Algoritma Klasifikasi

Nama : M. Iqbal Rivana

No Mahasiswa : 192420057

Mata Kuliah : Advanced Database

Soal

Algoritma Klasifikasi

Jawaban

Sistem kerja metode tidak terbimbing (Unsupervised) adalah melakukan pengelompokan nilai-nilai pixel suatu citra oleh komputer kedalam kelas-kelas spektral dengan menggunakan algoritma klusterisasi. Dalam metode ini, diawal proses biasanya analis (orang yang melakukan analisis) akan menentukan jumlah kelas (cluster) yang akan dibuat. Kemudian setelah mendapatkan hasil, analis menetapkan kelas-kelas lahan terdapat kelas-kelas spektral yang telah dikelompokan oleh komputer. Dari kelas yang dihasilkan, analis bisa menggabungkan beberapa kelas yang dianggap memiliki informasi yang sama menjadi satu kelas. Misal class 1, class 2, dan class 3 misal nya masing-masing adalah sawah, perkebunan dan hutan maka bisa dikelompokkan menjadi satu kelas yaitu kelas vegetasi.

Metode tidak terbimbing terdiri dari dua jenis yaitu :

1. **IsoData** = Mengklasifikasikan kelas secara merata, setiap pixel diklasifikasikan ke kelas terdekat. Setiap interaksi akan dikalkulasi ulang dan mereklasifikasi pixel ke bentuk baru. Memisah kelas, menggabungkan dan menghapus dilakukan berdasarkan parameter input. Semua pixel diklasifikasikan ke kelas terdekat kecuali deviasi standar atau ambang batas jarak yang telah ditentukan, dalam hal ini beberapa pixel mungkin tidak diklasifikasikan jika tidak memenuhi kriteria yang ditentukan. Proses ini berlanjut sampai jumlah pixel dalam setiap perubahan kelas kurang dari ambang perubahan pixel yang dipilih atau jumlah maksimum interasi tercapai.
2. **K-Means** = Hampir sama dengan metode IsoData, bedanya dengan menggunakan metode ini analis mengharuskan untuk memilih jumlah kelas yang berlokasi di data, kemudian sistem akan mengelompokkan data ke dalam kelas kelompok yang telah ditentukan. Pada setiap kelas akan terdapat titik tengah (centroid) yang mempresentasikan kelas tersebut.

No.

No.

Date . .

ADVANCED DATABASE (MTIK 112)

Nama : Rahmi

NM : 1924 - 20046

Tugas 2

Algoritma klasifikasi !

Jawab

Algoritma Naive - Bayes.

Naive - Bayes classifier merupakan sebuah metoda klasifikasi yang berakar pada teorema Bayes. Metode pengklasifikasian dengan menggunakan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai Teorema Bayes.

Keuntungan penggunaan adalah bahwa metoda ini hanya membutuhkan jumlah data pelatihan (training data) yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian.

Scanned with CamScanner

Nama : Rani Okta Felani

Nim : 192420048

Mata Kuliah : Advanced Database

Dosen Pengampu : M. Akbar,S.Ti, M.Ti

Tugas Topik 3

Algoritma Klasifikasi

Dalam Klasifikasi mesin belajar banyak sekali algoritma yang digunakan, coba cari Algoritma terbaik dan Tuliskan Argumennya di Tautan Tugas !

Jawab :

Menurut saya Algoritma terbaik dalam klasifikasi mesin belajar (machine learning) untuk tehnik Unsupervised learning adalah Algoritma **K-Means**

Algoritma K-Means seringkali digunakan karena mudah di aplikasikan dan di mengerti.

Klasifikasi Machine Learning dalam Algoritma K-Means

K-Means Clustering adalah salah satu *unsupervised machine learning* algoritma yang paling sederhana dan populer. Tujuan dari algoritma ini adalah untuk menemukan grub dalam data, dengan jumlah grub yang diwakili oleh variabel variabel K sendiri adalah jumlah cluster yang kita inginkan.

