KLASIFIKASI DALAM DATA MINING MENGGUNAKAN K-NEAREST NEIGHBOR BASED ASSOCIATION (KNNBA)

Arvica Suchiany Puspita Hati[1], Imelda Atastina[2], Kemas Rahmat Saleh Wiharja[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Seiring dengan perkembangan teknologi dalam hal pengumpulan dan penyimpanan data menyebabkan tumpukan data yang sangat banyak. Dengan adanya kumpulan data yang banyak, maka timbulah suatu kebutuhan untuk bisa memanfaatkan data tersebut. Pemanfaatan data tersebut tentunya bertujuan untuk mendapatkan informasi yang penting dari pola-pola data yang terbentuk. Proses untuk mendapatkan informasi atau pola-pola berharga dari sekumpulan data tersebutlah yang dinamakan Data mining. Klasifikasi merupakan salah satu metode dari data mining. Salah satu algoritma klasifikasi yang terkenal adalah K-nearest neigbor (KNN). Algoritma KNN sangatlah sederhana, bekerja berdasarkan jarak terpendek dari query instance ke training sample untuk menentukan KNN-nya dan mudah untuk di implementasikan. Salah satu masalah pada algoritma KNN adalah efek yang sama dari semua atribut dalam menghitung jarak antara dokumen yang baru dan dokumen yang tersedia dalam data training, mungkin beberapa dari atribut ada yang kurang penting untuk proses klasifikasi dan beberapa atribut lebih penting . Dalam penelitian ini akan menggunakan algoritma K-Nearest Neighbor Based Association (KNNBA) untuk mengatasi kekurangan dari KNN, dimana pada KNNBA tiap-tiap atribut diberikan bobot yang berbeda dengan menggunakan metode association rules. Pada tugas akhir menggunakan enam dataset yang memiliki karakteristik yang berbeda-beda. Dari hasil pengujian dan analisis didapat bahwa association dapat meningkatkan akurasi KNN dengan menggunakan parameter minimum support dan minimum confidence yang sesuai dengan jenis data atributnya

Kata Kunci : Data mining, klasifikasi, K-nearest neigbor, K-Nearest Neighbor Based Association association rules, support, confidence

Abstract

The development of technology for the collection and storage of data causes the stack of data very much. Given that many data sets, creating a need to be able to utilize the data. Utilization data is of course intended to get important information from the data patterns are formed. The process to obtain information or patterns from a collection of valuable data is called data mining. Classification is one method of data mining. One well-known classification algorithms are K-nearest neigbor (kNN). KNN algorithm is very simple, works based on the shortest distance from the query instance to the training sample to determine its kNN and easy to implement. One of the problems in the kNN algorithm is the same effect of all attributes in calculating the distance between new documents and documents that are available in training data, there are probably some of the attributes that are less important to the process of classification and some attributes are more important. In this research we will use K-Nearest Neighbor Algorithm Based Association (KNNBA) to overcome the shortcomings of the kNN, where the KNNBA of each attribute are given different weights by using the method of association rules.In the final project using six datasets that have different characteristics. From the results of testing and analysis shows that the association can improve the accuracy of kNN by using the parameters of minimum support and minimum confidence appropriate to the type of data attributes.

Keywords : Data mining, classification, K-nearest neigbor, K-Nearest Neighbor Based Association, association rules, support, confidence

1 Pendahuluan

1.1 Latar Belakang Masalah

Seiring dengan perkembangan teknologi dalam hal pengumpulan dan penyimpanan data menyebabkan tumpukan data yang sangat banyak. Dengan adanya kumpulan data yang banyak, maka timbulah suatu kebutuhan untuk bisa memanfaatkan data tersebut. Pemanfaatan data tersebut tentunya bertujuan untuk mendapatkan informasi yang penting dari pola-pola data yang terbentuk. Proses untuk mendapatkan informasi atau pola-pola berharga dari sekumpulan data tersebutlah yang dinamakan *Data mining.*

Data mining adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar [9]. Klasifikasi merupakan salah satu metode dari *data mining* . Klasifikasi adalah metode prediktif yang melakukan pembelajaran terhadap data-data yang sudah ada sehingga menghasilkan suatu model yang digunakan untuk memprediksi data-data baru. Salah satu algoritma klasifikasi yang terkenal adalah K-nearest neigbor (KNN).

