Jurnal Inovasi Teknik dan Edukasi Teknologi, 2(12), 2022, 550-556
ISSN: 2797-7196 (online)
DOI:10.17977/um068v2i122022pxxx-xxx

Algoritma – Algoritma Data Mining untuk Klasifikasi Data

Nur A'yuni Ramadhani, Harits Ar Rosyid*
Universitas Negeri Malang, Jl. Semarang No. 5 Malang, Jawa Timur, Indonesia
*Penulis korespondensi, Surel: harits.ar.ft@um.ac.id

Paper received: 06-12-2022; revised: 15-12-2022; accepted: 29-12-2022

Abstract

Classification is a data processing technique by grouping the data according to the criteria possessed by each data. In data processing, the data to be processed does not have certain requirements. In short, all data can be classified. The data classification technique can use various algorithms. There are many algorithms that can be used in classification, such as Decision Tree, Support Vector Machine, Neural Network, and K-Nearest Neighbor. This study will review classification algorithms. Researchers will show the differences of each algorithm by showing the advantages and disadvantages of the algorithm. This study will also demonstrate a Case Based Reasoning system that can improve the results of the classification algorithm.

Keywords: datamining; classification; support vector machines; neural networks; decision trees; k-nearest neighbor; case based reasoning

Abstrak

Klasifikasi merupakan teknik pengolahan data dengan cara mengelompokkan data-data tersebut sesuai dengan kriteria yang dimiliki oleh masing-masing data. Dalam pengolahan datanya, data yang akan diolah tidak memiliki persyaratan tertentu. Singkatnya semua data dapat diklasifikasikan. Teknik klasifikasi data tersebut dapat menggunakan berbagai macam algoritma. Terdapat banyak algoritma yang dapat digunakan dalam klasifikasi, seperti Decision Tree, Support Vector Machine, Neural Network, dan K-Nearest Neighbor. Penelitian ini akan mengulas tentang algoritma – algoritma klasifikasi tersebut. Peneliti akan menunjukkan perbedaan dari masing – masing algoritma dengan menunjukkan kelebihan dan kekurangan algoritma tersebut. Pada penelitian ini juga akan ditunjukkan sistem Case Based Reasoning yang dapat memperbaiki hasil dari algoritma klasifikasi.

Kata kunci: data mining; klasifikasi; support vector machine; neural network; decision tree; k-nearest neighbor; case based reasoning

1. Pendahuluan

Data Mining adalah satu proses logis yang digunakan untuk menemukan data yang diharapkan dari keseluruhan data yang berjumlah besar(Neelamegam & Ramaraj, 2013). Salah satu metode yang dimiliki oleh data mining adalah klasifikasi. Klasifikasi merupakan teknik supervised dalam data mining. Klasifikasi sendiri adalah cara pengelompokkan benda berdasarkan ciri – ciri yang dimiliki oleh objek yang akan diklasifikasikan. Dalam prosesnya, klasifikasi dapat dilakukan dengan banyak cara baik secara manual ataupun dengan bantuan teknologi. Klasifikasi yang dilakukan secara manual adalah klasifikasi yang dilakukan oleh manusia tanpa adanya bantuan. Sedangkan klasifikasi yang dilakukan dengan bantuan teknologi, memiliki beberapa algoritma yang dapat digunakan seperti Decision Tree, Support Vector Machine(SVM), K-Nearest Neighbor(KNN), Neural Network(NN), dan lain –lain [1] [2].

