan diperbaiki kembali sampai hasil proses sesuai dengan apa yang diharapkan.

3. HASIL DAN PEMBAHASAN

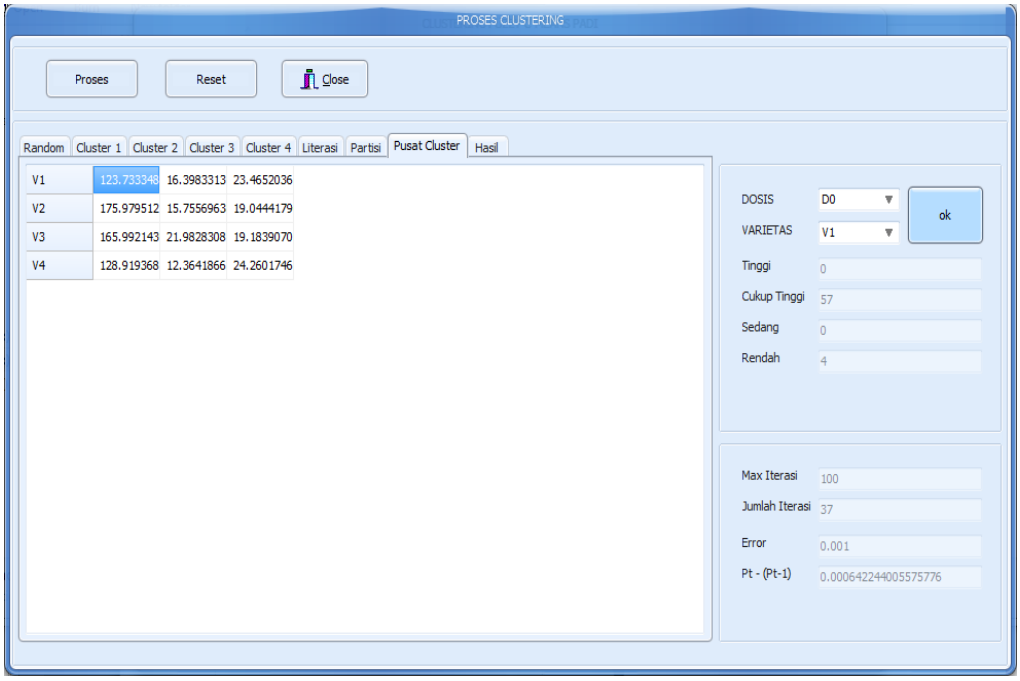
Dariproses clustering Fuzzy C-Means dengan batasan jumlah cluster sebanyak 4 cluster, pangkat 2, masimal iterasi 100, error 0,001 proses berhenti pada iterasi ke-37 dengan error 0,000642 dan diperoleh pusat cluster iterasi terakhir yaitu pada Tabel 1 atau pada Gambar 2 dengan menggunakan rumus dibawah ini (persamaan 2):

$$V_{kj} = \frac{\sum_{i=1}^{2420}\left((\mu_{ik})^2 * X_{ij}\right)}{\sum_{i=1}^{2420}(\mu_{ik})^2}$$

Keterangan :
V_{kj}: pusat cluster dengan k=1,2,...,c dan j = 1,2,...,n
X_{ij} : data ke-i (i=1,2,...m), atribut ke-j (j=1,2,...n)
μ_{ik} : bilangan acak dengan i=1,2,...m dan k=1,2,...c

Tabel 1. Pusat Cluster Delphi

	Umur Panen	Anakan Produktif	Bobot 1000
Klaster 1	165.992143	21.9828308	19.1839070
Klaster 2	123.733348	16.3983313	23.4652036
Klaster 3	128.919368	12.3641866	24.2601746
Klaster 4	175.979512	15.7556963	19.0444179



Gambar 2. Form Proses Fuzzy C-Means

Dari Tabel 1 dapat dilihat bahwa klaster 1 memiliki umur panen terendah ketiga, anakan produktif terbanyak pertama, dan bobot 1000 terberat ketiga. Klaster 2 memiliki umur panen terendah pertama, anakan produktif terbanyak kedua, dan bobot 1000 terberat kedua. Klaster 3 memiliki umur panen terendah kedua, anakan produktif terbanyak keempat, dan bobot 1000 terberat pertama.

Klaster 4 memiliki umur panen terendah keempat, anakan produktif terbanyak ketiga, dan bobot 1000 terberat keempat.

Dilihat dari umur panen klaster 2 lebih unggul karena memiliki umur panen terendah dibandingkan klaster 1, klaster 3, dan klaster 4. Untuk anakan produktif klaster 1 lebih unggul karena memiliki anakan produktif terbanyak. Dan untuk bobot 1000 klaster 3 lebih unggul karena memiliki bobot seribu terberat apabila dibandingkan dengan klaster lainnya.

