OPTIMASI TEKNIK KLASIFIKASI MODIFIED K NEAREST NEIGHBOR MENGGUNAKAN ALGORITMA GENETIKA

Optimization Techniques Modified k Nearest Neighbor Classification Using Genetic Algorithm

Siti Mutrofin[1], Abidatul Izzah[2], Arrie Kurniawardhani[3], Mukhamad Masrur[4]

Jurusan Sistem Informasi, Fakultas Teknik, Unipdu, Kompleks Ponpes Darul 'Ulum Peterongan Jombang, 61481

ABSTRACT

One of the tasks of data mining is classification, many researchers are already conducting research on the method of classification. Classification method used is k-Nearest Neighbor (kNN). kNN has several advantages, including the training is fast, simple and easy to learn, resistant to the training data that has noise, and effectively if the large training data. Meanwhile, the lack of kNN is the value of k bias, complex computing, memory limitations, and easily fooled by irrelevant attributes. One improvement is the Modified kNN kNN (MKNN), which aims to improve the accuracy of the kNN, by adding the calculation of validity, because it is considered the weight calculations contained in kNN, have problems outlier. However, MKNN also have the same disadvantages as kNN k value bias and complex computing. Based on the MKNN problems, the author intends to make improvements in terms of, optimization Genentic value k using Algorithm (GA), because GA has proven it can be used to optimize the value of k for kNN. Furthermore, the algorithm will be called GMKNN algorithm (Genetic Modified k Nearest Neighbor). Evaluation level of truth results will be based on the value of accuracy, either using kNN algorithm, MKNN and GMKNN use data UCI machine learning.

Keywords : kNN , Modified kNN , Genetic Algorithm , Genetic Modified kNN , UCI Machine Learning

ABSTRAK

Salah satu tugas dari data mining adalah klasifikasi, banyak peneliti yang sudah melakukan penelitian tentang metode klasifikasi. Metode klasifikasi yang biasa digunakan adalah k-Nearest Neighbor (kNN). kNN memiliki beberapa kelebihan, diantaranya adalah pelatihan sangat cepat, sederhana dan mudah dipelajari, tahan terhadap data pelatihan yang memiliki derau, dan efektif jika data pelatihan besar. Sedangkan, kekurangan dari kNN adalah nilai k bias, komputasi kompleks, keterbatasan memori, dan mudah tertipu dengan atribut yang tidak relevan. Salah satu perbaikan kNN adalah Modified kNN (MKNN), yang bertujuan untuk meningkatkan akurasi dari kNN, dengan menambahkan perhitungan validity, karena dianggap perhitungan bobot yang terdapat pada kNN, memiliki permasalahan outlier. Namun, MKNN juga memiliki kelemahan yang sama dengan kNN yaitu nilai k bias dan komputasi yang kompleks. Berdasarkan permasalahan MKNN tersebut, penulis bermaksud untuk melakukan perbaikan dalam hal, optimasi nilai k menggunakan Genentic Algorithm (GA), karena GA sudah terbukti dapat digunakan untuk melakukan optimasi pada nilai k untuk kNN. Selanjutnya algoritma tersebut akan dinamakan algoritma GMKNN (Genetic Modified k Nearest Neighbor). Evaluasi tingkat kebenaran hasil akan didasarkan pada nilai akurasi, baik menggunakan algoritma kNN, MKNN maupun GMKNN menggunakan data UCI machine learning.

Kata Kunci : kebijakan publik, Supiory, daerah tertinggal dan pengembangan

PENDAHULUAN

Klasifikasi adalah salah satu tugas dari data mining yang bertujuan untuk memprediksi label kategori benda yang tidak diketahui sebelumnya, dalam membedakan antara objek yang satu dengan yang lainnya berdasarkan atribut atau fitur [1,2]. Salah satu teknik klasifikasi yang paling dasar dan sederhana adalah k Nearest Neighbor (kNN) [1]. kNN memiliki beberapa keunggulan dan kelemahan [3], keunggulannya yaitu:

1) Pelatihan sangat cepat, 2) Sederhana dan mudah dipelajari, 3) Tahan terhadap data pelatihan yang memiliki derau, dan 4) Efektif jika data pelatihan besar. Sedangkan, kekurangan dari kNN adalah: 1) Nilai k bias, 2) Komputasi kompleks, 3) Keterbatasan memori, dan 4) Mudah tertipu dengan atribut yang tidak relevan.

