PREDIKSI PENYAKIT MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOUR DAN ALGORITMA GENETIKA UNTUK DATA BERDIMENSI TINGGI

DISEASE PREDICTION USING K-NEAREST NEIGHBOUR AND GENETIC ALGORITHM FOR HIGH DIMENSIONAL DATA

[1]Hana Rufaidha, [2]Fhira Nhita ST., MT., [3]Danang Triantoro M, S.si., MT.
Ilmu Komputasi Fakultas Informatika Universitas Telkom, Bandung
[1]hanarufaidha@gmail.com, [2]fhiranhita@telkomuniversity.ac.id, [3]danangtriantoro@telkomuniversity.ac.id

ABSTRAK

Data mining menjadi salah satu cara untuk memahami teknik-teknik tertentu dalam mengolah data, sehingga dapat diperoleh informasi yang tersembunyi pada suatu data. Dalam proses nya, pengolahan data memiliki dimensi yang tinggi sehingga sulit untuk ditangani. Curse of dimensionality atau kutukan dimensi merupakan permasalahan nyata yang terkait dengan dimensi tinggi, sehingga proses pengolahan data menjadi kurang efektif. Evolutionary Data Mining merupakan salah satu solusi yang dapat mengatasi permasalahan data berdimensi tinggi. Genetic Algorithm (GA) merupakan salah satu algoritma Eas yang sangat berguna untuk memecahkan masalah pada proses pencarian (searching) dan proses optimasi (optimization). Algoritma K-Nearest Neighbour (KNN) merupakan algoritma data mining yang dapat digunakan untuk melakukan klasifikasi pada data dan Genetic Algorithm (GA) dapat membantu memaksimalkan klasifikasi akurasi subset dari atribut. Dengan menggunakan metode tersebut, diharapkan akan menghasilkan suatu akurasi prediksi diatas 75%.

Kata kunci : data dimensi tinggi, evolutionary data mining, k-nearest neighbor, algoritma genetika

ABSTRACT

Data mining can understand certain techniques in data processing, in order to obtain the hidden information. In the data processing, the data has a high dimensionality and it's really difficult to handle. Curse of dimensionality are the real problems related to high dimensional data, so that the data processing becomes less effective. Evolutionary of Data Mining can be the one solution that can overcome the problems of high dimensional data. Evolutionary Algorithms (EAs) is a population-based meta heuristic optimization algorithm generic that can help in reducing the dimensions of data mining. Genetic Algorithm (GA) is one of the EAs very useful algorithm for solving the problem in the search process (searching) and process optimization (optimization). Algorithm K-Nearest Neighbor (KNN) is a data mining algorithm that can be used to classify the data and Genetic Algorithm (GA) can help maximize the classification accuracy of a subset of attributes. Hypothesis of this thesis is the analysis and implementation of algorithms K-Nearest Neighbor optimized with Genetic algorithms can predict the level of performance of a high-dimensional disease data above 75%.

***Keywords**: high dimensional data, evolutionary data mining, k-nearest neighbor, genetic algorithm*

1. Pendahuluan

1.1 Latar Belakang

Data mining memiliki banyak sekali manfaat dalam berbagai masalah pengolahan data, sehingga data- data yang tidak memiliki informasi penting dapat digali dan dianalisa. Data dimensi tinggi memiliki dimensi yang sangat banyak yang jumlah dimensinya mencapai ratusan bahkan ribuan dimensi, sehingga kompleksitas dari data tersebut menjadi sangat besar. Tidak menutup kemungkinan bahwa dari keseluruhan dimensi pada data, sebenarnya ada dimensi yang tidak perlu digunakan pada proses data mining. Curse of Dimensionality atau kutukan dimensi dapat membuat proses pengolahan data menjadi kurang efektif dan efisien, sehingga diperlukan teknik tertentu untuk mereduksi dimensi sehingga memiliki tingkat akurasi (performansi) yang baik [4]. Selain itu, mereduksi dimensi juga bertujuan untuk memudahkan pengolahan data yang dilakukan data mining dan membuat model klasifikasi lebih mudah untuk dipahami[12].