Dari hasil pusat klaster (Tabel 1) tersebut dilakukan perangkingan dengan metode *Simple Additive Weighting* (SAW) untuk melihat klaster mana yang memiliki umur panen, anakan produktif dan bobot 1000 terbaik. Perhitungan nilai prefrensi dilakukan menggunakan rumus dibawah ini (persamaam 9) :

$$V_i = \sum_{j=1}^{n} w_j r_{ij}$$

Keterangan :
V_i : nilaipreferensiuntuksetiapalternatif
w_j : nilai bobot preferensi
r_{ij} : matriks ternormalisasi

Berikut tabel nilai prefrensi dan tabel hasil pengelompokan per varietas :

Tabel 2. Nilai Prefensi Klaster

Klaster	Nila Prefensi	Kategori
Klaster 1	2.535898966	Cukup Tinggi
Klaster 2	2.712760587	Tinggi
Klaster 3	2.53314209	Sedang
Klaster 4	2.209208036	Rendah

Tabel 3. Hasil Pengelompokan Varietas Siam Harli

Kategori	Varietas Siam Harli (V1)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	-	-	-	46
Cukup Tinggi	57	77	32	-
Sedang	-	-	-	44
Rendah	4	2	3	-

Tabel 4. Hasil Pengelompokan Varietas Siam Unus

Kategori	Varietas Siam Unus (V2)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	-	1	-	-
Cukup Tinggi	48	205	48	86
Sedang	-	1	4	-
Rendah	4	-	5	8

Tabel 5. Hasil Pengelompokan Varietas Kuatek

Kategori	Varietas Kuatek (V3)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	-	-	-	371
Cukup Tinggi	7	152	18	176
Sedang	-	-	-	-
Rendah	41	5	132	-

Tabel 6. Hasil Pengelompokan Varietas Siam Sebelas

Kategori	Varietas Siam Sebelas (V4)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	5	-	-	25
Cukup Tinggi	1	28	9	-
Sedang	1	-	-	22
Rendah	25	158	78	-

Tabel 7. Hasil Pengelompokan Varietas Siam Gumpal

Kategori	Varietas Siam Gumpal (V5)			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Tinggi	-	-	-	13
Cukup Tinggi	-	-	-	-
Sedang	-	-	-	116
Rendah	54	173	105	-

Pengetahuan yang didapat dari hasil penambangan data/klastering ini berupa varietas unggul yaitu varietas yang masuk kategori tinggi. Varietas yang masuk kategori tinggi terdiri atas rata-rata umur panen sebesar 123,7665737; rata-rata anakan produktif sebesar 16,2834365; dan rata-rata bobot 1000 sebesar 23,46984343 dengan nilai preferensi sebesar 2,712760587. Diantara kelima varietas tersebut yang paling banyak masuk kategori tinggi yaitu varietas kuatek dengan dosis 30 krad yakni sebanyak 371 data padi atau sebesar 67,82 % data padi masuk kategori tinggi

Tabel 8. Presentase Data Padi Yang Termasuk Kategori Tinggi

Varietas	Dosis			
	Dosis 0 krad (D0)	Dosis 10 krad (D1)	Dosis 20 krad (D2)	Dosis 30 krad (D3)
Siam Harli (V1)	-	-	-	51,11%
Siam Unus (V2)	-	0,93%	-	-
Kuatek (V3)	-	-	-	67,82%
Siam Sebelas (V4)	15,61%	-	-	53,19%
Siam Gumpal (V5)	-	-	-	10,08%

4. KESIMPULAN

Dari hasil penelitian dan pengamatan dari sistem yang telah dibuat, maka dapat diperoleh kesimpulan sebagai berikut :

1. Metode Fuzzy C-Means berhasil diimplemantasikan pada clustering data varietas padi lokal pasang surut kalimantan selatan yang dimutasi dengan sinar gamma.
2. Dilihat dari umur panen klaster 2 lebih unggul karena memiliki umur panen terendah, untuk anakan produktif klaster 1 lebih unggul karena memiliki anakan produktif terbanyak, dan untuk bobot 1000 klaster 3 lebih unggul karena memiliki bobot seribu terberat apabila dibandingkan dengan klaster lainnya.
3. Pengetahuan yang didapat mengenai varietas unggul berdasarkan umur panen, anakan produktif dan bobot 1000 terbaik menggunakan SAW memiliki nilai preferensi sebesar 23,46984343 yaitu kategori tinggi.
4. Diantara kelima varietas yang paling banyak masuk kategori tinggi yaitu varietas kuatek dengan dosis 30 krad yakni sebanyak 371 data padi atau sebesar 67,82 % data padi masuk kategori tinggi.

DAFTAR PUSTAKA

[1]. Budi Santosa, "**Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis***",* Graha Ilmu, Yogyakarta, 2007.

[2]. Iko Pramudiono. "***Pengantar Data Mining : Menambang Permata Pengetahuan di Gunung Data***" 2003.

[3]. Noor, M., Y. Rina dan Noorginayuwati, "**Persepsi petani tentang lahan gambut dan pengelolaannya",** Laporan Penelitian, Balai Penelitian Pertanian Lahan Rawa, Banjarbaru, 2007.

[4]. Nurjanah, "**Clustering Varietas Padi Lokal Pasang Surut Kalimantan Selatan Yang Dimutasi Dengan Sinar Gamma Menggunakan Metode Fuzzy C-Means**" Naskah Skripsi Universitas Lambung Mangkurat, Banjarbaru, 2014.