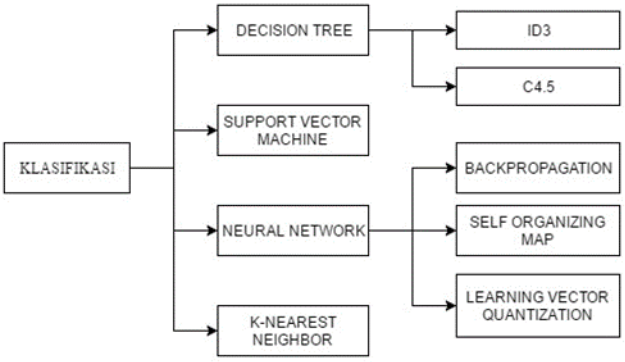
Algoritma Decision Tree sendiri terdiri dari berbagai macam seperti C4.5 dan ID3(Id et al., 2008). Sedangkan untuk algoritma Neural Network terdapat berbagai macam jenis seperti Backpropagation (Aryasa, 2012; Suwardi et al., 2012; Zamani & Amaliah, 2012), Learning Vector Quantization [7] [8], Multi Layer Perceptron (Wirakusuma, 2015), dan lain-lain. Algoritma-algoritma tersebut dapat digunakan sebagai algoritma untuk mengklasifikasikan berbagai macam data, seperti mengklasifikan gambar (Kim et al., n.d.), verifikasi teks (Colas & Brazdil,

2006) , klasifikasi spam-mail(Id et al., 2008), klasifikasi penyakit (Fernanda, 2012; Lee & To, 2010; Shanthi, 2014) dan lain – lain.

Dari berbagai macam algoritma yang dapat digunakan untuk klasifikasi tersebut, terdapat kelemahan dan kelebihan yang dimiliki oleh masing-masing algoritma. Untuk itu, peneliti akan mereview algoritma -algoritma klasifikasi untuk mendapatkan kekurangan serta kelebihan dari masing-masing algoritma. Sehingga akan lebih memudahkan dalam memilih algoritma klasifikasi tersebut.

2. Metode

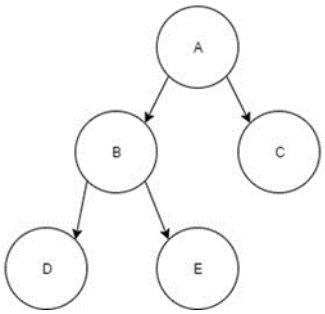
Klasifikasi merupakan teknik pengelompokan data yang memiliki ciri – ciri yang hampir sama. Seperti yang telah dijelaskan sebelumnya, terdapat beberapa macam teknik klasifikasi yang telah diulas pada penelitian sebelumnya. Oleh karena itu, pada penelitian ini peneliti akan mereview algoritma klasifikasi pada data mining. Untuk lebih jelasnya dapat dilihat pada gambar 1. Pada gambar tersebut algoritma-algoritma klasifikasi telah dikelompokkan menurut jenisnya.



Gambar 1. Beberapa algoritma klasifikasi

2.1. Decision Tree

Salah satu metode klasifikasi data mining adalah *Decision Tree*. *Decision Tree* memiliki konsep untuk mengubah data yang dimilikinya menjadi sebuah pohon keputusan yang didalamnya terdapat *rule* yang merupakan aturan keputusan (Rohman, 2013). Pemodelan pohon yang digunakan adalah pohon berakar yang simpulnya dianggap sebagai akar dan sisinya akan diberi arah untuk membentuk sebuah *graph* berarah (Suwondo et al., 2013). Untuk lebih jelasnya dapat dilihat pada gambar di bawah ini :



Gambar 2. Skema Decision Tree

Algoritma Decision Tree ini memiliki beberapa jenis algoritma seperti ID3 dan C4.5. Berikut merupakan cara kerja algoritma Decision Tree ID3 (Id et al., 2008):

1. Perhitungan nilai *Gain :*

$$Gain(S, A) = \text{Entropy (S)} - \sum_{v(A1)} \left|\frac{Si}{S}\right| \; Entropy\,(Si) \quad (1)$$

Dimana :

$$\text{Entropy(S)} = \text{- P} + \log_2 \text{ P} + \text{- P} - \log_2 \text{P} \quad (2)$$

2. Pemilihan atribut yang memiliki *gain* terbesar.
3. Pembentukan simpul.
4. Mengulangi proses dengan tidak mengikut sertakan atribut yang memiliki nilai *gain* tertinggi.