Penggabungan antara algoritma data mining dengan algoritma Evolutionary Algorithms (EAs) menjadi salah satu solusi untuk mengatasi permasalahan yang berkaitan dengan fenomena kutukan dimensi. EAs adalah algoritma-algoritma optimasi yang yang berbasis evolusi biologi yang ada pada dunia nyata. EAs bekerja dengan cara membangkitkan, menguji, dan berusaha memperbaiki sekumpulan kandidat solusi sampai

ditemukan satu solusi yang bisa diterima [2]. Algoritma K-Nearest Neighbour (KNN) merupakan algoritma data mining yang dapat digunakan untuk melakukan klasifikasi dan Genetic Algorithm (GA) yang dapat membantu dalam memaksimalkan akurasi klasifikasi dari atribut sehingga tingkat keakuratan (performansi) menjadi lebih optimal[10]. Penelitian dengan menggunakan Algoritma KNN dengan GA untuk data berdimensi tinggi sudah pernah dilakukan sebelumnya pada jurnal Classification of Heart Disease Using K-Nearest Neighbour and Genetic Algorithm [1].

1.2 Rumusan Masalah

Rumusan masalah dari tugas akhir ini adalah sebagai berikut :

1. Bagaimana mengimplementasikan algoritma KNN dan algoritma genetika untuk memprediksi data penyakit berdimensi tinggi?
2. Bagaimana performansi dari algoritma KNN dan GA dalam memprediksi penyakit dengan data berdimensi tinggi?

Adapun batasan masalah dari tugas akhir ini adalah sebagai berikut :

1. Data yang digunakan merupakan data beberapa penyakit, antara lain penyakit Leukimia dan Colon Tumor yang diambil dari Biomedical Dataset pada Kent Ridge [6].
2. Tidak ada penanganan outlier dan missing value pada data.
3. Perhitungan jarak data uji ke data latih pada algoritma KNN menggunakan metode Euclidean distance.
4. KKN digunakan sebagai evaluasi individu pada GA.

1.3 Tujuan

Tujuan dari tugas akhir ini adalah sebagai berikut :

1. Mengimplementasikan algoritma KNN dan algoritma genetika untuk memprediksi data penyakit berdimensi tinggi.
2. Mengetahui performansi dari algoritma KNN dan GA dalam memprediksi penyakit dengan data berdimensi tinggi.

2. Dasar Teori

2.1 *Classification*

Pada data mining, terdapat beberapa teknik dalam penggalian data, antara lain adalah klasifikasi. Teknik klasifikasi dapat didefinisikan secara detail sebagai suatu pekerjaan yang melakukan pelatihan atau pembelajaran terhadap fungsi target f yang memetakan setiap vektor (set fitur) x ke dalam satu dari jumlah label kelas y yang tersedia [13]. Tujuan dari klasifikasi antara lain adalah agar data record yang sebelumnya tidak terlihat dapat dinyatakan kelasnya seakurat mungkin.

2.1.1 K-Nearest Neighbour (KNN)

K-Nearest Neighbour (KNN) termasuk kedalam kelompok instance-based learning yang bertujuan untuk mengklasifikasikan suatu objek baru berdasarkan atribut dan pelatihan KNN digunakan dalam permasalahan aplikasi, salah satunya untuk masalah klasifikasi dan sebagai fungsi learning dan training [3]. KNN melakukan klasifikasi suatu data baru berdasarkan data pembelajaran yang memiliki jarak paling dekat data baru tersebut. Apabila diberikan suatu titik uji, akan ditemukan sejumlah K titik yang paling dekat dengan titik uji tersebut. Menentukan nilai K terbaik untuk algoritma ini tergantung pada data yang akan di proses. Pada umumnya, pemilihan nilai K yang tinggi dapat mengurangi efek noise pada klasifikasi, tetapi membuat batasan antara setiap klasifikasi menjadi lebih jauh. Tingginya tingkat sensitivitas lokal membuat pengklasifikasi KNN sangat rentan terhadap noise pada data pelatihan [11].