Banyak peneliti yang sudah melakukan penelitian tentang perbaikan kNN, baik dalam memperbaiki nilai akurasi kNN maupun dalam hal optimasi nilai k pada kNN. Penelitian terkait peningkatan nilai akurasi kNN adalah Modified kNN (MKNN), di mana dalam MKNN ditambahkan perhitungan nilai validity yang berguna untuk mengatasi permasalahan outlier dalam perhitungan nilai bobot pada kNN tradisional. Namun MKNN juga memiliki kelemahan, yaitu nilai k yang bias dan komputasi yang kompleks [1]. Sedangkan, penelitian terkait optimasi nilai k pada kNN adalah Genetic kNN (GKNN), selain dapat menentukan nilai k secara otomatis, GKNN dapat meningkatkan nilai akurasi dan dapat mengurangi kompleksitas [4].

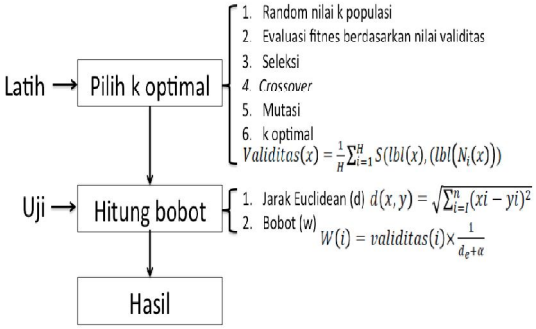
Berdasarkan latar belakang tersebut, penulis mengusulkan perbaikan pada algoritma MKNN, dengan cara mengoptimasi nilai k menggunakan algoritma genetika, selanjutnya algoritma tersebut dinamakan algoritma Genetic Modified k Nearest Neighbor (GMKNN).

METODE PENELITIAN

Untuk mendukung penelitian, maka perlu mengkaji berbagai referensi tentang tugas dari data mining, di mana salah satu tugas dari data mining adalah klasifikasi. Tenik klasifikasi yang umum digunakan adalah kNN, namun dengan segala kelebihan dari kNN, kNN masih memiliki kelemahan, diantara banyak penelitian tentang perbaikan kNN, algoritma GA diusulkan untuk penentuan nilai k otomatis pada kNN (GKNN), yang juga berdampak dapat meningkatkan akurasi dan mengurangi kompleksitas [4]. Selain GKNN sebagai perbaikan akurasi dari kNN, MKNN juga bertujuan untuk memperbaiki algoritma kNN dalam segi perbaikan nilai akurasi kNN, karena nilai k dari MKNN masih bersifat bias [1], maka penulis mengusulkan peningkatan kinerja MKNN dengan optimasi nilai k menggunakan GA.

Data yang akan diteliti adalah data UCI machine learning yaitu data iris dan data wine. Data iris adalah data bunga iris yang terdiri dari 4 atribut, 3 kelas dan 150 data. Sedangkan, data wine terdiri dari 13 atribut, 3 kelas, dan 178 data.

Tahap awal desain sistem adalah merumuskan kontribusi utama dari penelitian ini. Dari studi literatur yang telah dilakukan, diketahui bahwa MKNN merupakan algoritma yang memiliki akurasi yang lebih baik dibandingkan kNN. Namun, MKNN masih belum mampu mengatasi permasalahan kNN dalam hal nilai k yang masih bias. Oleh karena itu perlu dilakukan optimasi menggunakan GA sebagai operator dalam menentukan nilai k yang bertujuan mengatasi kelemahan tersebut (untuk selanjutnya MkNN dengan adanya penambahan operator GA disebut dengan GMKNN). Kemudian algoritma hybrid ini digunakan sebagai model pada saat klasifikasi. Model GMKNN ditunjukkan pada Gambar 1.



