Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 1, No. 2, Februari 2017, hlm. 75-79 http://j-ptiik.ub.ac.id

Implementasi *Modified K-Nearest Neighbor* Dengan Otomatisasi Nilai K Pada Pengklasifikasian Penyakit Tanaman Kedelai

Tri Halomoan Simanjuntak [1], Wayan Firdaus Mahmudy [2], Sutrisno [3]

Fakultas Ilmu Komputer, Universitas Brawijaya
Email: trihalomoans[at]gmail.com[1], wayanfm[at]ub.ac.id[2], trisno[at]ub.ac.id[3]

Abstrak

Berbagai serangan penyakit dan hama dapat menimbulkan masalah yang serius terhadap tanaman kedelai. Salah satu ancaman pengembangan tanaman kedelai bagi balai-balai penelitian dan pihak pengembang tanaman tersebut adalah gangguan hama. Serangan hama dapat menurunkan hasil kedelai hingga 80% bahkan lebih jika tidak ada pengendalian yang serius. Diperlukan klasifikasi untuk menentukan jenis penyakit yang menyerang tanaman kedelai. Penelitian ini menggunakan *Soybean Disease Data Set* yang terdiri dari 266 data latih dan akan dibangun aplikasi berbasis *desktop* dengan mengimplementasikan algoritma *Modified K-Nearest Neighbor*, parameter nilai *K* ditentukan oleh sistem dengan menggunakan metode *Brute Force* sehingga menemukan nilai *K* terbaik. Setiap nilai *K* dengan akurasi hasil terbaik akan disimpan dan digunakan sebagai parameter nilai *K* pada proses pengujian data baru. Nilai *K* pada metode ini mendefinisikan jumlah tetangga terdekat yang digunakan untuk proses klasifikasi. Hasil pengujian menunjukkan bahwa parameter nilai *K* sangat berpengaruh terhadap hasil klasifikasi dan akurasi yang dihasilkan. Rata-rata akurasi cenderung menurun seiring dengan penambahan nilai *k* sedangkan peningkatan jumlah data latih turut disertai dengan peningkatan hasil akurasi, untuk data latih dengan kelas tidak seimbang mengalami penurunan nilai akurasi seiring dengan bertambahnya jumlah data. Hasil akurasi tertinggi pada pengujian ini sebesar 100% dengan nilai *k*=1 dan rata-rata akurasi dari 5 percobaan sebesar 98,83%.

Kata kunci: *Klasifikasi, Modified K-Nearest Neighbor, Penyakit Tanaman Kedelai*

Abstract

Various diseases and pest attacks can cause serious problems to the soybean crop. One threat to the soybean crop development research centers and is the developer of the plant pests. Pests can reduce soybean yields by 80 % even if no serious control. Classification is needed to determine the types of diseases that attack soybean plants. This research use of Soybean Disease Data Set consisting of 266 training data and desktop-based applications to be built by implementing the algorithm Modified K - Nearest Neighbor, the parameter value of K is determined by the system using brute force methods to find the best K value. Each value of K with accuracy the best results will be recorded and used as the parameter value of K in the process of testing new data. K values in this method to define the number of nearest neighbors used for the classification process. The test results showed that the value of the parameter K affects the classification results and the accuracy result. Average accuracy tends to decrease with the addition of the value of k, while increasing the number of training data also accompanied by an increase in the accuracy of the results, for training data with imbalanced class accuracy values decreased with increasing amount of data. The results of the highest accuracy on the test at 100 % with a value of k = 1 and an average accuracy of 5 times the experimental is 98.83 %.

Keywords: *Classification, Modified K-Nearest Neighbor, Soybean Plant Diseases*

1. PENDAHULUAN

Kedelai merupakan salah satu komoditas yang mempunyai fungsi multiguna dan sumber utama minyak nabati dunia. Berbagai serangan penyakit dan hama dapat menimbulkan masalah yang serius terhadap tanaman kedelai. Salah satu ancaman pengembangan tanaman kedelai bagi balai-balai penelitian dan pihak pengembang tanaman tersebut adalah gangguan

hama. Serangan hama dapat menurunkan hasil kedelai hingga 80% bahkan lebih jika tidak ada pengendalian yang serius (Marwoto, 2007). Hingga saat ini balai-balai penelitian maupun pihak pengembang tanaman kedelai masih kesulitan untuk mengidentifikasi jenis penyakit tanaman tersebut. Beberapa kendala adalah lemahnnya dalam identifikasi hama dan gejala serangan, dan tindakan pengendalian yang terlambat (Marwoto, 2007). Mengidentifikasi jenis penyakit dengan tepat menjadi masalah penting yang harus diselesaikan saat ini.