Sedangkan untuk algoritma Decision Tree C4.5 pemilihan atribut dilakukan dengan menggunakan rumus *gain ratio* seperti berikut ini :

$$\text{Gain Ratio(S, A)} = \frac{Gain(S,A)}{SplitInfo(S,A)} \quad (3)$$

Dimana :

$$\text{SplitInfo(S, A)} = -\sum_{i=1}^{n} \frac{Si}{S} \log 2 \frac{Si}{S} \quad (4)$$

Sehingga algoritma Decision Tree C4.5 lebih unggul dengan kemampuan mengolah data numerik dan diskrit jika dibandingkan dengan algoritma Decision Tree ID3 (Id et al., 2008).

2.2. Support Vector Machine

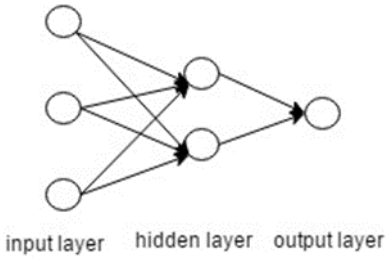
Algoritma Support Vector Machine merupakan salah satu algoritma klasifiasi yang menggunakan fungsi pemisah(klasifier/*hyperplane*) terbaik untuk memisahkan dua macam objek pada *input space*(Gitasari, 2015). Support Vector Machine merupakan metode supervised learning yang digunakan untuk klasifikasi dan regresi dengan keefektifan proses dalam mengenali pola berdasarkan prinsip meminimalisasi struktural (Lee & To, 2010). Hyperplane akan memisahkan training sample yang bernilai postif dan negatif dan menghasikan jarak maksimum antara margin dengan hyperplane. Margin adalah jarak antara hyperplane dengan data kelas (Kinerja et al., 2016). Jika tidak ada hyperplane yang mampu memisahkan nilai positif dan negatif, SVM akan memilih hyperplane yang memisahkan *sample* sebisa mungkin (Nayak et al., 2015). Support Vector Machine tidak sensitif dengan variasi parameter yang digunakan serta tidak rentan dengan *overfitting* seperti penggunaakn polinomial kernel tingkat tinggi (Burbidge et al., n.d.). Parameter-parameter yang berpengaruh dalam kinerja *Support Vector Machine* adalah(Kinerja et al., 2016) :

1. Alpha, nilai multiplier yang didapat dari proses *training.*
2. Bias, nilai bias yang didapatkan dari *training*.

Untuk fungsi kernel, dapat digunakan kernel linear, quadratic, rbf, dan polynomial.

2.3. Neural Network

Algoritma *Neural Network* (NN) merupakan salah satu algoritma data mining yang memiliki konsep rekayasa pengetahuan dengan mengadopsi sistem syaraf manusia (Kinerja et al., 2016). *Neural Network* adalah set unit input dan output yang terhubung dimana tiap relasinya memiliki bobot yang terdiri dari sejumlah besar unit pemroses yang disebut *neuron*. *Neuron* memiliki relasi dengan *synapse* yang mengelilingi *neuron-neuron* lainnya. Susunan syaraf tersebut dipresentasikan dalam *Neural Network* berupa *graph* yang menghubungkan neuron-neuron dan berkorespondensi dengan *synapse* (Saputra, 2014). Untuk lebih jelasnya dapat dilihat pada Gambar 3 berikut ini:



Gambar 3. Skema Algoritma Neural Network

Terdapat beberapa jenis neural network, diantaranya yaitu *self organizing map*, *backpropagation*, dan *learning vector quantization*.

Algoritma self organizing map sendiri merupakan salah satu jenis algoritma neural network unsupervised learning yang digunakan untuk mempelajari ditribusi himpunan pola tanpa informasi. Algoritma ini memiliki dua buah layer yaitu layer input dan output dimana di dalam layer-layer tersebut terdapat neuron yang terhubung satu sama lain (Hidayat & Shofa, 2016). Perhitungannya dilakukan dengan menginisialisasi bobot w_0 dengan nilai random. Kemudian mengatur besarnya nilai dari *learning rate*, pengurangan *learning rate* dan *MaxEpoch*.