Proses K-Means *cluestering* itu sendiri adalah untuk memperoses data algoritma K-Means *clustering* , data di mulai dengan kelompok pertama *centroid* yang di pilih secara acak, yang digunakan sebagai titik awal untuk setiap *cluster* dan kemudian melakukan perhitungan berulang untuk mengoptimalkan posisi *centroid.*

Proses ini berhenti atau tidak selesai dalam mengoptimalkan *cluster* ketika :

1. C*entroid* telah stabil tidak ada perubahan dalam nilai-nilai mereka karena pengelompokkan sudah berhasil.

2. jumlah iterasi yang di tentukan telah tercapai.

Hasil dari K-Means *Clustering* adalah

1. *Centroid* dari *cluster* K , yang dapat digunakan untuk memberi label data baru.

- Label untuk data pelatihan (setiap titik data di tugaskan ke satu *Cluster)*

Algoritma terbaik menurut saya adalah Dynamic Programming

Dynamic programming problems adalah masalah multi tahap(multistage) dimana keputusan dibuat secara berurutan (in sequence).

Pemrograman dinamis (dynamic programming) adalah metode pemecahan masalah dengan cara menguraikan solusi menjadi sekumpulan langkah (step) atau tahapan (stage) sedemikian rupa sehingga solusi dari permasalahan ini dapat dipandang dari serangkaian keputusan-keputusan kecil yang saling berkaitan satu dengan yang lain. Penyelesaian persoalan dengan pemrograman dinamis ini akan menghasilkan sejumlah berhingga pilihan yang mungkin dipilih, lalu solusi pada setiap tahap-tahap yang dibangun dari solusi pada tahap sebelumnya, dan dengan metode ini kita menggunakan persyaratan optimasi dan kendala untuk membatasi sejumlah pilihan yang harus dipertimbangkan pada suatu tahap.

Karakteristik Persoalan yang dimiliki oleh Program Dinamis:

a. Persoalan dapat dibagi menjadi beberapa tahap (stage), yang pada setiap tahap hanya dapat diambil satu keputusan.
b. Masing-masing tahap terdiri dari sejumlah status (state) yang berhubungan dengan tahap tersebut. Secara umum, status merupakan bermacam kemungkinan masukan yang ada pada tahap tersebut. Jumlahnya bisa berhingga atau tak berhingga.
c. Hasil dari keputusan yang diambil pada setiap tahap ditransformasikan dari status yang bersangkutan ke status berikutnya pada tahap berikutnya.
d. Ongkos (cost) pada suatu tahap meningkat secara teratur (steadily) dengan bertambahnya jumlah tahapan.
e. Ongkos pada suatu tahap bergantung pada ongkos tahap-tahap yang sudah berjalan dan ongkos pada tahap tersebut.
f. Keputusan terbaik pada suatu tahap bersifat independen terhadap keputusan yang dilakukan pada tahap sebelumnya.
g. Adanya hubungan rekursif yang mengidentifikasikan keputusan terbaik untuk setiap status pada tahap k memberikan keputusan terbaik untuk setiap status pada tahap k + 1.
h. Prinsip optimalitas berlaku pada persoalan tersebut.

Kelebihan dan kekurangan System Dynamic Programming

Kelebihannya :

a. Mengoptimalkan penyelesaian suatu masalah tertentu yang diuraikan menjadi sub-submasalah yang lebih kecil yang terkait satu sama lain dengan tetap memperhatikan kondisi dan batasan permasalahan tersebut.
b. Proses pemecahan suatu masalah yang kompleks menjadi sub-sub masalah yang lebih kecil membuat sumber permasalahan dalam rangkaian proses masalah tersebut menjadi lebih jelas untuk diketahui.
c. Pendekatan dynamic programming dapat diaplikasikan untuk berbagai macam masalah pemrograman matematik, karena dynamic programming cenderung lebih fleksibel daripada teknik optimasi lain.
d. Prosedur perhitungan dynamic programming juga memperkenankan bentuk analisissensitivitas terdapat pada setiap variabel status (state) maupun pada variabel yang ada di masing-masing tahap keputusan (stage).
e. Dynamic programming dapat menyesuaikan sistematika perhitungannya menurut ukuran masalah yang tidak selalu tetap dengan tetap melakukan perhitungan satu persatu secara lengkap dan menyeluruh.