[5]. Nuqson Masykur Huda, "***Aplikasi Data Mining Untuk Menampilkan Informasi Tingkat Kelulusan Mahasiswa (Studi Kasus Di Fakultas MIPA Universitas Diponegoro)",*** Naskah Skripsi Universitas Diponegoro, Semarang, 2010.

[6]. Sri Kusumadewi and Hari Purnomo, "**Aplikasi Logika Fuzzy Untuk Pendukung Keputusan***",* Yogyakarta, Graha Ilmu, 2010.

[7]. Wibowo S, Henry., Amalia, Riska., Fadlun M, Andi., Arivanty, Kurnia, "**Sistem Pendukung Keputusan Untuk MenentukanPenerima Beasiswa Bank BRI Menggunakan FMADM (Studi Kasus: Mahasiswa Fakultas Teknologi Industri Unversitas Islam Indonesia)***",* Seminar Nasional Aplikasi Teknologi Informasi, Yogyakarta, 2008.

2011 IEEE International Conference on Fuzzy Systems
June 27-30, 2011, Taipei, Taiwan

On Hard and Fuzzy c-Means Clustering with Conditionally Positive Definite Kernel

Yuchi Kanzawa
Shibaura Institute of Technology
Tokyo, Japan 135-8548
Email: kanzawa@sic.shibaura-it.ac.jp

Yasunori Endo
University of Tsukuba
Ibaraki, Japan

Sadaaki Miyamoto
University of Tsukuba
Ibaraki, Japan

***Abstract*—In this paper, we investigate three types of c-means clustering algorithms with a conditionally positive definite kernel. One is based on hard c-means, and the others are based on standard and entropy-regularized fuzzy c-means. First, based on a conditionally positive definite kernel describing a squared Euclidean distance between data in the feature space, these algorithms are derived from revised optimization problems of the conventional kernel c-means. Next, based on the relationship between the positive definite kernel and conditionally positive definite kernel, the revised dissimilarity between a datum and a cluster center in the feature space is shown. Finally, it is shown that a conditionally positive definite kernel c-means algorithm and a kernel c-means algorithm with a positive definite kernel derived from the conditionally positive definite kernel are essentially identical to each other. An explicit mapping for a conditionally positive definite kernel is also described geometrically.**

***Index Terms*—Clustering, Fuzzy c-means, Conditionally positive definite kernel**

I. Introduction

Fuzzy c-means (FCM) clustering [1] is a well-known fuzzy clustering technique, and many FCM variants have been proposed. One of the authors has proposed an FCM algorithm based on the concept of entropy regularization [2]; this algorithm is known as the entropy-regularized FCM (eFCM) algorithm. It is a useful algorithm that is mathematically related to other techniques.

Three algorithms [3], [4] have been proposed to cluster data with nonlinear borders via nonlinear transformation from the original pattern space into a higher dimensional feature space with support vector machine (SVM) kernel functions [5]. They are the K-HCM algorithm derived from the hard c-means (HCM) algorithm, the K-sFCM algorithm derived from the standard FCM (sFCM) algorithm, and the K-eFCM algorithm derived from the eFCM algorithm. These three algorithms are known as kernel c-means (K-CM) algorithms. For kernel data analysis, an explicit mapping is unknown, whereas the inner product is known. An explicit mapping has been introduced by one of the authors, and it is used to describe the situation of kernel fuzzy c-means in higher dimensional space via kernel principal component analysis [6], [7].

In kernel data analysis, the kernel must be positive definite so that it describes the inner product between the data in the feature space. On the other hand, there exists a class of conditionally positive definite (cpd) kernel [8]. In this paper, we investigate c-means clustering algorithms with cpd kernel. First, we develop three c-means clustering algorithms with cpd kernel by solving the revised optimization problem of K-HCM, K-sFCM, and K-eFCM. Accordingly, the cpd kernel is the squared Euclidean distance between the data in the feature space. We call these three algorithms cpdK-HCM, cdK-sFCM, and cpdK-eFCM, respectively; and sum up to cpdK-CM. Next, we show how the cpd kernel is used in K-CM algorithms, with the positive definite kernel obtained by arranging the cpd kernel. Furthermore, we show the connection between the cpdK-CM algorithms and the K-CM algorithms with the positive definite kernel obtained from the cpd kernel, and we interpret a mapping from the original data space to the feature space, which is latent in the cpdK-CM algorithms. Finally, we show a numerical example of the cpdK-CM algorithms.

II. Preliminaries

In this section, we introduce three K-CM clustering algorithms with an explicit mapping from the original data space to the feature space [6],[7]