Seiring dengan perkembangan teknologi yang semakin pesat, maka dibuatlah suatu *software* yang dapat mengklasifikasikan penyakit pada tanaman kedelai. Sehingga dapat membantu peneliti untuk mengetahui jenis dari suatu penyakit tanaman kedelai. Klasifikasi merupakan proses mengidentifikasi obyek ke dalam sebuah kelas, grup, atau ketegori berdasarkan prosedur, karakteristik & definisi yang telah ditentukan sebelumnya (Han dan Kamber. 2006).

Algoritma *Modified K-Nearest Neighbor (MKNN)* merupakan pengembangan performansi dari metode *K-Nearest Neighbor (KNN)*. Pemikiran utama dari metode ini adalah pengklasifikasian *sample* uji sesuai tag tetangganya. *MKNN* terdiri dari dua pemrosesan, pertama validasi data *training* dan yang kedua adalah menerapkan pembobotan *KNN* (Parvin, Alizadeh dan Minae-Bidgoli, 2008).

Nilai k sangat berpengaruh terhadap hasil keakurasian data dan harus melalui serangakaian percobaan pendahuluuan. Tingkat akurasi *MKNN* terbukti lebih baik jika dibandingkan dengan metode sebelumnya yaitu *KNN*. Terbukti dari penelitian sebelumnya pada dataset *Balance-sc* metode *KNN* mempunyai tingkat akurasi sebesar 80.69% sedangkan metode *MKNN* 85.49%, begitu juga pada data set *Monk 1* metode *KNN* memiliki tingkat akurasi 84.49% sedangkan metode *MKNN* 87.81% (Parvin, Hoseinali dan Minati. 2010). Berdasarkan uraian tersebut pada penelitian ini dibangun *software* yang dapat mengklasifikasi jenis penyakit pada tanaman kedelai dengan metode *Modified K-Nearest Neighbor (MKNN)* dengan otomatisasi nilai K.

2. TANAMAN KEDELAI

Kedelai , *(Glycine max (L) Merril),* merupakan komoditi tanaman pangan nomor tiga setelah padi dan jagung. Sampai saat ini diduga berasal dari kedelai liar China, Manchuria dan korea. Klasifikasi tanaman kedelai adalah sebagai berikut (Widodo, 1987):

Divisio : Spermatophyta
Classis : Dicotyledoseae
Ordo : Rosales
Familia : Papilionaceae
Genus : Glycine
Species : *Glycine mas* (L.) Merill

3. K – NEAREST NEIGHBOR (KNN)

K-Nearest Neighbor (KNN) merupakan metode yang biasa digunakan pada klasifikasi data. Algoritma ini digunakan untuk mengklasifikasikan terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut.

KNN merupakan suatu metode yang menggunakan algoritma *supervised* dengan hasil dari *query instance* yang baru diklasifikasikan berdasarkan mayoritas dari kategori pada KNN. Tujuan dari algoritma ini ialah mengklasifikasikan objek baru berdasarkan atribut dan *training sample* (Larose, 2005).

Prinsip umum dari algoritma ini adalah menemukan *k* data training untuk menentukan *k-nearest neighbor* berdasarkan ukuran jarak. Selanjutnya mayoritas dari *k* tetangga terdekat akan menjadi dasar untuk memutuskan kategori dari sample berikutnya (Yu, 2010). Selain itu algoritma ini sendiri sering digunakan untuk klasifikasi pada teknik data mining meskipun dapat digunakan untuk estimasi dan prediksi data.

Metode ini adalah contoh dari *instance-based learning* dimana data training di simpan, sehingga proses klasifikasi dari data yang belum diklasifikasi dapat dengan mudah ditemukan dengan membandingkan data tersebut dengan data yang paling mirip di data training yang ada (Larose, 2005).