Pengukuran kedekatan dapat diukur berdasarkan jarak antara data pertama terhadap data lainnya. Semakin dekat jarak antara data, maka semakin besar pula kemiripannya dan begitu pula sebaliknya [9]. Ada banyak cara untuk mengukur jarak kedekatan antara data, diantara nya dengan menggunakan teknik Euclidean distance. Fungsi dari Euclidean distance antara lain adalah sebagai berikut.

$$\sqrt{\qquad\qquad\qquad\qquad\qquad\qquad} \qquad (2.2)$$

merupakan sampel atribut pertama *merupakan sampel atribut kedua dan* *merupakan jumlah total dari sampel atribut [13].*

2.2 *Genetic Algorithm* (GA)

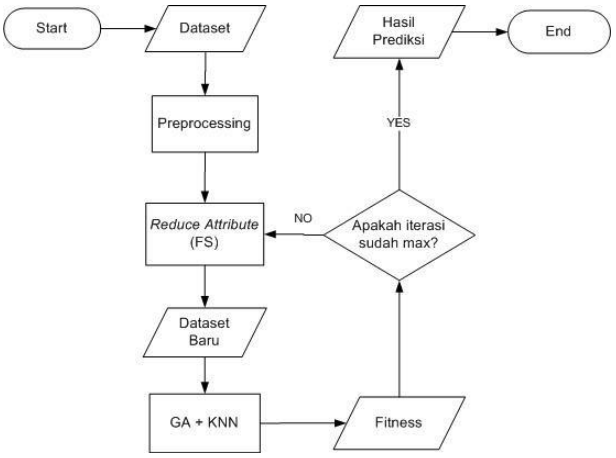
Genetic Algorithm merupakan salah satu algoritma dari EAs yang merupakan sebuah metodologi yang dibangun dengan proses evolusi komputasi. Algoritma ini didasarkan pada proses genetik yang terjadi pada makhluk hidup, dimana perkembangan generasi pada suatu populasi yang alami secara lama kelamaan akan mengikuti seleksi alam yaitu dimana yang kuat yang akan bertahan. Proses menghasilkan populasi baru berdasarkan populasi aturan sebelumnya akan berulang sampai suatu populasi P, P dapat berkembang dimana setiap aturan dalam P memenuhi nilai fitness sudah ditentukan oleh ambang batas [5]. Dengan mengikuti teori evolusi tersebut, algoritma genetik ini dapat digunakan untuk memecahkan masalah yang terjadi pada sehari-hari.

Selain itu, GA sangat berguna untuk memecahkan masalah pada proses pencarian (searching) dan proses optimasi (optimization). Oleh karena itu GA memainkan peran penting sebagai model untuk memecahkan suatu masalah [1]. Algoritma GA merupakan algoritma yang paling sederhana dibandingkan dengan algoritma EAs lainnya. GA merupakan algoritma yang mudah untuk di paralel kan dan telah digunakan untuk masalah optimasi klasifikasi lainnya. Dalam data mining, GA juga dapat digunakan mengevaluasi nilai fitness algoritma lainnya [5].

3. Pembahasan

3.1 Deskripsi Sistem

Pada tugas akhir ini, akan dirancang suatu sistem untuk menentukan prediksi suatu penyakit dengan menggunakan data berdimensi tinggi. Dataset yang digunakan merupakan data penyakit yang diambil dari Biomedical Dataset pada Kent Ridge. Berikut ini adalah skema dari perancangan sistem yang akan dibangun.



Gambar 1 Gambaran Umum Sistem

Pada gambar 3-1 terdapat beberapa tahapan dari sistem yang akan dibangun. Dataset yang akan digunakan pada sistem di *preprocessing* terlebih dahulu. Selanjutnya dataset akan direduksi dimensinya menggunakan *Feature Selection* sehingga menghasilkan dataset baru. Dataset baru akan masuk kedalam sistem GA+KNN untuk mendapatkan hasil prediksi sehingga menghasilkan *output* performansi dari hasil prediksi yang telah diperoleh.

3.2 Dataset

Data yang digunakan pada tugas akhir ini merupakan data beberapa penyakit yang memiliki atribut yang sangat banyak (berdimensi tinggi). Data ini diambil dari beberapa data penyakit yang tersedia pada *Biomedical Dataset* dari *Kent Ridge Repository*. Dataset yang ada pada *repository* tersebut merupakan *gene expression data*, *protein profiling data* dan *genomic sequence data* yang beberapa data diantaranya sudah pernah digunakan dalam berbagai jurnal ilmiah.