Kelemahannya yaitu :

Penggunaan dynamic programming jika tidak dilakukan secara tepat, akan mengakibatkan ketidakefisienan biaya maupun waktu. Karena dalam menggunakan dynamic programming diperlukan keahlian, pengetahuan, dan seni untuk merumuskansuatu masalah yang kompleks, terutama yang berkaitan dengan penetapan fungsi transformasi dari permasalahan tersebut.

Program Pascasarjana Magister Teknik informatika

Universitas Bina Darma Palembang

Nama : Suwani

Nim : 192420049

Mata Kuliah :Advanced Database

Soal :

Dalam Klasifikasi Machine learning ada banyak sekali algoritma yang dapat di gunakan, coba cari algoritma terbaik untuk di tuliskan algoritmanya.

Jawab :

Adaboost

Model standard dari algoritma adaboost terdiri dari dua bagian, yaitu bagian *offline training* dan bagian *online recognizing*. Bagian *offline training* adalah bagian proses pelatihan data yang tidak bekerja secara *realtime*. Bagian ini meliputi penginputan sampel gambar positif dan sampel gambar negatif, *preprocessing,* pelatihan data oleh algoritma adaboost sampai membangun detektor. Setelah detektor terbentuk kita bisa melakukan pendeteksian secara *realtime/online recognizing* terhadap data pengujian. Sebelum melakukan pendeteksian dengan algoritma adaboost, terlebih dahulu data pengujian sudah harus mengalami *preprocessing.*

Metode AdaBoost merupakan salah satu algoritma *supervised* pada data mining yang diterapakan secara luas untuk membuat model klasifikasi. AdaBoost sendiri pertama kali diperkenalkan oleh Yoav Freund dan Robert Schapire(1995). Walaupun pada awalnya algoritma ini diterapkan pada model regresi, seiring dengan perkembangan teknologi komputer yang cepat, metode ini juga dapat diterapkan pada model statistik lainnnya. Metode adaBoost merupakan salah satu teknik *ensamble* dengan menggunakan *loss function fungsi exponential* untuk memperbaiki tingkat akurasi dari prediksi yang dibuat.

Nama : Theo Vhaldino
Nim : 192420058
Angkatan/Reguler : 22 / A R1
Mata Kuliah : *Advanced Database* (MTIK112)

Tugas Sesi 3

Dalam Klasifikasi *machine learning* terdapat banyak algoritma yang dapat digunakan. Secara Universal, algoritma tertentu tidak bisa dikatakan algoritma terbaik karena setiap algoritma memiliki kelebihan dan kekurangan. Tetapi diantara algoritma untuk klasifikasi saya memilih algoritma *Naive Bayes*.

Mengapa saya memilih *Naive Bayes* ? karena algoritma ini merupakan salah satu metode pembelajaran paling populer yang dikelompokkan berdasarkan kesamaan, yang bekerja pada *Bayes Theorem of Probability* - untuk membangun model pembelajaran mesin. Teknik klasifikasi berdasarkan teorema Bayes sangat mudah dibangun dan sangat berguna untuk set data yang sangat besar. Seiring dengan kesederhanaan, *Naive Bayes* dikenal mengungguli metode klasifikasi yang sangat canggih. *Naive Bayes* juga merupakan pilihan yang baik ketika CPU dan sumber daya memori merupakan faktor pembatas.

