View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Open Library

KLASIFIKASI DATA DENGAN MENGGUNAKAN ALGORITMA C4.5 DAN TAN (TREE AUGMENTED NAIVE BAYES)
DATA CLASSIFCATION USING C4.5 AND TAN (TREE AUGMENTED NAIVE BAYES) ALGORITHMS

Anggi Fitrining Tyas[1], Imelda Atastina[2], Adiwijaya[3]

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

---

Abstrak

Data mining merupakan serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui. Ada beberapa task dalam data mining, salah satunya adalah klasifikasi. Dalam Tugas Akhir ini akan digunakan metode klasifikasi C4.5 dan salah satu metode dalam Bayesian Network, yaitu Tree Augmented Naive Bayes (TAN).

Algoritma C4.5 menggambarkan suatu distribusi joint probability dari sebuah set atribut. TAN merupakan graf asiklik berarah yang node-nodenya merepresentasikan variable pada data set sedangkan busur-busurnya (arc) merepresentasikan relasi ketergantungan diantara variable tersebut, dan algoritma C4.5+TAN adalah penggabungan kedua fungsionalitas di atas.

Tugas Akhir ini bertujuan untuk menganalisis performansi waktu klasifikasi dan akurasi, serta bentuk pohon keputusan dari gabungan algoritma C4.5 dan TAN Classifier yang pemodelannya dibangun menggunakan algoritma C4.5 dan conditional independence test.

Kata Kunci : Kata Kunci: C4.5, Conditional independence test, TAN, data mining, klasifikasi.

Abstract

Data mining is a process to find out the potential of information implicitly from database which unknown identifier before. One of many tasks in data mining that would be the subject of this final project is classification. The subjects of this final project are C4.5 Tree Augmented Naive Bayes (TAN) classifier.

TAN is a directed acyclic graph whose nodes represent variables and arcs represent statistical dependence relations among the variables and local probability distributions for each variable given values of its parents.

This final project analyzes the performance and accuracy of Naïve Bayes classifier and Tree Augmented Naïve Bayes classifier as classification technique of BN which build using conditional independence test based algorithms.

Keywords : Keywords: C4.5, Conditional independence test, TAN, data mining, klasifikasi.

BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Dalam beberapa dekade terakhir, teknologi informasi dan basis data telah berkembang dari sistem pemrosesan file primitif menjadi sistem basis data yang canggih. Namun, seiring dengan berjalannya waktu, jumlah data yang sangat besar yang dikumpulkan dan disimpan dalam gudang penyimpanan data sering kali tidak dipakai oleh para *analist* dalam membuat karena adanya kesulitan dalam mengekstrak informasi dari data yang jumlahnya sangat besar. Akibatnya, yang dibuat hanya berdasarkan intuisi bukan berdasarkan informasi dari data yang ada. Oleh karena itu, pembuat membutuhkan *tool* untuk mengekstrak pengetahuan berharga dari data yang sangat besar. Salah satu *tool* tersebut yaitu data mining yang dapat menganalisis data dan menemukan pola data yang penting untuk pengambilan .

Beberapa *task data mining* yang ada yaitu klasifikasi, regresi, asosiasi, klasterisasi, dan *anomaly detection*. Dalam tugas akhir ini dilakukan penelitian tentang klasifikasi. Klasifikasi merupakan proses data mining bersifat prediksi. Klasifikasi memiliki tujuan akhir membentuk pola sederhana/model berupa kelas dari distribusi data input. Model yang ditemukan dapat berupa aturan *"if-then" decision tree*, formula matematis atau *neural network, genetic algorithm, fuzzy, case-based reasoning, k-nearest neighbor*, dan *bayesian*.