Data Leukimia ALL-AML merupakan salah satu dataset yang akan digunakan dalam tugas akhir ini. Didalam data tersebut, terdapat 38 sampel *data training* dan 34 sampel *data testing* dengan 7129 atribut yang memiliki dua kelas, yaitu ALL dan AML. Sedangkan data tumor usus merupakan data penyakit selanjut nya yang akan digunakan pada tugas akhir ini. Didalam data tersebut terdapat 62 jumlah sampel data dengan 2000 atribut yang memiliki kelas porsitif mengidap tumor dan negatif.

4. Pembahasan

4.1 Skenario Pengujian

Pada parameter GA terdapat nilai probabilitas crossover (Pc) dan probablitas mutasi (Pm) yang akan digunakan pada sistem. Nilai Pc yang akan digunakan yaitu 0.6 dan 0.8 dan nilai Pm yang akan digunakan yaitu 0.05 dan 0.1. Jumlah individu yang akan dievaluasi sebanyak 1000 individu. Berikut tabel dari skenario kombinasi dari Pc dan Pm yang akan digunakan pada sistem ini :

Tabel 4-1: Tabel Kombinasi Pengujian

Kombinasi	Ukuran Populasi	Pc	Pm
1	50	0.6	0.05
2	100	0.6	0.05
3	200	0.6	0.05
4	50	0.6	0.1
5	100	0.6	0.1
6	200	0.6	0.1
7	50	0.8	0.05
8	100	0.8	0.05
9	200	0.8	0.05
10	50	0.8	0.1
11	100	0.8	0.1
12	200	0.8	0.1

Berdasarkan pembagian data, maka skenario yang akan dilakukan menjadi 2 bagian yaitu :

a. Skenario 1
   Data yang akan digunakan pada skenario pertama antara lain adalah data Colon Tumor dan Leukimia. Pada kedua data tersebut dilakukan pembagian data training dan data testing menggunakan *Percentage Split* dengan rasio pembagian sebesar 70:30. Kedua data tersebut telah melalui tahap *preprocessing* terlebih dahulu.
b. Skenario 2
   Data yang akan digunakan pada skenario kedua antara lain adalah data Colon Tumor dan Leukimia. Pada kedua data tersebut dilakukan pembagian data training dan data testing menggunakan *Cross Validation* dengan metode *K-Fold* dan parameter K bernilai 3. Kedua data tersebut telah melalui tahap *preprocessing* terlebih dahulu.

4.2 Hasil dan Analisis

4.2.1 Analisis Parameter Pc dan Pm

Kombinasi pengujian data pada sistem salah satu tujuannya antara lain untuk mengetahui pengaruh dari parameter Pc dan Pm terhadap nilai perfomansi yang dihasilkan. Terdapat 1000 individu yang akan dievaluasi pada setiap kombinasi. Berikut ini adalah tabel performansi pada data Colon Tumor dan Leukemia pada saat nilai K bernilai 3.

Tabel 4-2: Tabel Performansi dari Pengaruh Pc dan Pm

Pc	Pm	Rata-rata Performansi	
		Colon Tumor	Leukemia
0.6	0.05	90.16	97.22
0.6	0.1	**93.57**	**100**
0.8	0.05	86.90	97.22
0.8	0.1	85.48	98.61

Pada Tabel 4-2, diperoleh perhitungan rata-rata performansi dari skenario 2 yang dilakukan pada data Colon Tumor dan Leukimia. Parameter K yang digunakan adalah 3. Hal ini dikarenakan hasil performansi yang diperoleh cenderung beragam, sehingga analisis untuk mengetahui pengaruh dari parameter Pc dan Pm menjadi lebih mudah.

Dapat diketahui dari Tabel 4-6 diatas, parameter Pc dan Pm yang memiliki rata-rata akurasi tertinggi pada kedua data adalah saat Pc bernilai 0.6 dan Pm bernilai 0.1. Pada saat Pc 0.8, rata-rata nilai akurasi yang dihasilkan pada kedua data berbeda. Parameter Pm 0.05 menghasilkan rata-rata yang lebih kecil pada Colon Tumor. Berbeda dengan rata-rata pada Leukemia, Pm 0.1 menghasilkan data yang lebih besar. Hal ini dapat terjadi akibat jumlah sampel data yang berbeda. Parameter Pm akan menghasilkan nilai rata-rata akurasi yang lebih tinggi pada data sampel yang ukurannya sedikit.