4.1. *Modified K-Nearest Neigbor* (MKNN)

Ide utama dari metode ini adalah hal pertama yang dilakukan adalah perhitungan validitas untuk semua data yang terdapat pada data latih. Selanjutnya, dilakukan perhitungan *Weight Voting* pada semua data uji menggunakan validitas data **(**Parvin, Hoseinali dan Minati, 2010).

4.2. Normalisasi Data

Normalisasi pada penelitian ini digunakan untuk mempersempit *range* data morfologi tanaman kedelai. Normalisasi yang digunakan pada penelitian ini adalah *min-max normalization* yang merupakan proses transformasi nilai dari data yang dikumpulkan pada range *value* antara 0.0 dan 1.0, dimana nilai terkecil (*min*) adalah 0.0 dan nilai tertinggi (*max)* adalah 1.0 (Chandrasekhar, Thangavel dan Elayaraja, 2011).

$$v' = \frac{v - min_A}{max_A - min_A} (new\_max_A - new\_min_A) + new\_min_A$$

Dimana :
v' : Data Baru Setelah Normalisasi
v : Data Sebelum Normalisasi
new_max_A : Batas Nilai Max Baru adalah 1
new_min_A : Batas Nilai Min Baru adalah 0
max_A : Nilai Maximun Pada Kolom
min_A : Nilai Minimum Pada Kolom

4.3. Validitas Data Training

Validitas digunakan untuk menghitung jumlah titik dengan label yang sama untuk semua data pada data latih. Validitas setiap data tergantung pada setiap tetangga terdekatnya. Setelah dilakukan validasi data, selanjutnya data tersebut digunakan sebagai informasi lebih mengenai data tersebut. Persamaan yang digunakan untuk menghitung validitas setiap data latih adalah (Parvin, Hoseinali dan Minati, 2010):

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

Dimana :
H : Jumlah titik terdekat
Lbl(x) : Kelas x
Ni(x) : Label kelas titik terdekat x

Fungsi *S* digunakan untuk menghitung kesamaan antara titik a dan data ke-b tetangga terdekat. Persamaan untuk mendefinisikan fungsi *S* terdapat dalam persamaan dibawah ini :

$$S(a, b) = \begin{cases} 1 & a = b \\ 0 & a \neq b \end{cases} \quad (3)$$

Dimana :
a : kelas a pada data training
b = kelas lain selain a pada data training

2.5.1 *Weight Voting*

Dalam metode *MKNN*, pertama *weight* masing-masing tetangga dihitung dengan menggunakan $1 / (d_e + 1)$. Kemudian, validitas dari setiap data pada data latih dikalikan dengan *weight* berdasarkan pada jarak Euclidean. Sehingga metode *MKNN*, didapatkan persamaan *weight voting* tiap tetangga sebagai berikut :

$$W(x) = Validasi\ (x) \times \frac{1}{d_e + 0,5} \quad (4)$$

Dimana :
W(i): Perhitungan Weight Voting
Validasi (x): Nilai Validasi
d_e: Jarak Euclidean

4.4. Akurasi Sistem

Perhitungan ini dilakukan untuk mengetahui tingkat akurasi metode *MKNN* dalam hasil klasifikasi. Akurasi dapat diperoleh dari presentase kebenaran, yaitu perbandingan antara jumlah data uji dengan jumlah data keseluruhan dikalikan 100%. Akurasi bisa didapat melalui persamaan berikut :

$$Akurasi = \frac{Jumlah\ Data\ Uji\ Benar}{Jumlah\ Seluruh\ Data\ Uji} \times 100\%$$

4. METODOLOGI PENELITIAN

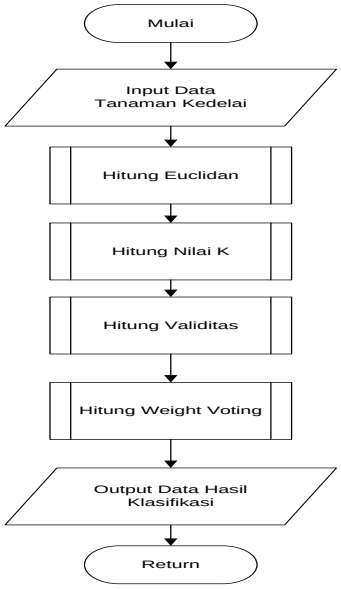
Secara umum sistem ini akan melakukan diagnosa penyakit pada tanaman kedelai menggunakan algoritma *Modified K-Nearest Neighbor*. Nilai k yang dibutuhkan pada sistem ini tidak diinputkan secara manual oleh *user*, melainkan sistem akan melakukan perhitungan dengan menggunakan algoritma *brute force*.