Naive Bayes sangat sederhana, Anda hanya melakukan banyak perhitungan. Jika asumsi independensi bersyarat *Naive Bayes* benar-benar berlaku, klasifikasi *Naive Bayes* akan menyatu lebih cepat daripada model diskriminatif seperti regresi logistik, sehingga Anda membutuhkan lebih sedikit data pelatihan. Dan bahkan jika asumsi *Naive Bayes* tidak berlaku, pengklasifikasi *Naive Bayes* masih sering melakukan pekerjaan yang baik dalam praktik. Taruhan yang bagus jika menginginkan sesuatu yang cepat dan mudah yang berkinerja cukup baik. Kerugian utamanya adalah tidak bisa mempelajari interaksi antar fitur.

Kapan menggunakan algoritma *Naive Bayes* dipergunakan ?

1. Jika memiliki kumpulan data pelatihan sedang atau besar.
2. Jika instance memiliki beberapa atribut.
3. Mengingat parameter klasifikasi, atribut yang menggambarkan instance harus independen secara kondisional.

Naive Bayes dapat digunakan dalam aplikasi dunia nyata seperti :

1. Analisis sentimen dan klasifikasi teks
2. Sistem rekomendasi seperti Netflix, Amazon
3. Untuk menandai email sebagai spam atau bukan spam
4. Pengenalan wajah

Nama : Arpa Pauziah

Ruangan : U 705

Mata Kuliah : Advanced Database

Pertanyaan : Dalam klasifikasi mesin learning banyak sekali algoritma yang dapat digunakan, coba cari algoritma terbaik dan tuliskan argumen nya di link tugas?

Pembahasan:

Algoritma klasifikasi pada machine learning yang terawasi atau supervised learning salah satu nya adalah algoritma K-Nearest Neighbor (K-NN) adalah pendekatan untuk mencari kasus dengan menghitung kedekatan antara kasus baru dengan kasus lama, yaitu berdasarkan pada pencocokan bobot dari sejumlah fitur yang ada. Kedekatan biasanya berada pada nilai antara 0 s/d 1. Nilai 0 artinya kedua kasus mutlak tidak mirip, sebaliknya untuk nilai 1 kasus mirip dengan mutlak.

$$similarity(T,S) = \frac{\sum_{i=1}^{n} f(T_i,S_i)\ x\ w_i}{w_i}$$

dengan
T : kasus baru
S : kasus yagn ada dalam penyimpanan
n : jumlah atribut dalam masing-masing kasus
i : atribut individu antara 1 s/d n
f : fungsi similarity atribut i antara kasus T dan kasus S
w : bobot yang diberikan pada atribut ke i

Contoh kasus sederhana pada perusahaan Finance yang memiliki kriteria untuk memberikan pinjaman dana kepada nasabahnya, kriteria tersebut adalah sebagai berikut :

Pekerjaan/Usaha	Jumlah Tanggungan	Kepemilikan Rumah	Pinjaman
PNS	5	Milik sendiri	Ya
Pedagang	3	Sewa	Tidak
Pengusaha	5	Milik sendiri	Ya
PNS	5	Sewa	Tidak

Apabila terdapat data baru, yaitu sebagai berikut :

Pekerjaan/Usaha	Jumlah Tanggungan	Kepemilikan Rumah	Pinjaman
Pengusaha	3	Sewa	?
Pedagang	5	Milik Sendiri	?

Langkah pertama yang dilakukan pada algoritma knn adalah menentukan bobot dan kedekatan tiap kriteria.

Bobot antar variabel

Kriteria	Bobot
Pekerjaan/Usaha	0,75
Jumlah Tanggungan	0,5
Kepemilikan Rumah	1

Kedekatan Pekerjaan/Usaha

Nilai1	Nilai2	Kedekatan
PNS	PNS	1
PNS	Pedagang	0,5
PNS	Pengusaha	0,5
Pedagang	Pedagang	1
Pedagang	Pengusaha	0,5
Pedagang	PNS	0,5
Pengusaha	PNS	0,5
Pengusaha	Pedagang	0,5
Pengusaha	Pengusaha	1