Desain sistem yang diusulkan pada penelitian ini (dapat dilihat pada Gambar 1) terdiri dari beberapa tahap yakni sebagai berikut:

- Masukkan data latih. Misal data latih yang di gunakan adalah data Iris sebanyak 125 data.
- Tentukan populasi dari kromosom

(solusi). Misal populasi yang diinginkan adalah 3. Maka secara random akan dibangkitkan kromosom (kemungkinan solusi) sebanyak 3 buah, dengan ketentuan, nilai k < 125. Misal hasil kromosom didapatkan 3, 9, 4. Selanjutnya nilai kromosom tersebut dibinerkan, misal menjadi 0011 untuk kromosom 3, 0100 untuk kromosom 4 dan 1001 untuk kromosom 9.

- Hitung nilai fitness dengan menggunakan nilai validitas (validity) [1], nilai validity tertinggi adalah nilai fitness terbaik. Misal yang terbaik adalah kromosom 4.

$$S(a,b) = \begin{cases} 1 \; jika \; a = b \\ 0 \; jika \; a \neq b \end{cases} \quad (1)$$

Keterangan:
S : nilai similaritas antar data latih
a dan b : label kelas antar data latih.

$$Validity(x) = \frac{1}{H}\sum_{i=1}^{H} S(lbl(x), lbl(N_i(x))) \quad (2)$$

Keterangan:
i : banyaknya data latih
k : jumlah tetangga terdekat antar data latih dari similaritas yang terbaik (nilai 1 = terbaik).
N_i adalah banyaknya label kelas.

- Lakukan proses seleksi menggunakan roulette wheel, misalkan yang terpilih adalah kromosom 4 dan 9.
- Lakukan proses crossover dari dua kromosom yang telah terpilih pada langkah no. 4, dikarenakan operasi ini diharapkan memiliki keberhasilan tinggi, maka digunakan probabilitas sebesar 0.8.
- Lakukan proses mutasi dari hasil anakan pada langkah no. 5, dikarenakan operasi ini diharapkan memiliki kemungkinan kecil, maka digunakan probabilitas sekecil-kecilnya, misal 0.001.
- Didapatkan individu baru dari nilai fitness yang terbaik.
- Ulangi tahapan tahapan di operasi GA sampai memiliki nilai k (kromosom) yang optimal.
- Lakukan perhitungan jarak euclidean [1] dari data uji ke data latih.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(xi - yi)^2} \quad (3)$$

Keterangan:
d(x,y) : jarak antara data uji dengan data latih
x : data uji,
y : data latih,
n : jumlah data latih.

- Lakukan perhitungan bobot (w) [1] dengan mempertimbangkan nilai validitas dan jarak.

$$w(i) = validitas(i) \times \frac{1}{d(i)+\alpha} \quad (4)$$

Keterangan:
w : bobot
α : nilai smooting (pemulusan), dalam penelitian nilai yang digunakan adalah 0,5 [1].

- Nilai bobot terbesar adalah prediksi kelas dari data uji

HASIL DAN PEMBAHASAN

Dataset yang digunakan adalah dataset bunga Iris yang terdiri dari 150 data, 4 atribut, 3 kelas, setiap kelas memiliki 50 data. Sedangkan, data Wine terdiri dari 178 data, 13 atribut, 3 kelas, dan setiap kelas memiliki jumlah data yang berbeda, kelas 1 memiliki 59 data, kelas 2 memiliki 71 data, dan kelas 3 memiliki 48 data. Berdasarkan uji coba pada kedua dataset, dengan beberapa skenario uji coba baik untuk data latih 90%, 80%, dan 70%. Untuk dataset bunga Iris memiliki akurasi 100% baik menggunakan algoritma kNN, MKNN, maupun GMKNN dengan baik pada k = 1, 2, ataupun 3. Hal itu disebabkan tiap kelas memiliki jumlah data yang sama, yaitu sejumlah 50 data (perhatikan Tabel 1). Namun, ketika setiap kelas memiliki jumlah data yang tidak sama, seperti pada kasus data Wine, maka akurasi terbaik hanya mampu mencapai di angka 80-an saja, dan untuk