The data set $X = \{x_i \mid x_i \in \mathbb{R}^p, i \in \{1, \ldots, N\}\}$ is given. The membership by which x_i belongs to the j-th cluster is denoted by $u_{i,j}$ ($i \in \{1, \cdots, N\}, j \in \{1, \cdots, C\}$), and the set of $u_{i,j}$, denoted by $u \in \mathbf{R}^{N \times C}$, is known as the partition matrix. The constraint on u is

$$\sum_{j=1}^{C} u_{i,j} = 1 \quad (0 \leq u_{i,j} \leq 1).$$

A high-dimensional feature space used in SVM is denoted by $\mathbb{H}$, whereas the original space, which is known as the data space, $\mathbb{R}^p$. $\mathbb{H}$ may be infinite dimensional. Let the inner product be denoted by $\langle \cdot, \cdot \rangle$. The norm of $\mathbb{H}$ for an element $g \in \mathbb{H}$ is given by

$$\|g\|_{\mathbb{H}} = \sqrt{\langle g, g \rangle}. \tag{1}$$

A transformation $\Phi : \mathbb{R}^p \to \mathbb{H}$ is employed, whereby x_i is mapped into $\Phi(x_i)$. In general, Φ is not usable; however, the inner product $\langle \Phi(x), \Phi(y) \rangle$ can be expressed by a kernel function

$$\mathcal{K}(x, y) = \langle \Phi(x), \Phi(y) \rangle. \tag{2}$$

The matrix whose (i, j)-element is $\mathcal{K}(x_i, x_j)$ is denoted by K. The cluster center set in $\mathbb{H}$ is denoted by $W = \{W_1, \cdots, W_C\}$.

An explicit function Φ corresponding to an introduced kernel function has been proposed as

$$\Phi(x_i) = e_i \tag{3}$$

with

$$\langle e_i, e_j \rangle = e_i^{\mathsf{T}} K e_j, \tag{4}$$

where $\mathbb{H} = \mathbb{R}^N$ and e_i is the k-th unit vector whose i-th component is unity and remaining components are zero. Note that $\Phi : X \to \mathbb{R}^N$, i.e., Φ is not defined on $\mathbb{R}^p$ but on finite data X.

Three K-CM algorithms with explicit mapping are obtained by solving the following optimization problem:

$$\underset{u,W}{\text{minimize}} \sum_{i=1}^{N} \sum_{j=1}^{C} u_{i,j} \|\Phi(x_i) - W_j\|_{\mathbb{H}}^2, \tag{5}$$

$$\text{subject to } \sum_{j=1}^{C} u_{i,j} = 1, \quad u_{i,j} \in \{0, 1\}, \tag{6}$$

for K-HCM,

$$\underset{u,W}{\text{minimize}} \sum_{i=1}^{N} \sum_{j=1}^{C} u_{i,j}^m \|\Phi(x_i) - W_j\|_{\mathbb{H}}^2, \tag{7}$$

$$\text{subject to } \sum_{j=1}^{C} u_{i,j} = 1, \tag{8}$$

for K-sFCM, and

$$\underset{u,W}{\text{minimize}} \sum_{i=1}^{N} \sum_{j=1}^{C} u_{i,j} \|\Phi(x_i) - W_j\|_{\mathbb{H}}^2 + \lambda^{-1} \sum_{i=1}^{N} \sum_{j=1}^{C} u_{i,j} \log(u_{i,j}), \tag{9}$$

$$\text{subject to } \sum_{j=1}^{C} u_{i,j} = 1. \tag{10}$$

for K-eFCM. These algorithms are described as follows:

Algorithm 1 (K-CM): [6] [7]

STEP 1. Set $m > 1$ for K-sFCM. Set $\lambda > 0$ for K-eFCM. Set the kernel matrix K and initial partition matrix.

STEP 2. Update the cluster centers W:

$$W_j = \left(\sum_{i=1}^{N} u_{i,j} \right)^{-1} (u_{1,j}, \ldots, u_{N,j})^{\mathsf{T}}, \tag{11}$$

for K-HCM and K-sFCM, and

$$W_j = \left(\sum_{i=1}^{N} u_{i,j}^m \right)^{-1} (u_{1,j}^m, \ldots, u_{N,j}^m)^{\mathsf{T}}, \tag{12}$$

for K-eFCM.

STEP 3. Update the squared Euclidean distance between the cluster center W_j and data $\Phi(x_i)$ in the feature space:

$$d_{i,j} = (e_i - W_j)^{\mathsf{T}} K (e_i - W_j). \tag{13}$$

STEP 4. Update the membership $u_{i,j}$:

$$u_{i,j} = \begin{cases} 1 & (i = \arg\ \min_k d_{k,j}), \\ 0 & (\text{otherwise}), \end{cases} \tag{14}$$

for K-HCM,

$$u_{i,j} = 1 / \sum_{k=1}^{C} \left(\frac{d_{i,j}}{d_{i,k}} \right)^{1/(m-1)}, \tag{15}$$

for K-sFCM, and

$$u_{i,j} = \frac{\exp(-\lambda d_{i,j})}{\sum_{k=1}^{C} \exp(-\lambda d_{i,k})}, \tag{16}$$

for K-eFCM.

STEP 4. If (u, W, d) converge, terminate this algorithm. Otherwise, go to STEP 2.

III. CONDITIONALLY POSITIVE DEFINITE KERNEL c-MEANS

In this section, we investigate K-CM clustering algorithms with a conditionally positive definite kernel. First, we derive conditionally positive definite K-CM algorithms; thus, the conditionally positive definite kernel is the squared Euclidean distance in a feature space. Next, we show how the cpd kernel is used in a K-CM algorithm with a positive definite kernel obtained by arranging the cpd kernel. Furthermore, we determine the relation between the cpdK-CM and K-CM algorithms with a positive definite kernel obtained from a cpd kernel, and we interpret a mapping from the original data space to the feature space, which is latent in cpdK-CM algorithms. Finally, we show a numerical example of the cpdK-CM algorithms.