Kedekatan Jumlah Tanggungan

Nilai1	Nilai2	Kedekatan
3	3	1
3	5	0,4
5	5	1
5	3	0,4

Kedekatan Kepemilikan Rumah

Nilai1	Nilai2	Kedekatan
Milik sendiri	Milik sendiri	1
Milik sendiri	Sewa	0,75
Sewa	Sewa	1
Sewa	Milik sendiri	0,75

Langkah selanjutnya membandingkan data baru dengan data yang sudah ada dan memperoleh nilai kedekatannya

Pekerjaan /Usaha	Jumlah Tanggungan	Kepemilikan Rumah	Pinjaman	Kedekatan Pekerjaan/Usaha	Kedekatan Jumlah Tanggungan	Kedekatan Kepemilikan rumah	Similararity
PNS	5	Milik sendiri	Ya	0.5	0.4	0.75	0.588888889
Pedagang	3	Sewa	Tidak	0.5	1	1	0.833333333
Pengusaha	5	Milik sendiri	Ya	1	0.4	0.75	0.755555556
PNS	5	Sewa	Tidak	1	1	0.75	0.888888889
Pengusaha	3	Sewa	?	Tidak			

Pekerjaan /Usaha	Jumlah Tanggungan	Kepemilikan Rumah	Pinjaman	Kedekatan Pekerjaan/Usaha	Kedekatan Jumlah Tanggungan	Kedekatan Kepemilikan rumah	Similararit
PNS	5	Milik sendiri	Ya	0.5	1	1	0.833333
Pedagang	3	Sewa	Tidak	1	0.4	0.75	0.755556
Pengusaha	5	Milik sendiri	Ya	0.5	1	1	0.833333
PNS	5	Sewa	Tidak	0.5	1	0.75	0.722222
Pedagang	5	Milik Sendiri	?	Ya			

Nilai kedekatan yang diperoleh pada tabel tersebut adalah dengan membandingkan data lama dengan data baru, contohnya untuk kriteria pekerjaan/usaha data baru adalah pengusaha sedangkan data lama pertama adalah PNS, maka lihat pada tabel kedekatan pekerjaan/usaha berapa kedekatan untuk nilai1 pengusaha dan nilai2 PNS sehingga diperoleh 0,5 begitu seterusnya. Sedangkan untuk nilai similaraty diperoleh berdasarkan rumus similarity pada penjelasan diawal.

Jadi, untuk data baru tersebut, maka diambil keputusan sebagai berikut :

Pekerjaan/Usaha	Jumlah Tanggungan	Kepemilikan Rumah	Pinjaman
Pengusaha	3	Sewa	Tidak
Pedagang	5	Milik Sendiri	Ya

Nama : Elpina Sari

Nim : 192420050

Tugas 3 (E-L)

Dalam klasifikasi mesin learning banyak sekali algoritma yang dapat digunakan, coba cari algoritma terbaik dan tuliskan argumen nya di link tugas

Jawaban :

Algoritma C4.5 / Decision Tree Algoritma C4.5 merupakan salah satu teknik klasifikasi pada machine learning yang digunakan pada proses data mining dengan membentuk sebuah pohon keputusan (decision tree) yang direpresentasikan dalam bentuk aturan. Algoritma C4.5 merupakan kelompok algoritma dengan menggunakan pohon keputusan. Pohon keputusan merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal. Semakin kaya informasi atau pengetahuan yang dikandung oleh data training, maka akurasi akan semakin meningkat. Pohon dalam analisis pemecahan masalah pengambilan keputusan adalah pemetaan mengenai alternatif-alternatif pemecahan masalah yang dapat diambil dari masalah tersebut. Pohon tersebut juga memperlihatkan faktor-faktor kemungkinan/probabilitas yang akan mempengaruhi alternatif-alternatif keputusan tersebut, disertai dengan estimasi hasil akhir yang akan didapat bila kita mengambil alternatif keputusan tersebut. Pengambilan keputusan merupakan masalah penting bagi organisasi untuk menemukan alternatif terbaik dari alternatif yang ada.