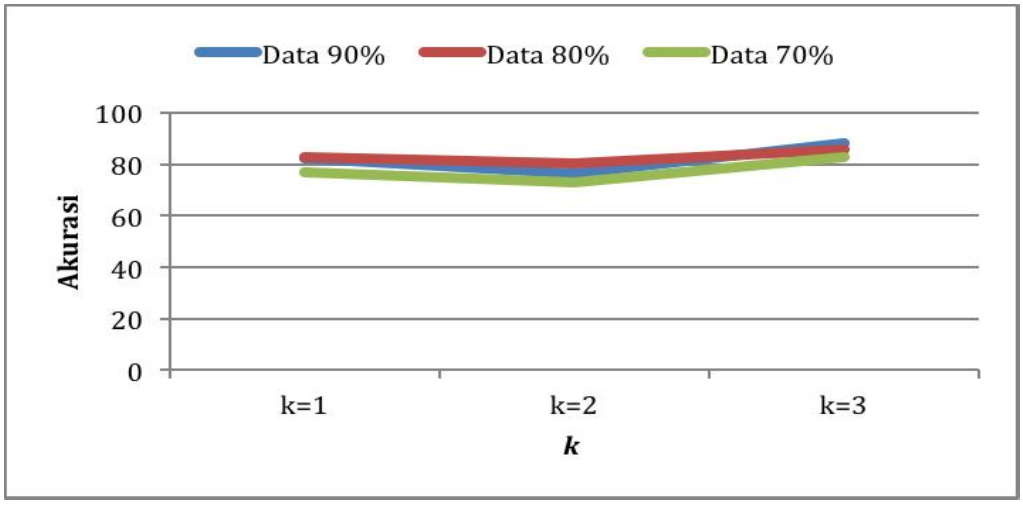
semua pengujian k yang optimal adalah k = 3. Pada percobaan data Wine dengan menggunakan kNN, hasil menunjukkan bahwa jumlah data berbanding lurus dengan tingkat akurasi. Lebih jelas akan diperlihatkan pada Tabel 2 dan Grafik 1.

Tabel 1 Hasil Uji Coba Klasifikasi kNN, MKNN pada data iris

Jumlah Data	Jumlah Per kelas			Data Latih(%)	k	Akurasi (%)
	Kelas1	Kelas2	Kelas3			
150	50	50	50	90	1	100
					2	100
					3	100
				80	1	100
					2	100
					3	100
				70	1	100
					2	100
					3	100

Tabel 2 Hasil Uji Coba Klasifikasi kNN pada data Wine

Jumlah Data	Jumlah Per kelas			Data Latih (%)	k	Akurasi (%)
	Kelas1	Kelas2	Kelas3			
178	59	71	48	90	1	82.35
					2	76.47
					3	88.23
				80	1	82.86
					2	80
					3	85.71
				70	1	76.92
					2	73.02



Grafik 1 Jumlah data berbanding lurus dengan tingkat akurasi

KESIMPULAN DAN SARAN

Algoritma kNN, MKNN dan GMKNN memiliki kinerja yang sama baiknya, dalam melakukan klasifikasi data Iris dengan hasil akurasi 100%. Keunggulan dari algoritma GMKNN adalah dapat menentukan nilai k optimal pada MKNN dengan otomatis, tanpa harus mencoba satu persatu dalam menentukan nilai k.

DAFTAR PUSTAKA

Parvin H, Alizadeh H, Bidgoli B M. MKNN: Modified K-Nearest Neighbor. Proceedings of the Word Congress on Engineering and Computer Science 2008 (WCECS 2008). San Francisco. 2008: 831-834.

Ngai E W T, Hu Y, Wong Y H, Chen Y, Sun X. The Application of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and An Academic Review of Literature. Decision Support Systems. 2011; 50(3): 559-569.

Bhatia N, Vandana. Survey of Nearest Neighbor Techniques. International Journal of Computer Science and Information Security. 2010. 8(2): 302-305.

Suguna N, Thanushkodi K. An Improved k-Nearest Neighbor Classification Using Genetic Algorithm. International Journal of Computer Science Issues. 2010. 7(4): 18-21.