A. *c-Means algorithms with cpd kernel*

In this subsection, we develop c-means clustering algorithms with a cpd kernel by solving the revised optimization problems of K-HCM, K-sFCM, and K-eFCM; thus, the cpd kernel is the squared Euclidean distance between the data in the feature space.

The cpd kernel is defined as follows:

Definition 1 (cpd matrix): [8]

A symmetric $N \times N$ matrix K ($N \geq 2$) taking values in $\mathbb{R}$ and satisfying

$$\sum_{i=1}^{N} \sum_{k=1}^{N} \zeta_i \zeta_k K_{i,k} \geq 0 \tag{17}$$

for all $\zeta_i \in \mathbb{R}$, with

$$\sum_{i=1}^{N} \zeta_i = 0, \tag{18}$$

is said to be conditionally positive definite (cpd).

Definition 2 (cpd kernel): [8]

Let $\mathcal{X}$ be a nonempty set. A function $\tilde{\mathcal{K}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, which yields a conditionally positive definite Gram matrix for all $N \geq 2$, $x_1, \cdots, x_N \in \mathcal{X}$, is known as a conditionally positive definite (cpd) kernel.

A cpd kernel describes the squared Euclidean distance in a feature space as follows:

Theorem 1 (Hilbert space representation of cpd kernels): [8]

Let $\tilde{\mathcal{K}}$ be a real-valued conditionally positive definite kernel on $\mathcal{X}$, satisfying $\tilde{\mathcal{K}}(x, x) = 0$ for all $x \in \mathcal{X}$. Then, there exists a Hilbert space $\mathbb{H}$ of real-valued functions on $\mathcal{X}$ and a mapping $\Phi : \mathcal{X} \to \mathbb{H}$ such that

$$\|\Phi(x) - \Phi(x')\|_{\mathbb{H}}^2 = -\tilde{\mathcal{K}}(x, x'). \tag{19}$$

If we drop the assumption $\tilde{\mathcal{K}}(x, x) = 0$, the Hilbert space representation reads

$$\|\Phi(x) - \Phi(x')\|_{\mathbb{H}}^2 = -\tilde{\mathcal{K}}(x, x') + \frac{1}{2}(\tilde{\mathcal{K}}(x, x) + \tilde{\mathcal{K}}(x', x')). \tag{20}$$

We cannot utilize the squared Euclidean distance between the data in the feature space for K-CM (Algorithm 1) because the optimization problems (5)–(10) include the squared Euclidean distance between the data and the cluster center in the feature space. Therefore, we reformulate the K-CM optimization problems (5)–(10) into the following problems:

$$\underset{u}{\text{minimize}} \sum_{j=1}^{C} \frac{\sum_{i=1}^{N}\sum_{k=1}^{N} u_{i,j}\|\Phi(x_i) - \Phi(x_k)\|_{\mathbb{H}}^2}{\left(\sum_{t=1}^{N} u_{t,j}\right)^2}, \tag{21}$$

$$\text{subject to} \sum_{j=1}^{C} u_{i,j} = 1, \quad u_{i,j} \in \{0, 1\}, \tag{22}$$

for K-HCM,

$$\underset{u}{\text{minimize}} \sum_{j=1}^{C} \frac{\sum_{i=1}^{N}\sum_{k=1}^{N} u_{i,j}^m\|\Phi(x_i) - \Phi(x_k)\|_{\mathbb{H}}^2}{\left(\sum_{t=1}^{N} u_{t,j}\right)^2}, \tag{23}$$

$$\text{subject to} \sum_{j=1}^{C} u_{i,j} = 1, \tag{24}$$

for K-sFCM, and

$$\underset{u}{\text{minimize}} \sum_{j=1}^{C} \frac{\sum_{i=1}^{N}\sum_{k=1}^{N} u_{i,j}\|\Phi(x_i) - \Phi(x_k)\|_{\mathbb{H}}^2}{\left(\sum_{t=1}^{N} u_{t,j}\right)^2} + \lambda^{-1}\sum_{i=1}^{N}\sum_{j=1}^{C} u_{i,j}\log(u_{i,j}), \tag{25}$$

$$\text{subject to} \sum_{j=1}^{C} u_{i,j} = 1. \tag{26}$$

for K-eFCM. The form $\|\Phi(x_i) - \Phi(x_k)\|_{\mathbb{H}}^2$ in the above objective functions can be translated into $-\tilde{K}_{i,k}$ if $\tilde{K}_{i,i} = 0$, and $-\tilde{K}_{i,k} + 1/2(\tilde{K}_{i,i} + \tilde{K}_{k,k})$ otherwise, with the cpd kernel matrix $\tilde{K}$. Such optimization problems are identical to those for RHCM and RFCM [9] if the corresponding relational data matrix is $\tilde{K}$, and they are solved using the following algorithms:

Algorithm 2 (cpdK-{H,sF,eF}CM):