K-Nearest Neighbor (KNN) adalah suatu metode yang menggunakan algoritma *supervised* dimana hasil dari *query instance* yang baru diklasifikan berdasarkan mayoritas dari kategori pada KNN. Tujuan dari algoritma ini adalah mengklasifikasikan obyek baru bedasarkan atribut dan *training sample*. Algoritma KNN sangatlah sederhana, bekerja berdasarkan jarak terpendek dari *query instance* ke *training sample* untuk menentukan KNN-nya dan mudah untuk di implementasikan. KNN memiliki kemampuan kerja yang rendah ketika *training dataset* besar [9]. Salah satu masalah pada algoritma ini adalah bobot yang sama dari semua atribut dalam menghitung jarak antara data *testing* dan data *training* , bagaimana pun, mungkin dari semua atribut ada beberapa atribut yang kurang penting untuk proses klasifikasi dan ada beberapa atribut yang lebih penting untuk proses klasifikasi[12]. Sehingga tidak jelas jarak mana yang harus digunakan dan atribut mana yang harus digunakan untuk mendapatkan hasil terbaik[8]. Hal ini dapat menyesatkan proses klasifikasi dan dapat menurunkan akurasi dari klasifikasi [12].

Pendekatan yang banyak dilakukan untuk mengatasi masalah ini adalah dengan memberi bobot yang berbeda pada tiap-tiap atribut ketika mengukur jarak dua record. Pembobotan berguna untuk menentukan jarak antar atribut tetangga dengan record baru berdasarkan similarity. Dalam penelitian ini akan menggunakan algoritma *K-Nearest Neighbor Based Association* (KNNBA), dimana pada KNNBA tiap-tiap atribut diberikan bobot yang berbeda dengan menggunakan metode *association rules* [12]. *Association* termasuk metode *data mining* yang deskriptif dimana pada metode ini dilakukan penemuan pola-pola dari sekumpulan data yang ada, sehingga dengan pola tersebut dapat menggambarkan atau mendeskripsikan data tersebut. Dimana pada penelitian ini *association rules* berguna untuk memprediksi label kelas dari tiap atribut [12].

1.2 Perumusan Masalah

Tugas Akhir ini mempunyai perumusan masalah sebagai berikut :

1. Bagaimana menerapkan *K-Nearest Neighbor Based Association* (KNNBA) untuk klasifikasi data.
2. Bagaimana menganalisis performansi dan pengaruh parameter-parameter dari KNNBA seperti ukuran *neighbor* , *minimum support*, *minimum confidence* dan ukuran *training set* terhadap hasil prediksi system dibandingkan dengan KNN.

Hipotesa awal dari penelitian ini adalah *K-Nearest Neighbor Based Association* (KNNBA) menghasilkan akurasi yang lebih baik dibandingkan KNN .

Adapun batasan masalah dari tugas akhir ini adalah :

1. *Dataset* yang digunakan adalah *dataset* sintetik yang dihasilkan oleh data generator ataupun *dataset* yang berasal dari *UCI Machine Learning Repository*. *Dataset* yang dipakai berjumlah enam *dataset*.
2. Tidak menangani tahap *pre-processing*.
3. Inputan yang diterima berupa tabel yang telah tersedia .

1.3 Tujuan

Tujuan dari Tugas Akhir ini adalah :

1. Menerapkan *association rules* dalam algoritma KNNBA untuk klasifikasi data dan menganalisis performansinya.
2. Menganalisis performansi dan pengaruh parameter-parameter dari KNNBA seperti ukuran *neighbor* , *minimum support*, *minimum confidence* dan ukuran *training set* terhadap hasil prediksi sistem dibandingkan dengan KNN dengan parameter pengujian akurasi , *precision*, *recall* dan *F-measure*.

1.4 Metodologi Penyelesaian Masalah

Metodologi yang digunakan dalam memecahkan masalah di atas adalah dengan menggunakan langkah-langkah berikut:

1. Studi Literatur
   Pencarian referensi dan sumber-sumber yang berhubungan dengan *Data mining*, Algoritma KNN dan algoritma KNNBA dalam menyelesaikan tugas akhir ini.
2. Pengumpulan data
   Mengumpulkan dataset yang diperlukan sebagai *training set* dan *test set*.
3. Perancangan Sistem
   Analisis dan perancangan perangkat lunak yang akan dibangun dengan menggunakan metode berorientasi objek.
4. Implementasi Sistem
   Tahap pembangunan perangkat lunak dengan melakukan pembangunan terhadap pembuatan algoritma, melakukan implementasi algoritma pada *classificasion engine*, serta pembuatan antar muka untuk aplikasi *client*.
5. Pengujian Sistem dan Pengambilan kesimpulan
   Melakukan Pengujian dan analisis terhadap perangkat lunak yang dibangun. Melakukan analisis dari implementasi sistem dan Pengujian

hasil. Menganalisa kebenaran klasifikasi dan mengambil kesimpulan berdasarkan parameter akurasi yang dihasilkan Pengambilan Kesimpulan dan penyusunan laporan Tugas Akhir.