Teknik klasifikasi yang digunakan pada pengerjaan tugas akhir ini adalah teknik *decision tree* algoritma C4.5. Yang menjadi perhatian yaitu akurasi *decision tree* pada algoritma C4.5 yang dihasilkan dari data input. Untuk jumlah data yang sangat banyak, pada algoritma C4.4 sering terjadi overlap, dan terdapat kesulitan dalam mendesain pohon keputusan yang optimal. Proses untuk membentuk *decision tree* pada jumlah *record* data yang banyak dapat diminimalkan dengan pengoptimalan algoritma *Tree Augmented Naive Bayes (TAN)*, sebab dalam pembentukan graf TAN hanya mengeksekusi atribut sebanyak satu kali dan menghasilkan satu level *tree*, berbeda dengan C4.5 yang dapat mengakses atribut beberapa kali dalam pembentukan *tree*. Selain itu, *Tree Augmented Naive Bayes (TAN) classifier* merupakan salah satu tipe *Bayesian Belief Network* (BBN) dan merupakan pengembangan dari Naive *Bayes classifier* yang memiliki node-node yang dapat memiliki keterkaitan satu sama lain, sehingga proses pembentukan graf semakin sederhana.

Tujuan dari tugas akhir ini adalah membandingkan proses pembentukan *decision tree* dengan menggunakan algoritma C4.5 dengan pembentukan *decision tree* menggunakan

gabungan algoritma TAN dan C4.5, dari segi waktu proses pembentukan graf yang dibutuhkan (optimalisasi) dan bentuk graf serta nilai akurasi (ketepatan rule yang didapat) untuk data uji.

1.2 Tujuan Penulisan

Adapun tujuan tugas akhir ini adalah :

a. Merancang dan membangun aplikasi pembentukan graf dengan menggunakan kombinasi antara algpritma C4.5 dengan TAN.
b. Menganalisa tingkat optimalisasi pembentukan graf berdasarkan waktu yang dibutuhkan, bentuk graf yang dihasilkan dan akurasi graf (ketepatan pembentukan *rule*) dengan menggunakan algoritma C4.5 dan kombinasi C4.5 dan TAN.
c. Mengevaluasi graf akhir yang terbentuk dari kedua algoritma yang digunakan.

1.3 Perumusan Masalah

Adapun rumusan masalah dalam tugas akhir ini adalah :

a. Bagaimana waktu yang dibutuhkan untuk pembentukan graf serta graf akhir yang dibentuk oleh algoritma C4.5.
b. Bagaimana waktu yang dibutuhkan untuk pembentukan graf serta graf akhir yang terbentuk dari gabungan algoritma C4.5 dan TAN.
c. Bagaimana pengaruh penggunaan ukuran pengambilan sampel data untuk proses pembentukan graf pada gabungan algoritma TAN dan C4.5 dalam akurasi dan waktu pembentukan graf.

1.4 Batasan Masalah

Dalam penulisan tugas akhir ini, ruang lingkup pembahasan masalah hanya dibatasi pada :

a. Tidak menangani proses pre-processing data.
b. Data latih dan data uji yang akan digunakan adalah data sintetik yang dihasilkan oleh generator atau yang berasal dari UCI *Machine Learning Repository*.
c. Data yang digunakan merupakan data kategorikal.

1.5 Metodologi Penyelesaian Masalah

Metode yang digunakan dalam menyelesaiakan tugas akhir ini adalah :

a. Studi literature
   - Pencarian referensi

Mencari referensi dan sumber-sumber lain yang berhubungan dengan masalah *data mining*, klasifikasi, *bayesian network*, *C4.5, TAN*, dan *CI Test Based Algorithms.*

- Pendalaman materi
  Mempelajari dan memahami materi yang berhubungan dengan tugas akhir ini.

b. Mempelajari konsep tentang *data mining*, klasifikasi C4.5 dan TAN.

c. Melakukan implementasi perancangan perangkat lunak sesuai dengan tujuan yang telah disampaikan.

d. Melakukan pengujian perangkat lunak.

e. Menganalisis hasil klasifikasi berdasarkan akurasi dan waktu pembentukan graf, serta bentuk akhir dari graf klasifikasi.

f. Penyusunan Laporan Tugas Akhir dan pengambilan kesimpulan akhir.

BAB V
KESIMPULAN DAN SARAN

4.1 Kesimpulan

Berdasarkan pengujian dan analisa maka dapat diambil beberapa kesimpulan :

1. Secara umum Model klasifikasi pada algoritma C4.5+TAN pada pengujian data *training* menghasilkan nilai akurasi lebih rendah di bandingkan dengan algoritma C4.5 .