STEP 1. Set the fuzzifier parameter m and λ for cpdK-sFCM and cpdK-eFCM, respectively. Set the cpd kernel matrix $\tilde{K}$ and the initial membership $u_{i,j}$. Set the N-dimensional squared matrix R as

$$R_{i,k} = -\tilde{K}_{i,k} + 1/2(\tilde{K}_{i,i} + \tilde{K}_{k,k}) \tag{27}$$

STEP 2. Update the intermediate variable $w_j \in \mathbb{R}^N$ as

$$w_j = \left(\sum_{i=1}^{N} u_{i,j}\right)^{-1} (u_{1,j}, \ldots, u_{N,j})^{\mathsf{T}}, \tag{28}$$

for cpdK-HCM,

$$w_j = \left(\sum_{i=1}^{N} u_{i,j}^m\right)^{-1} (u_{1,j}^m, \ldots, u_{N,j}^m)^{\mathsf{T}}, \tag{29}$$

for cpdK-sFCM, and

$$w_j = \left(\sum_{i=1}^{N} u_{i,j}\right)^{-1} (u_{1,j}, \ldots, u_{N,j})^{\mathsf{T}}. \tag{30}$$

for cpdK-eFCM.

STEP 3. Update the intermediate variable $\tilde{d}_{i,j}$ as

$$\tilde{d}_{i,j} = (Rw_j)_i - \frac{1}{2}w_j^{\mathsf{T}}Rw_j. \tag{31}$$

STEP 4. Update the membership $u_{i,j}$ as

$$u_{i,j} = \begin{cases} 1 & (i = \arg\min_k \tilde{d}_{k,j}), \\ 0 & (\text{otherwise}), \end{cases} \tag{32}$$

for cpdK-HCM,

$$u_{i,j} = 1/\sum_{k=1}^{C}\left(\frac{\tilde{d}_{i,j}}{\tilde{d}_{i,k}}\right)^{1/(m-1)}, \tag{33}$$

for cpdK-sFCM, and

$$u_{i,j} = \frac{\exp(-\lambda \tilde{d}_{i,j})}{\sum_{k=1}^{C} \exp(-\lambda \tilde{d}_{i,k})}, \tag{34}$$

for cpdK-eFCM.

STEP 5. If a stopping criterion is satisfied, terminate this algorithm. Otherwise, go to STEP. 2.

B. c-Means with pd kernel obtained from cpd kernel

In this subsection, we show how the cpd kernel is used in K-CM algorithms with a positive definite kernel obtained by arranging the cpd kernel. First, we show the relation between the cpd kernel and the pd kernel. Next, for K-CM (Algorithm 1) with a pd kernel obtained by reformulating the cpd kernel, we show how the cpd kernel is used to calculate the squared Euclidean distance between the data and the cluster center in the feature space.

The cpd kernel and the pd kernel have the following relation:

Theorem 2 (Connection cpd — pd): [8]

Let K be an N-dimensional symmetric matrix; $\mathbf{1} \in \mathbb{R}^N$, an N-dimensional real-valued vector whose elements are all 1; $E \in \mathbb{R}^{N \times N}$, the identity matrix; and $\eta \in \mathbb{R}^N$, a vector satisfying

$$\mathbf{1}^\mathsf{T} \eta = 1. \tag{35}$$

Then,

$$K = (E - \mathbf{1}\eta^\mathsf{T})\tilde{K}(E - \eta\mathbf{1}^\mathsf{T}) \tag{36}$$

is positive definite if and only if $\tilde{K}$ is conditionally positive definite.

Given the cpd kernel $\tilde{K}$ for a data set X, we can obtain the pd kernel

$$K = (E - \mathbf{1}\mathbf{1}^\mathsf{T}/N)\tilde{K}(E - \mathbf{1}\mathbf{1}^\mathsf{T}/N) \tag{37}$$

from Theorem 2 and use it K-CM (Algorithm 1). Then, STEP 3 of Algorithm 1 is rewritten as

STEP 3. Update the intermediate variable $\tilde{d}_{i,j}$ as

$$d_{i,j} = (e_i - W_j)^\mathsf{T} \tilde{K}(e_i - W_j) \tag{38}$$

from that

$$(E - \eta\mathbf{1}^\mathsf{T})(e_i - W_j) = (e_i - W_j). \tag{39}$$

C. Discussion

In this subsection, we show the relation between the K-CM algorithm with a pd kernel obtained from the cpd kernel and the cpdK-CM algorithm, and we geometrically interpret the explicit mapping from the data space to the feature space, hiding in cpdK-CM.