6. Penyusunan laporan tugas akhir
   Pembuatan laporan tugas akhir yang mendokumentasikan tahap-tahap kegiatan dan hasil dalam tugas akhir ini.

1.5 Sistematika Penulisan

Tugas akhir ini disusun dengan sistematika penulisan sebagai berikut:

BAB I **PENDAHULUAN**
Berisi pemaparan mengenai latar belakang permasalahan, tujuan yang ingin dicapai dengan adanya penelitian ini, perumusan masalah, batasan masalah, metodologi tugas akhir, dan sistematika penulisan.

BAB II **LANDASAN TEORI**
Berisi uraian mengenai landasan teori yang akan digunakan, meliputi teori tentang algoritma *clustKNN* dan teori-teori lain yang berkaitan dengan penelitian tugas akhir ini

BAB III **ANALISIS DAN PERANCANGAN SISTEM**
Berisi tentang analisa dan perancangan terhadap *recommender system* yang akan dibangun.

BAB IV **ANALISIS DAN PENGUJIAN SISTEM**
Berisi implementasi dari hasil analisa dan perancangan sistem yang dibuat, serta Pengujian sistem.

BAB V **KESIMPULAN DAN SARAN**
Berisi kesimpulan dan saran-saran untuk pengembangan lebih lanjut terhadap hasil penelitian ini.

5 Kesimpulan dan Saran

5.1 Kesimpulan

Berdasarkan hasil analisis terhadap pengujian maka dapat ditarik kesimpulan sebagai berikut:

1. Akurasi menjadi optimal saat atribut yang diberi bobot merupakan atribut yang dapat mewakili dataset atau atribut tersebut penting untuk proses klasifikasi.
2. Untuk atribut bertipe numerik parameter minimum *confidence* tidak berpengaruh.
3. Secara umum KNNBA memiliki akurasi lebih tinggi dibanding KNN saat K=1 dan jumlah data *training* = 70%.
4. Akurasi KNNBA akan semakin tinggi jika jumlah atribut dan jumlah kelas dari data kecil.

5.2 Saran

1. *K-Nearest Neighbor Based Association* dapat dikembangkan sehingga bisa menangani berbagai tipe data yang diinputkan, termasuk menangani data yang memiliki *missing value.*
2. Proses *association rules* dengan menggunakan algoritma *association* yang lain, seperti algoritma Apriori atau penggunaan *tree* sehingga dapat menghasilkan akurasi yang lebih optimal.

Daftar Pustaka

[1] A. Asuncion dan D.J. Newman.2006. "*UCI Machine Learning Repository*", Irvine, CA: University of California, School of Information and Computer Science.

[2] Agusta Agus,"*The Elements of Statistical Learning: Data Mining, Inference, and Prediction*",Springer.

[3] Bing Liu, Wynne Hsu dan Yiming Ma,1998,*Integrating Classification and Association Rule Mining*, Singapore: Department of Information Systems and Computer Science National University of Singapore.

[4] Bing Liu, Yiming Ma, dan Ching Kian Wong,1999,*Improving an Association Rule Based Classifier*, Singapore: School of Computing National University of Singapore.

[5] Carl, 2009, *Precision, Recall, and the F-Measure*, Available: measure1442

[6] Cover dan Hart, 1967,"*Nearest neighbor pattern classification*", *IEEE Transactions on Information Theory*.

[7] D. T. LAROSE.2005."*Discovering knowledge in data: an introduction to data mining*", New Jersey: John Wiley & Sons.

[8] Evan," *K-Nearest Neigbor(KNN)"*,Evan's blog.

[9] Ian H.Witten dan Eibe Frank.2005."*Data mining*",2nd ed.Amsterdam: Morgan Kaufmann Publishers.

[10] Iperpin, *Recall & Precision*, Available: http://iperpin.wordpress.com/2008/03/27/*recall-precision*/

[11] J. Han dan M. Kamber, 2006."*Data mining Concepts and Techniques*", 2nd ed. Amsterdam: Morgan Kaufmann Publishers.

[12] Mehdi Moradian, Ahmad Baraani.2009. *KNNBA: K-Nearest Neigbor Based Association Algorithm*. Isfahan: *Department of Computer Engineering, University of Isfahan.*

[13] Microsoft, *Partitioning Data into Training and Testing Sets (Analysis Services – Data Mining)*, Available: http://technet.microsoft.com/en-us/library/bb895173.aspx.

[14] Pang-Ning Tan, M. Steinbach dan V. Kumar. 2006. *Intoduction to Data mining* .

[15] Santosa, Budi.2007. Data mining Teknik Pemanfaatan Data Untuk Keperluan Bisnis. Graha Ilmu.