2. Model klasifikasi pada algoritma C4.5+TAN pada pengujian data testing menghasilkan nilai akurasi lebih tinggi di bandingkan dengan algoritma C4.5 untuk data yang bersih dari outlier.

3. Dalam pembangunan model klasifikasi menggunakan algoritma C4.5 dan *CI Test*, algoritm C4.5+TAN *classifier* membutuhkan waktu yang lebih besar daripada C4.5 dengan algoritm C4.5+TAN membutuhkan waktu sekitar 2 hingga ratusan kali lipat lebih lambat dari C4.5.

4. Penambahan jumlah *record* pada data *training* sebanyak 2 sampai 3 kali lipat sangat berpengaruh terhadap nilai akurasi. Pada algoritm C4.5+TAN penambahan jumlah *record* pada data *training* dapat menaikkan nilai akurasi.

5. Bentuk graf pada algoritma C4.5+TAN lebih sederhana dibandingkan dengan algoritma C4.5, sebab pada algoritma C4.5+TAN antar simpul atribut dapat saling terhubung dan memiliki ketergantungan.

4.2 Saran

Berikut ini hal-hal yang disarankan penulis untuk penelitian selanjutnya :

1. Dapat dikembangkan dengan pembangunan model klasifikasi menggunakan jenis data lain, misalnya jenis data *imbalance class.*

2. Dapat dikembangkan pada tipe model *bayesian network* yang lain misalnya menggunakan GBN.

3. Dapat juga dikembangkan dengan menggunakan tipe data yang bersifat *continuous.*

DAFTAR PUSTAKA

[1]. Agustina Ratna Puspitasari, 2005, "Klasifikasi Pada Data Mining Menggunakan Naive Bayesian Classifier", STT Telkom Bandung.

[2]. Andrew W. Moore, "*Decision trees*", Carnegie Mellon University. http://www.cs.cmu.edu/~awm. Diakses tanggal 11 Desember 2009.

[3]. Baesens, B., M. Egmont Petersen., R. Castelo., J. Vanthienen. "Learning Bayesian Network Classifiers for Credit Scoring using Markov Chain Monte Carlo Search". K.U.Leuven Dept. of Applied Economic Sciences Naamsestraat, Leuven, Belgium. *www.cs.uu.nl/research/techreps/repo/CS-2001/2001-58.pdf*. Diakses tanggal 24 Februari 2010.

[4]. Charles River Analytics, Inc, 2004, "About Bayesian Belief Networks", https://www.cra.com/pdf/BNetBuilderBackground.pdf. Diakses tanggal 23 Desember 2009.

[5]. Cheng, Jie, dkk, "An Algorithms for Bayesian Belief Network Construction from Data". School of Information and Software Engineering University Ulster. Northern Ireland.

[6]. Cheng, Jie, Russel Greiner, "Learning Bayesian Belief Network Classifiers: Algorithms and System", Department of Computer Science, University of Alberta Edmonton, Canada. www.ee.bgu.ac.il/~boaz/LPGM/ChengGreinerA101Algorithms.pdf. Diakses tanggal 30 Mei 2010.

[7]. Chia-Ping Chen, "Entropy and Mutual Information Notes on Information Theory", http://www.slpl.cse.nsysu.edu.tw/cpchen/courses/ita/entropy.pdf. Diakses tanggal 11 Desember 2009.

[8]. Heckerman, David, 1995, "A Tutorial on Learning With Bayesian Networks", Advanced Technology Division. Microsoft Corporation.

[9]. Jiang, Liangxiao, Harry Zhang, Jiang Su, "Learning *Tree* Augmented Naïve Bayes for Ranking", Department of Computer Science, China University of Geosciences. Wuhan, China. www.ai.mit.edu/projects/jmlr/papers/volume3/ling02a/top.pdf . Diakses tanggal 7 Maret 2010.

[10]. Jiawei Han, Micheline Kamber, 2001, "Data Mining : Concepts and Techniques", Simon Fraser University.

[11]. Pang-Ning Tan, Vipin Kumar, Michael Steinbach, 2004, "Introduction to Data Mining", Michigan State University, University of Minnesota.