Algoritma backpropagation merupakan salah satu jenis algoritma neural network yang menggunakan beberapa layer dalam akan lebih baik dalam mengenali pola input jika dibandingkan dengan algoritma self organizing map yang hanya memiliki 2 buah layer saja (Afifah, n.d.). Algoritma ini sukses dalam penyelesaian berbagai macam masalah yang sulit untuk dipecahkan (Fernanda, 2012). Backpropagation mempunyai pengaturan hubungan yang sederhana dengan mengatur jika terdapat keluaran yang salah, maka bobot akan dikoreksi dengan harapan respon neuron akan menghasilkan hasil yang mendekati nilai yang benar (Rifai, 2011). Langkah-langkah algoritma backpropagation adalah sebagai berikut (Suwardi et al., 2012):

1. Inisialisasi bobot jaringan secara acak
2. Menghitung keluaran berdasarkan bobot jaringan
3. Menghitung nilai eror untuk setiap keluaran dengan persamaan :
4. Melakukan langkah 2 hingga mendapatkan kondisi yang diinginkan

$$Ei = 0i\ (1 - 0i)(Ti - 0i) \quad (5)$$

dan hidden node dengan persamaan :

$$\mathrm{Ei} = 0\mathrm{i}\,(1 - 0\mathrm{i}) \sum_j \mathrm{Ej}\,\mathrm{Wj} \quad (6)$$

Keterangan :

Ei = Eror pada output layer
Oi = keluaran dari output unit i
Ti = nilai dari output dalam data *training*
Ej = Eror pada unit j
Wj = Bobot antara dua node

Sedangkan algoritma learning vector quantization merupakan algoritma supervised learning yang perhitungannya lebih cepat dibandingkan algoritma neural network lain. Neuron yang ada pada input layer langsung terhubung dengan neuron yang ada pada output layer (Arifianto et al., 2014). Berikut merupakan langkah – langkah yang ada pada algoritma LVQ (Mafrur et al., 2008):

1. Menetapkan bobot awal dan *MaxEpoch*, *learning rate*, dan *error* yang diharapkan
2. Memasukkan data x dan juga targetnya (T)
3. Menetapkan kondisi awal (epoch = 0), *error* yang diharapkan = 1
4. Memproses jika nilai *epoch* < *MaxEpoch* dan learning rate lebih besar dari *error* yang diharapkan
5. Mengurangi nilai pengurangan learning rate.

Dari ulasan tiga jenis algoritma *Neural Network* tersebut, dapat disimpulkan bahwa algoritma *Learning Vector Quantization* lebih unggul jika dibandingkan dengan algoritma *Backpropagation* dan *Self Organizing Map*.

2.4. K-Nearest Neighbor

Algoritma K-Nearest Neighbor merupakan sebuah algoritma yang sederhana yang digunakan untuk mengklasifikasikan objek berdasarkan data pembelajaran yang jaraknya dekat dengan objek (Nursalim et al., 2014). Algoritma ini juga merupakan salah satu tenik dari algoritma *lazy learning* dan masuk dalam kelompok *instance-based learning* (Leidiyana, 2013). Pada proses klasifikasinya, objek query yang tidak memiliki label akan diberi label k dari tetangga terdekatnya. Jika nilai k =1, maka objek digolongkan sebagai kelas dari objek terdekat. Namun jika banyaknya kelas adalah 2, maka nilai k harus bilangan bulat ganjil (Kim et al., n.d.).

Algoritma ini tidak memiliki proses training(Syafitri, 2010), data akan langsung memasuki proses testing dengan menghitung jarak masing-masing *training sample*. Lalu akan ditentukan sejumlah k point yang dekat dengan objek query. Untuk mencari jarak, dapat digunakan rumus (Kim et al., n.d.) dimana d adalah jarak dan p adalah dimensi data :

$$d(x, y) = \sum_{i=1}^{m} (xi - yi) \quad (5)$$

Keterangan :

m = dimensi data
x dan y = koordinat histogra.

3. Hasil dan Pembahasan

3.1. Algoritma support vector machine dan sistem case based reasoning

Dari keseluruhan metode yang telah diulas sebelumnya, algoritma yang memiliki keunggulan lebih jika dibandingkan dengan algoritma lainnya adalah algoritma *Support Vector Machine*. Hal tersebut dapat lihat pada Tabel 1.