5.1. Data Penelitian

Dataset yang digunakan dalam penelitian adalah dataset penyakit tanaman kedelai. Kumpulan data tersebut diperoleh dari situs *Center for Machine Learning and Intelligent Systems* (http://archive.ics.uci.edu/ml/machine-learning-databases/soybean/) yang terdiri dari beberapa atribut. *Data set* ini dikategorikan menjadi 15 jenis penyakit tanaman kedelai hingga total data keseluruhan sebanyak 266 data.

5.2. Proses Klasifikasi

Algoritma MKNN terdiri dari 3 proses utama, yaitu perhitungan validitas data *training*, perhitungan jarak, dan perhitungan *weght voting*. Proses algoritma MKNN ditunjukkan pada Gambar 1.



Gambar 1. Flowchart Proses Klasifikasi MKNN

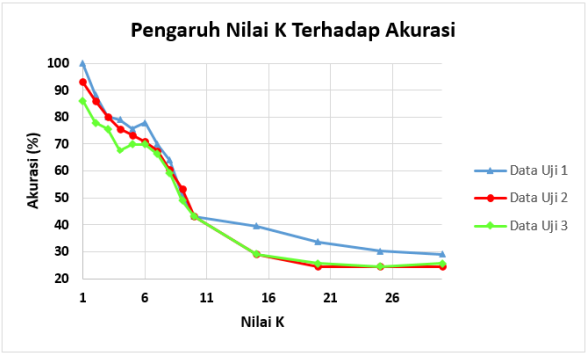
5. PENGUJIAN DAN ANALISIS

Dalam pengujian ini terdapat 266 *dataset* yang nantinya dapat diolah sebagai data latih maupun data uji. Pengujian dilakukan untuk mengetahui beberapa parameter yang mempengaruhi akurasi yang dihasilkan.

6.1 Pengujian Pengaruh Parameter K terhadap Akurasi

Berdasarkan hasil pengujian yang dilakukan terlihat bahwa nilai k sangat berpengaruh terhadap akurasi yang dihasilkan seperti ditunjukkan pada Gambar 2. Namun rata-rata akurasi cenderung menurun seiring dengan penambahan nilai k. Hal ini dikarenakan semakin besar nilai k maka semakin banyak tetangga yang digunakan untuk proses klasifikasi dan kemungkinan untuk terjadinya *noise* juga semakin besar ditambah lagi dengan adanya dominasi atau frekuensi kelas data latih yang tidak seimbang dari suatu kelas tertentu sehingga hasilnya data cenderung diklasifikasikan pada data kelas yang mendominasi. Pada pengujian yang dilakukan nilai akurasi maksimum cenderung terjadi saat nilai k = 1 atau k <= 5. Semakin kecil nilai k berarti semakin sedikit jumlah tetangga yang digunakan untuk proses klasifikasi data baru.

Metode *euclidean distance* digunakan untuk mencari kedekatan antar data dimana semakin kecil nilainya maka jarak antar 2 data semakin dekat. Dengan demikian ketika nilai k kecil maka hanya tetangga yang memiliki kedekatan data terbaik saja yang digunakan untuk proses klasifikasi, Hal ini sebenarnya tergantung dengan bentuk data latih yang digunakan.



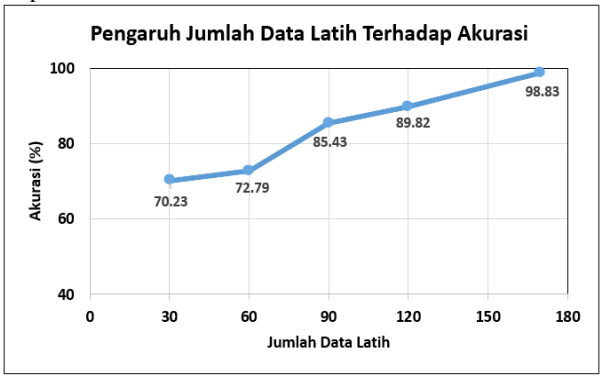
Gambar 2. Grafik Pengaruh Nilai K Terhadap Akurasi

6.2. Pengaruh Jumlah Data Uji Tetap dengan Jumlah Data Latih Berbeda

Berdasarkan hasil uji coba yang telah dilakukan, terlihat bahwa jumlah data latih sangat berpengaruh pada besar nilai akurasi yang dihasilkan terlebih jika data latih dengan kelas data seimbang seperti ditunjukkan pada Gambar 3.