First, the updating equations (28) and (29) of the intermediate variable w_j in cpdK-CM (Algorithm 2) and the updating equations (11) and (12) of the cluster center W_j in K-CM (Algorithm 1) are identical. Next, the updating equations (32), (33), and (34) of the membership $u_{i,j}$ in cpdK-CM (Algorithm 2) and the updating equations (14), (15), and (16) of the membership $u_{i,j}$ in K-CM (Algorithm 1) are also identical. Furthermore, the updating equation of $d_{i,j}$ in K-CM (Algorithm 1) with a pd kernel obtained by reformulating the cpd kernel is expressed as Eq. (38). On the other hand, the updating equation of $\tilde{d}_{i,j}$ in cpdK-CM (Algorithm 2) can be expressed as

$$\begin{aligned} \tilde{d}_{i,j} =& (Rw_j)_i - \frac{1}{2} w_j^\mathsf{T} R w_j \\ =& \frac{1}{2}(e_i - w_j)^\mathsf{T} \tilde{K}(e_i - w_j) \end{aligned} \tag{40}$$

from $R = -\tilde{K}$, with $\tilde{K}_{i,i} = 0$. Therefore, $\tilde{d}_{i,j}$ in cpdK-CM (Algorithm 2) and $d_{i,j}$ in K-CM (Algorithm 1) with a pd kernel obtained by reformulating the cpd kernel have the following relation:

$$\tilde{d}_{i,j} = \frac{1}{2} d_{i,j}. \tag{41}$$

Comparing these relations with that between the pd kernel and the cpd kernel, cpdK-CM (Algorithm 2) for the cpd kernel $\tilde{K}$ with $\tilde{K}_{i,i} = 0$ and K-CM (Algorithm 1) for the pd kernel K with

$$K = 2(E - \mathbf{1}\eta^\mathsf{T})\tilde{K}(E - \eta\mathbf{1}^\mathsf{T}) \tag{42}$$

are identical.

Next, we geometrically interpret the explicit mapping from the data space to the feature space, hiding in cpdK-CM. Since cpdK-CM (Algorithm 2) and K-CM (Algorithm 1) with a pd kernel obtained from the cpd kernel are identical, we interpret cpdK-CM (Algorithm 2) from the geometrical interpretation of K-CM (Algorithm 1) with a pd kernel obtained from the cpd kernel.

Fig. 1 shows the feature space $\mathbb{H}$ used in K-CM (Algorithm 1). The mapping $e_i = \Phi(x_i)$ from each data x_i by an explicit mapping Φ is expressed along the i-th axis. From the updating equation of the cluster center W_j in K-CM (Algorithm 1), every cluster center W_j is in the $N-1$-dimensional subspace spanned by each data e_i. Hence $e_i - W_j$ for all $i \in \{1, \cdots, N\}$ and $j \in \{1, \cdots, C\}$ is orthogonal to $\mathbf{1}$. Therefore, it is guaranteed that $d_{i,j}$ is non-negative, even if $\tilde{K}$ is conditionally positive definite (and not positive definite) in K-CM (Algorithm 1).

IV. NUMERICAL EXAMPLE

In this section, we present an example of cpdK-CM. Fig. 2 shows 147 data points in two-dimensional space, which form doughnut-shaped and ball-shaped clusters. Such data cannot be clustered correctly using the usual HCM, sFCM, and eFCM algorithms because these clusters have a nonlinear border. Introducing the cpd kernel

$$\tilde{\mathcal{K}}(x, x') = \log(1 + \|x - x'\|_2^{0.5}), \tag{43}$$

cpdK-sFCM (Algorithm 2) with the fuzzifier parameter $m = 1.1$ yielded a successful result, as shown in Fig. 3.

V. CONCLUSION

In this paper, we investigated a hard and two fuzzy c-means clustering algorithms with a conditionally positive definite