Tabel 1. Perbandingan Algoritma Klasifikasi

Algoritma	Kelebihan	Kekurangan
Decision Tree	Kecepatan klasifikasi (Kotsiantis, 2007) Kontinyu (Kotsiantis, 2007) Transparasi klasifikasi(Kotsiantis, 2007)	Rendanya akurasi yang dihasilkan (Suwondo et al., 2013)
Neural Network	Akurasi yang dihasilkan tinggi Performa lebih baik dari Decision Tree (Kotsiantis, 2007)	Tingkat konvergensi rendah (Lee & To, 2010) Membutuhkan data training yang banyak(Lee & To, 2010) Solusi optimal yang hanya bersifat lokal(Lee & To, 2010)
K-Nearest Neighbor	Dapat bekerja dengan baik dengan kelas multi modal (Kim et al., n.d.)	Penggunaan atribut yang dapat memicu eror (Kim et al., n.d.) Proses komputasi yang lama (Kinerja et al., 2016)
Support Vector Machine	Mampu bekerja dengan baik meskipun dengan sedikit data training (Kim et al., n.d.) Tingkat akurasi tinggi (Kotsiantis, 2007) Kecepatan klasifikasi (Kotsiantis, 2007)	Waktu pemrosesan yang lama (Colas & Brazdil, 2006)

Support Vector Machine sendiri memiliki prinsip SRM(Structural Risk Minimization) untuk menemukan hyperplane terbaik pada input space. Support Vector Machine sendiri merupakan algoritma yang menerapkan prinsip linear classifier, Suppot Vector Machine dapat menyelesaikan problem non-linear dengan memasukkan konsep kernel trick (Aulia et al., 2015). Hyperplane akan memisahkan training sample yang bernilai postif dan negatif dan menghasikan jarak maksimum antara margin dengan hyperplane. Margin adalah jarak antara hyperplane dengan data kelas (Kinerja et al., 2016). Jika tidak ada hyperplane yang mampu memisahkan nilai positif dan negatif, Support Vector Machine akan memilih hyperplane yang memisahkan sample sebisa mungkin (Nayak et al., 2015). Support Vector Machine tidak sensitif dengan variasi parameter yang digunakan serta tidak rentan dengan overfitting seperti penggunaakn polinomial kernel tingkat tinggi (Burbidge et al., n.d.). Parameter-parameter yang berpengaruh dalam kinerja Support Vector Machine adalah (Kinerja et al., 2016) :

1. Alpha, nilai multiplier yang didapat dari proses training.
2. Bias, nilai bias yang didapatkan dari training.
3. Untuk fungsi kernel, dapat digunakan beberapa kernel

Linear, bersifat intuitif, mudah untuk divisualisasikan dan mampu memberikan klasifikasi yang baik. Namun tidak mampu memberikan output probabilistik dan tidak dapat melakukan regresi dengan rumus :

$$K(xi, xj) = xiTxj \quad (6)$$

Rbf, memiliki ciri – ciri bahwa setiap fungsi hanya tergantung pada jarak radial dari pusat.

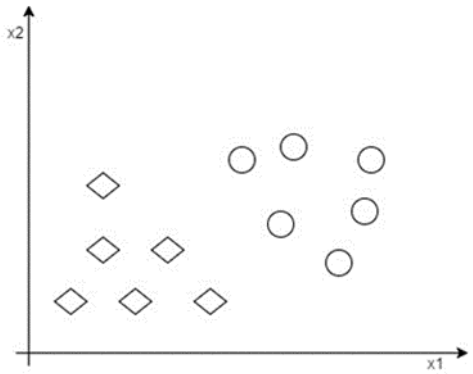
$$K(xi, xj) = (\gamma xiTxj + r)d, \gamma > 0 \quad (7)$$

Polynomial, memiliki tambahan parameter yaitu parameter fungsi Gauss yang akan mempengaruhi skala fungsi non-linear.

$$K(xi, xj) = \exp(-\gamma \|xi - xj\|2), \gamma > 0 \quad (8)$$

Keterangan :

γ = gamma untuk seluruh fungsi kernel kecuali linier
d = Derajat polinomial dalam fungsi kernel
γ, d, dan r merupakan parameter yang dikontrol oleh user untuk menambah tingkat akurasi.