Pada pengujian ini akurasi yang diambil merupakan akurasi terbaik sesuai dengan parameter k yang sudah ditentukan oleh aplikasi. Berdasarkan hasil uji coba peningkatan jumlah data latih turut disertai dengan peningkatan hasil akurasi. Hal ini dipengaruhi dikarenakan semakin banyak data latih yang digunakan maka semakin banyak data yang dibandingkan dan akan berpengaruh pada hasil akurasi.

Dalam pengujian ini data latih dalam keadaan kelas data seimbang *(Balanced Class)* artinya tidak ada kelas yang lebih mendominasi. Karena data latih yang tidak seimbang menimbulkan *noise* dalam penentuan keputusan.



Gambar 3. Pengaruh Jumlah Data Latih Terhadap Akurasi

6. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan maka didapatkan kesimpulan bahwa metode *Modified K-Nearest Neighbor* dapat diimplementasikan untuk klasifikasi *dataset* penyakit pada tanaman kedelai. Metode *Brute Force* dapat diimplementasikan untuk mendapatkan nilai k terbaik berdasarkan akurasi terbesar dari suatu proses percobaan pendahulu, sehingga tidak perlu memasukkan parameter nilai k secara manual untuk suatu pengujian data.

Rata-rata akurasi maksimum yang dihasilkan pada penelitian ini sebesar 98,83% pada saat jumlah data latih 170 data dan akurasi minimum sebesar 70,23% pada saat jumlah data latih 30 data. Secara umum penambahan nilai k akan menyebabkan penurunan akurasi.

Penelitian selanjutnya bisa dilakukan dengan hibridisasi KNN dan algoritma heuristik (Jain dan Mazumdar, 2003) seperti algoritma genetika yang terbukti efektif untuk permasalahan yang kompleks (Mahmudy, 2014). Secara fungsional sistem ini dapat dikembangkan agar dapat digunakan secara langsung oleh petani dengan mudah, melalui parameter-parameter yang mudah diketahui oleh petani.

7. DAFTAR PUSTAKA

Chandrasekhar, T. Thangavel, K. dan Elayaraja, E. 2011. *Effective Clustering Algorithms for Gene Expression Data. Department of Computer Science*. Periyar University. Tamil Nadu, India.

Han, J. dan Kamber, M. 2006. *Data Mining Concepts and Techniques Second Edition.* Morgan *Kaufmann* Publishers.

Jain, R. dan Mazumdar, J. 2003. A genetic algorithm based nearest neighbor classification to breast cancer diagnosis. *Australasian Physics & Engineering Sciences in Medicine*, vol 26, no. 6.

Larose, D. T. 2005. *Discovering Knowledge in Data: An Introduction to Data Mining*. Wiley, Chichester.

Mahmudy, W. F. (2014), *Optimisation of Integrated Multi-Period Production Planning and Scheduling Problems in Flexible Manufacturing Systems (FMS) Using Hybrid Genetic Algorithms, School of Engineering*, University of South Australia.

Marwoto. 2007. *Dukungan Pengendalian Hama Terpadu dalam Program Bangkit Kedelai*. Balai Penelitian Tanaman Kacang-kacangan dan Umbi-umbian. Malang.

Parvin, H.,, Alizadeh, H dan Minae-Bidgoli, B. 2008. *MKNN: Modification on K-Nearest Neighbor Classification.* San Francisco. USA.

Parvin, H., Hoseinali dan Minati, B. 2010. Modification on K-Nearest Neighbor Classification. Global Journal of Computer Science and Technology, vol. 10, no. 14.

Yu, C. 2010. *How KNN works* ?. Indiana *University*. USA.

Widodo, D. 1987. *Hama dan Penyakit Kedelai*. Penerbit Pustaka Buana. Bandung.