4.2.2 Analisis Parameter K Terhadap Performansi Sistem

Untuk mengukur seberapa baik sistem yang telah dibangun, maka data pelatihan akan digunakan sebagai data uji. Selain itu, ingin diketahui seberapa besar pengaruh dari parameter K terhadap performansi dari sistem yang telah dibangun. Berikut tabel performansi dari sistem berdasarkan parameter K yang digunakan.

Tabel 4-3: Performansi Sistem Pada Colon Tumor

Kombinasi	Skenario 1			Skenario 2		
	K = 3	K = 5	K = 7	K = 3	K = 5	K = 7
1	90.7	90.7	90.7	95.12	90.24	90.24
2	90.7	90.7	83.72	82.93	92.68	78.57
3	93.02	90.7	90.7	92.68	83.33	75.61
4	90.7	90.7	90.7	88.1	95.12	83.33
5	93.02	90.7	90.7	82.93	85.37	70.73
6	90.7	90.7	83.72	85.37	87.8	88.1
7	90.7	90.7	88.37	95.12	85.37	80.95
8	100	88.37	86.05	92.68	88.1	75.61
9	90.7	93.02	86.05	92.86	80.49	85.37
10	90.7	90.7	83.72	87.8	90.48	87.8
11	90.7	93.02	88.37	82.93	80.95	80.95
12	90.7	88.37	93.02	90.24	82.93	83.33
MEAN	**91.86**	90.7	87.98	**89.06**	86.9	81.72

Pada Tabel 4-3, terlihat bahwa performansi maksimum yang diperoleh pada Colon Tumor sebesar 91.86% dan pada Leukimia sebesar 89.06% adalah pada saat K bernilai 3. Pada tabel tersebut juga menunjukan bahwa semakin besar nilai K yang digunakan, semakin kecil nilai performansi yang diperoleh.

Tabel 4-4: Performansi Sistem Pada Leukimia

Kombinasi	Skenario 1			Skenario 2		
	K = 3	K = 5	K = 7	K = 3	K = 5	K = 7
1	98	98	94	100	97.92	91.67
2	98	96	94	97.92	93.75	93.75
3	100	98	96	93.75	97.92	95.83
4	96	96	98	93.75	97.92	93.75
5	98	98	98	95.83	95.83	91.67

6	96	98	96	89.58	97.92	87.5
7	96	98	98	95.83	95.83	93.75
8	96	98	98	97.92	91.67	93.75
9	100	98	92	93.75	95.83	95.83
10	98	96	94	97.92	95.83	89.58
11	100	98	94	97.92	91.67	89.58
12	98	96	94	95.83	93.75	93.75
MEAN	**97.83**	97.33	95.5	**95.83**	95.49	92.53

Pada Tabel 4-4, diketahui bahwa performansi maksimum yang diperoleh pada data Leukimia di kedua skenario adalah pada saat K bernilai 3. Pada skenario 1 diperoleh performansi sebesar 97.83% dan pada skenario 2 diperoleh performansi sebesar 95.83%. Pada tabel tersebut juga menunjukan bahwa semakin besar nilai K yang digunakan semakin kecil nilai performansi yang diperoleh, karena kesimpulan yang diperoleh pada kedua data sama, yaitu performansi sistem tertinggi diperoleh pada saat K=3.

Dapat disimpulkan bahwa parameter K yang paling mempengaruhi performansi sistem adalah pada saat K=3. Pada Tabel 4-7 dan Tabel 4-8, terlihat bahwa semakin besar nilai dari parameter K, maka semakin kecil hasil performansi sistem yang diperoleh. Terbukti dengan semakin besar kedekatan data, maka semakin besar peluang bahwa data tersebut berada pada label kelas yang sama dan begitu pula sebaliknya.

5. Kesimpulan

5.1 Kesimpulan

Berdasarkan hasil analisis pengujian yang telah dilakukan pada bab sebelumnya, maka diperoleh beberapa kesimpulan sebagai berikut.