Gambar 3 menunjukkan pengelompokan dua buah kelas yang dilakukan oleh *Support Vector Machine*.

Untuk mencari besar *dari hyperplane*, dapat digunakan persamaan berikut (Shanthi, 2014) :

$$w.\emptyset(x_i) + b = 0 \quad (9)$$

Dimana w dan b merupakan parameter klasifikasi dan ϕ adalah fungsi klasifikasi dalam dimensi yang lebih tinggi guna untk memisahkan x_i secara linear.

Support Vector Machine merupakan algoritma pembelajar yang kuat karena data latih hanya ada pada produk skalar untuk mempelajari *hyperplane* dalam dimensi tertentu. Sehingga *Support Vector Machine* dapat digunakan pada dua lapisan seperti jaringan syaraf sigmoid dan jaringan *radial basis function*.

Dalam implementasinya, algortima *Support Vector Machine* dapat diimplementasikan dengan algoritma-algoritma lain dengan tujuan untuk meningkatkan performa ataupun meningkatkan akurasi. Salah satu contohnya adalah dengan mengkombinasikan algoritma *K-Nearest Neighbor* dengan *Support Vector Machine* (K-SVNN) (Kinerja et al., 2016). Hasil dari penggabungan dua metode tersebut adalah dengan dihasilkannya akurasi yang lebih tinggi dan waktu training yang lebih singkat. Contoh lainnya adalah penggunaan *Case Based Reasoning*

dengan algoritma *K-Nearest Neighbor*(Zainuddin et al., 2016). Dari penelitian tersebut, ditunjukkan bahwa hanya dengan 15 kasus yang dipelajari oleh sistem *Case Based Reasoning*, mampu menghasilkan akurasi sebesar 93,3%.

Case Based Reasoning sendiri adalah salah satu sistem pemecahan masalah dengan mempelajari permasalahan-permasalahan yang lama untuk menemukan solusi dalam permasalahan yang baru (Aamodt, 1994). Terdapat beberapa tahap dalam penyelesaiannya, yaitu (Choudhury, 2016) :

1. *Retrieve*, untuk mendapatkan permasalahan-permasalahan serupa.
2. *Reuse*, pemakaian kembali informasi dari permaslaahan-permasalahan sebelumnya
3. *Revise*, menambahkan solusi yang berkaitan
4. *Retain*, penggunaan solusi baru yang dapat digunakan untuk kasus selanjutnya

Jika *Case Based Rasoning* akan digunakan bersamaan dengan algoritma lain, maka algoritma tersebut akan berjalan setelah *Case Based Reasoning* selesai.

4. Simpulan

Simpulan dapat bersifat generalisasi temuan sesuai permasalahan penelitian, dapat pula berupa rekomendatif untuk langkah selanjutnya.

Daftar Rujukan

Daftar rujukan ditulis menggunakan gaya APA edisi keenam

De Vaus, D. A. (2014). *Surveys in social research*. Sydney, Australia: Allen & Unwin.

McKenzie, H., Boughton, M., Hayes, L., & Forsyth, S. (2008). Explaining the complexities and value of nursing practice and knowledge. In I. Morley & M. Crouch (Eds.), *Knowledge as value: Illumination through critical prisms* (pp. 209-224). Amsterdam, Netherlands: Rodopi.

Putra, E. M., Handarini, D. M., & Muslihati, M. (2019). Keefektifan achievement motivation training untuk meningkatkan motivasi berprestasi siswa sekolah menengah pertama. *Jurnal Kajian Bimbingan dan Konseling, 4*(2), 62-68.

Scheinin, P. (2009). Using student assessment to improve teaching and educational policy. In M. O'Keefe, E. Webb, & K. Hoad (Eds.), *Assessment and student learning: Collecting, interpreting and using data to inform teaching* (pp. 12-14). Melbourne, Australia: Australian Council for Educational Research.

Makmara. T. (2009). *Tuturan persuasif wiraniaga dalam Berbahasa Indonesia: Kajian etnografi komunikasi*. (Unpublished master's thesis) Universitas Negeri Malang, Malang, Indonesia.