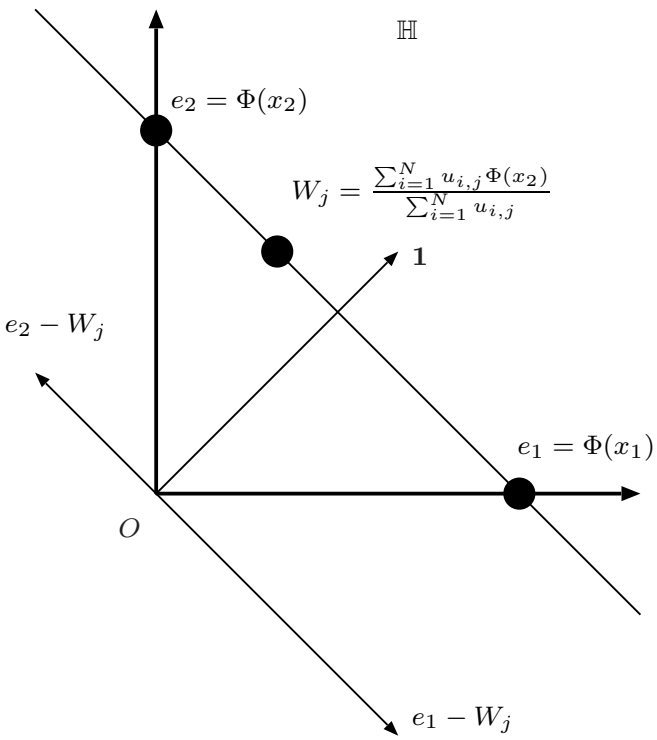


Fig. 1. Data and cluster centers in feature space $\mathbb{H}$

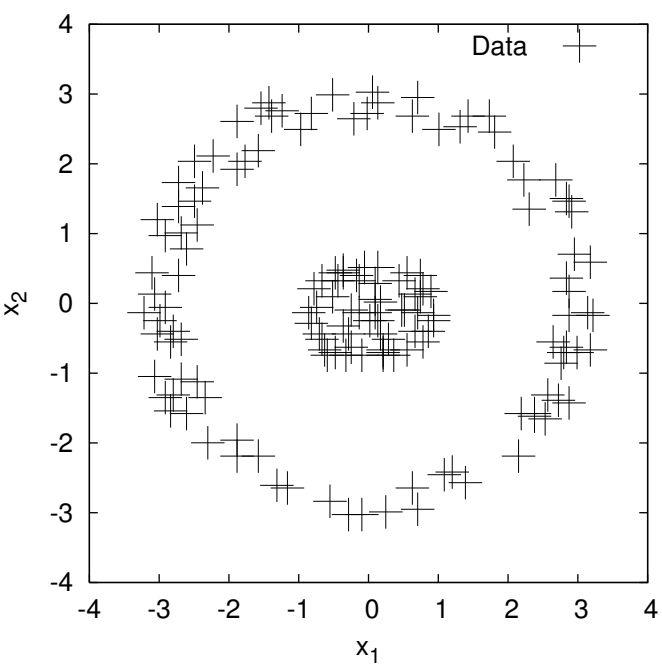


Fig. 2. Data

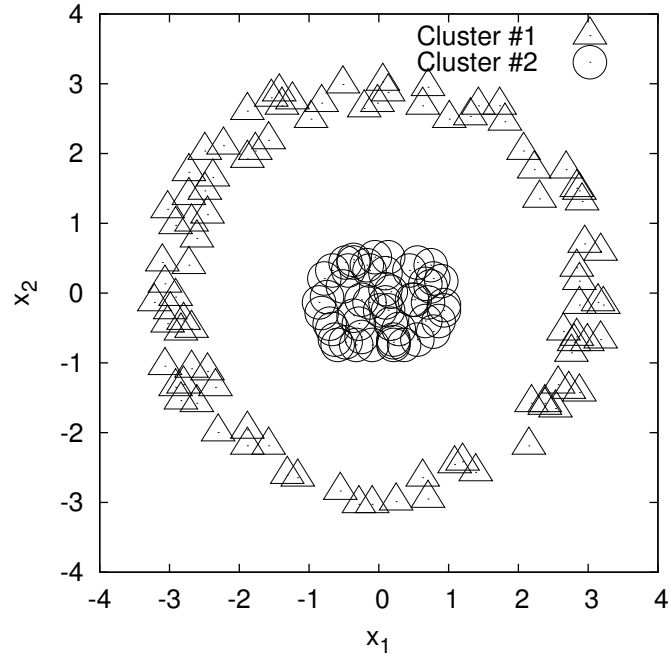


Fig. 3. Successful result ($\tilde{\mathcal{K}}(x, x') = \log(1 + \|x - x'\|_2^{0.5})$, $m = 1.1$

kernel. First, we developed c-means clustering algorithms using cpd kernel by solving the revised optimization problem of K-HCM, K-sFCM, and K-eFCM; thus, the cpd kernel is the squared Euclidean distance between the data in the feature space. Next, we showed how the cpd kernel is used in K-CM algorithms with a positive definite kernel obtained by arranging the cpd kernel. Furthermore, we showed the relation between the cpdK-CM algorithms and K-CM algorithms with a positive definite kernel obtained from the cpd kernel, and we interpreted a mapping from the original data space to the feature space, which is latent in cpdK-CM algorithms. Finally, we showed a numerical example of the cpdK-CM algorithms.

REFERENCES

[1] Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms, Plenun, New York, (1981).

[2] Miyamoto, S. and Umayahara, K.: "Methods in Hard and Fuzzy Clustering", in: Liu, Z.-Q. and Miyamoto, S. (eds), Soft computing and human-centered machines, Springer-Verlag Tokyo, (2000).

[3] Miyamoto, S. and Nakayama, Y.: "Algorithms of Hard c-Means Clustering Using Kernel Functions in Support Vector Machines", J. Advanced Computational Intelligence and Intelligent Informatics, Vol.7, No.1, pp.19–24, 2003.

[4] Miyamoto, S. and Suizu, D.: "Fuzzy c-Means Clustering Using Kernel Functions in Support Vector Machines", J. Advanced Computational Intelligence and Intelligent Informatics, Vol.7, No.1, pp.25–30, 2003.

[5] Vapnik, V., N.: Statistical Learning Theory, Wiley, New York, 1998.

[6] Miyamoto, S., Kawasaki, Y. and Sawazaki, K.: "An Explicit Mapping for Kernel Data Analysis and Application to Text Analysis", Proc. IFSA-EUSFLAT 2009, pp.618–623, 2009.

[7] Miyamoto, S. and Sawazaki, K.: "An Explicit Mapping for Kernel Data Analysis and Application to c-Means Clustering", Proc. NOLTA 2009, pp.556–559, 2009.

[8] Schoölkopf, B: "The Kernel Trick for Distances", In Advances in Neural Information Processing System 13, MIT Press, 2001.

[9] Hathaway, R.J., Davenport, J.W. and Bezdek, J.C.: "Relational Duals of the c-means Clustering Algorithms", Pattern Recognition, Vol.22, No.2, pp.205–212, 1989.