1. Berdasarkan skenario pengujian data Colon Tumor dan Leukimia yang telah dilakukan, pada skenario 1 diperoleh hasil performansi yang konstan pada setiap kombinasinya. Sedangkan pada skenario 2 diperoleh hasil performansi yang bervariasi pada setiap kombinasi pengujian. Hal tersebut dapat terjadi akibat pemilihan data yang kurang pada skenario 1.
2. Pengujian data Colon Tumor dan Leukimia pada skenario 1 menghasilkan rata-rata performansi yang sama dengan nilai tertinggi pada saat K=3. Sedangkan pada skenario 2 rata-rata performansi dengan nilai tertinggi dengan nilai yang sama pada saat K=3 dan K=7. Hal tersebut dapat terjadi akibat perbedaan banyaknya jumlah data yang berbeda sehingga menghasilkan nilai K yang berbeda pada skenario pengujian tertentu.
3. Pada kombinasi pengujian data Colon Tumor dan Leukimia diperoleh rata-rata performansi tertinggi pada saat parameter Pc sebesar 0.6 dan Pm sebesar 0.1, dengan rata-rata performansi Colon Tumor sebesar 93.57% dan Leukimia sebesar 100%.
4. Pengujian sistem yang dilakukan pada data Colon Tumor dan Leukimia diperoleh performansi tertinggi pada saat K=3 dengan rata-rata performansi sebesar 97.83% pada data Colon Tumor dan 95.83% pada data Leukimia. Hal tersebut dapat terjadi akibat pengaruh dari besar kecil nya parameter K. Semakin besar nilai K yang digunakan untuk menguji performansi sistem, menghasilkan performansi yang semakin kecil.

5.2 Saran

Berdasarkan pengujian yang sudah dilakukan, ada beberapa saran yang ingin disampaikan penulis mengenai tugas akhir ini, yaitu :

1. Data penyakit yang akan digunakan sebaiknya memiliki *record* yang jumlahnya banyak sehingga sistem memiliki proses pelatihan lebih banyak sehingga dapat meningkatkan performansi data dengan lebih baik lagi.

DAFTAR PUSTAKA

[1] M. Akhil J, B.L. Deekshatulu, Priti C. "Classification of Heart Disease Using K-Nearest Neighbour and Genetic Algorithm," Procedia Technology 10 (2013) 85 – 94.

[2] Suyanto, Soft Computing "Membangun Mesin Ber-IQ Tinggi," Bandung: Informatika (2008).

[3] Rhido. A, 2006. k-Nearest Neighbor. Soft Computing Researce Group. EEPIS-ITS.

[4] Michel Verleysen, "Learning high-dimensional data," IOS Press (2003) pp. 141 – 162.

[5] Han,J., Kamber,M., dan Pei,J. "Data mining: Concepts and Techniques 3rd edition" USA : The Morgan Kaufmann series in Data Management System (2012).

[6] Li, J., 2009. http://levis.tongji.edu.cn/gzli/data/mirror-kentridge.html. [Online] [Diakses pada tanggal 28 Oktober 2015].

[7] S. M. Suyanto, Soft Computing "Membangun Mesin Ber-IQ Tinggi", Bandung: INFORMATIKA, 2008.

[8] Suyanto, Evolutionary Computation "Komputasi Berbasis Evolusi dan Genetika" Bandung : Informatika (2008).

[9] Prasetyo, E. Data Mining: "Mengolah Data Menjadi Informasi Menggunakan Matlab" Yogyakarta : ANDI (2014).K

[10] Juluisdottir,T., Keedwell,E., Corne.D., Narayanan.A. "Two phase EA/k-NN for feature Selection and Classification in Cancer Microarray Datasets" IEEE Xplore (2005).

[11] C. Gunavathi, K. Premalatha, "Performance Analysis of Genetic Algorithm with KNN and SVM for Feature Selection in Tumor Classification," International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:8, No:8 (2014).

[12] Erick Cantu-Paz, "Feature Subset Selection, Class Separability, and Genetic Algorithms," Genetic and Evolutionary Computation Conference (2004)

[13] Mai S., Tim T., Rob S., "Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients," International Journal of Information and Education Technology, Vol. 2, No. (2012)