Algoritma C4.5 adalah algoritma yang dibuat oleh Ross Quinlan dan digunakan untuk membuat pohon keputusan. Algoritma ini sering dikategorikan sebagai pengklasifikasi statistik. Algoritma C4.5 merupakan pengembangan dari algoritma ID3 yang menggunakan entropi informasi, atribut continu dan diskret, atribut kategorial dan numerik, dan missing values. Algoritma ini membutuhkan set data latih karena termasuk algoritma pembelajaran yang terawasi (supervised learning algorithm). Set data latih berupa sampel yang sudah terklasifikasi. Algoritma C4.5 menganalisa set data latih dan membangun pengklasifikasian yang harus secara tepat mampu mengklasifikasikan data latih maupun data uji. Formulasi Matematis dari algoritma C 4.5 dapat dilihat pada formula (1)-(5).

$$Entropy(S) = \sum_{i=1}^{n} - pi * \log_2 pi \quad (1)$$

dengan penjelasan $\log_2$:

$$\log_2(x) = \frac{\ln(x)}{\ln(2)} \quad (2)$$

dengan:

S = ruang (data) sampel yang digunakan untuk training.

pi = proporsi dari S_i terhadap S. S_i diperoleh dari *Gain*.

$$Gain(S, A) = Entropy(S) - \sum_{i=1} \frac{|S_i|}{|S|} * Entropy(S_i) \quad (3)$$

dengan :

S = ruang (data) sampel yang digunakan untuk training.

A = atribut. V = suatu nilai yang mungkin untuk atribut A.

Nilai (A) = himpunan yang mungkin untuk atribut A.

|Si| = jumlah sampel untuk nilai i.

|S| = jumlah seluruh sampel data.

Entropy(Si) = entropi untuk sampel-sampel yang memiliki nilai i.

$$SplitInfo(S, A) = -\sum_{i=1} \frac{S_i}{S} \log_2 \frac{S_i}{S} \quad (4)$$

dengan:

S = ruang (data) sampel yang digunakan untuk training.

A = atribut.

Si = jumlah sampel untuk atribut i

$$GainRatio(S, A) = \frac{Gain(S, A)}{SplitInfo(S, A)} \quad (5)$$

dengan:

S = ruang (data) sampel yang digunakan untuk training.

A = atribut.

Gain (S,A) = information gain pada atribut A.

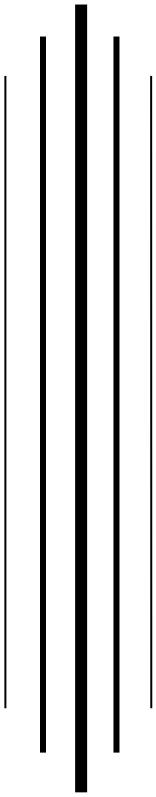
SplitInfo (S,A) = split information pada atribut A.

Secara umum langkah dalam membuat pohon keputusan menggunakan algoritma C4.5 adalah sebagai berikut:

1. Menghitung entropi total dari dataset dilanjutkan dengan entropi masing-masing atribut.
2. Setelah diperoleh entropi masing-masing atribut, menghitung information gain masing - masing.
3. Memilih atribut yang memiliki information gain paling besar sebagai akar.
4. Mengulangi perhitungan entropi dan gain untuk menentukan atribut berikutnya sebagai daun.

TUGAS ADVANCE DATABASE

Perbandingan Algoritma Klasifikasi



DISUSUN OLEH:

FADEL MUHAMMAD MADJID

192420052

PROGRAM PASCA SARJANA MAGISTER TEKNIK INFORMATIKA

UNIVERSITAS BINA DARMA

Perbandingan Algoritma Klasifikasi

Kelebihan dan kekurangan masing-masing algoritma

Logistic regresion (LR)

+

output memiliki intepretasi probabilistik bagus, algoritma dapat diatur untuk menghindari overfitting. model dapat diupdate dengan mudah dengan data baru menggunakan stochastic gradien descent.

-

LR cenderung buruk performanya ketika ada data multiple atau non-linear

Decission tree (DT)

+

performanya bagus saat diberi pelatihan. algortima kuat untuk outliers, scalable dan dapat digunakan untuk model non linear pada stuktur hierarkinya

-

cenderung overfitting jika cuma individual tree

Support Vector Machine (SVM)

+

dapat digunakan dalam model non-linear, dan ada banyak kernel yang dapat dipilih. algoritma jarang overfittig, terutama dalam high dimensional space

-

memakai memori yang banyak, harus cermat memilih kernel karena sangat penting, tidak dapat diskala dalam dataset yang lebih besar

Naive Bayes

+

sangat simpel, dapat diskala dan mudah diterapkan

-

kurang kompleks dan dapat dikalahkan algo lain yg lebih terlatih

K-nearest Neeighbour (KNN)

+

mudah dan simpel model

sedikit hyperparameter untuk di tuning

-

k harus dipilih secara teliti

beban komputasi beaar untuk running dalam ukuran data besar

scaling dibutuhkan untuk mengatasi semua variabel

Neural Network (NN)

+

adanya toleransi kesalahan untuk missing information

memori yang terdistribusi

-

susah untuk menunjukkan masalah dalam network

tidak ada aturan khusus untuk network karena hanya dilakukan dengan trial dan error

Perbandingan antara 2 algoritma

LR vs SVM

SVM dapat mengatasi model non-linear, LR tidak

SVM linear mengatasi outliers lebih bagus, karena dapat diperoleh margin maksimum

error SVM lebih besar ketimbang LR

LR vs DT

DT mengatasi lineaitas lebih baik

DT tidak memngartikan pentingnya variabel

DT lebih bagus untuk kategori

LR vs NN

NN dapat mengatasi solusi nonlinear

LR memiliki fungsi error yang dapat membesar, N tergantung

LR dapat melatih data kecil, NN harus banyak

LR vs NB

NB merupakan model generatif, LR merupakan model diskriminatif

NB bekerja bagus dalam dataset kecil

LR vs KKN

KNN merupakan model non paramteris, LR merupakan model paramteris

KNN lebih lambat ketimbang LR

KNN dapat mengatasi model nonlinear

KNN dapat untuk prediksi (level kepercayaan), KNN hanya untuk iutput yang berlabel

KNN vs NB

NB lebih cepat

NB merupakan parametris, KNN non paramteris

KNN vs LR

KNN lebi bagus untuk mengatasi regresi linear dengan data yang memiliki SNR tinggi

KNN vs SVM

SVM mengatasi outlier lebih bagus

Jika training data memiliki junmlah variabel tinggi, KNN lebih bagus. SVM lebih ungul jika variabel besar dan training data lebih sedikit

KNN vs NN

NN butuh training data yang besar, untuk mempertinggi akurasi

NN butuh tuning hyperparameter yang lebih banyak ketimbang KNN

DT vs KNN

sama-sama non parametris

DT menyediakan interaksi variabel yang otomatis

DT lebih cepat ketimbang KNN

DT vs NB

DT merupakan model diskriminatif, NB merupakan model generatif

DT lebih fleksibel dan mudah

dalam pemangkasan DT menyampingkan nilai penting dalam training data, dapat mempertinggi error

DT vs NN

sama-sama non-linear, dan memiliki interaksi antar variabel input

DT lebih bagus dalam set nilai kategori yang tinggi dalam training

DT lebih bagus jika skenario membutuhkan penjelasan dalam tujuan training

NN unggul dalam training data yang cukup

DT vs SVM

SVM menggunakan kernel untuk menyelesaikan masalah non-linear, dimana DT membutuhkan pemangkasn dalam input untuk menyelesaikan masalah

DT bagus untuk data kategori dan linearitas

SVM vs NB

sama-sama memiliki performa bagus dalam jumlah data sedikit dan banyak variabel

SVM adalah model diskriminatif dan NB adalah generatif

SVM vs NN

SVM memiliki optimization yang besar, NN bergantung pada local minima

SVM lebih bagus jika data training sedikit dan banyak variabel. NN butuh banyak data training untuk akurasi yang lebih bagus

Klasifikasi kelas banyak membutuhkan model yang banyak untuk SVM, sementara NN hanya dalam 1 model

Kesimpulan

Dalam perbandingan antara algoritma diatas dapat disimpulkan bahwa masing-masing algoritma memiliki kelebihan dan kekurangan dalam menyelesaikan masalah, kualitas dan kuantitas dari training, jumlah dan jenis variabel serta jenis dataset yang digunakan. Oleh karena itu, berlaku teori no free lunch yang berarti sebuah algoritma tidak dapat menyelesaikan semua masalah yang diberikan. Pemilihan algoritma yang sesuai akan mempermudah pekerjaan dan mendapatkan solusi yang akan digunakan dalam memecahkan masalah.

Daftar Pustaka

https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222

https://medium.com/@dannymvarghese/comparative-study-on-classic-machine-learning-algorithms-part-2-5ab58b683ec0

Nama : Isti Maátun Nasichah

NPM : 192420051

Program : Magister Teknik Informatika

TUGAS 2

Dalam klasifikasi mesin learning banyak sekali algoritma yang dapat digunakan, coba cari algoritma terbaik dan tuliskan argumennya!

Jawab :

Menurut saya algoritma klasifikasi terbaik adalah ANN (Artificial Neural Network), dimana ANN mempunyai kelebihan dalam hal kemampuan untuk generalisasi, yang bergantung pada seberapa baik ANN meminimalkan resiko empiris. Selain itu ANN mampu melakukan ekstraksi dari suatu pola data tertentu ANN dapat menciptakan suatu pola pengetahuan melalui pengaturan diri atau kemampuan belajar (self organizing).

ANN (Artificial Neural Network) adalah suatu sistem pemrosesan informasi yang mencoba meniru kinerja otak manusia. ANN merupakan generalisasi model matematis dengan asumsi :

a. Pemrosesan informasi terjadi pada elemen sederhana (=neuron).
b. Sinyal dikirimkan diantara neuron-neuron melalui penghubung (=dendrit dan akson).
c. Penghubung antar elemen memiliki bobot yang akan menambah atau mengurangi sinyal.
d. Untuk menentukan output, setiap neuron memiliki fungsi aktivasi (biasanya non linier) yang dikenakan pada semua input.
e. Besar output akan dibandingkan dengan threshold

Baik tidaknya suatu model ANN ditentukan oleh:

a. Pola antar neuron (arsitekur jaringan).
b. Metode untuk menentukan dan mengubah bobot (disebut metode learning).
c. Fungsi aktivasi ANN disebut juga: brain metaphor, computational neuroscience, parallel distributed processing

Kelebihan algoritma ANN :

a. Mampu mengakuisisi pengetahuan walau tidak ada kepastian.
b. Mampu melakukan generalisasi dan ekstraksi dari suatu pola data tertentu ANN dapat menciptakan suatu pola pengetahuan melalui pengaturan diri atau kemampuan belajar (self organizing).
c. Memiliki fault tolerance, gangguan dapat dianggap sebagai noise saja.
d. Kemampuan perhitungan secara paralel sehingga proses lebih singkat

Selain itu ANN mampu untuk :

a. Klasifikasi: memilih suatu input data ke dalam kategori tertentu yang sudah ditetapkan.
b. Asosiasi: menggambarkan suatu obyek secara keseluruhan hanya dengan bagian dari obyek lain.
c. Self organizing: kemampuan mengolah data-data input tanpa harus mempunyai target.
d. Optimasi: menemukan suatu jawaban terbaik sehingga mampu meminimalisasi fungsi biaya