STUDI KLASIFIKASI DENGAN BAYESIAN BELIEF NETWORKS MENGGUNAKAN NAIVE BAYES CLASSIFIER DAN TREE AUGMENTED NAIVE BAYES CLASSIFIER

Lasnita Y. Dahlia[1], Ririn Dwi Agustin[2], Moch Arif Bijaksana[3]

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

Abstrak

Data Mining merupakan ekstraksi informasi potensial yang terkandung secara implisit pada database. Salah satu task pada data mining yang menjadi pokok perhatian dalam Tugas Akhir ini adalah klasifikasi, khususnya teknik bayesian yang sedang berkembang yaitu Bayesian Belief Networks (BBN).

Bayesian Belief Networks (BBN) merupakan graf asiklik berarah yang simpul-simpulnya mewakili variabel-variabel pada dataset dan busur-busurnya mewakili relasi ketergantungan antar variabel dan distribusi probabilitas lokal untuk masing-masing variabel yang diberikan oleh orang tuanya. Tugas Akhir ini menganalisis performansi Naïve Bayes classifier dan Tree Augmented Naïve Bayes (TAN) classifier sebagai teknik klasifikasi BBN yang menggunakan restricted structure learning serta mengimplementasikannya untuk menyelesaikan persoalan klasifikasi dalam data mining.

Hasilnya, TAN classifier menunjukkan performansi yang lebih baik daripada Naive Bayes classifier dalam hal akurasi walaupun dari segi kecepatan pembangunan model klasifikasi membutuhkan waktu yang lebih lama.

Kata Kunci : Bayesian Belief Networks, TAN, Naive Bayes, classifier, klasifikasi,

Abstract

Data mining is an extraction of potential information implicitly from a database. One of many tasks in data mining that would be the subject of this final project is classification, especially Bayesian Belief Networks (BBN).

Bayesian Belief Networks (BBN) 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 of Naïve Bayes classifier and Tree Augmented Naïve Bayes classifier as classification technique of BBN which use restricted structure learning and implement these classifiers to solve classification problems in data mining.

As the result, it had been proved that TAN classifier performance better than Naïve Bayes classifier in accuracy although for construct classification model need longer time.

Keywords : Bayesian Belief Networks, TAN, Naive Bayes, classifier,

1. Pendahuluan

1.1 Latar belakang

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 keputusan karena adanya kesulitan dalam mengekstrak informasi dari data yang jumlahnya sangat besar. Akibatnya, keputusan yang dibuat hanya berdasarkan intuisi bukan berdasarkan informasi dari data yang ada. Oleh karena itu, pembuat keputusan 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 keputusan.

Beberapa *task data mining* yaitu klasifikasi, regresi, asosiasi, klasterisasi, dan *anomaly detection.* Dalam Tugas Akhir ini, akan dilakukan penelitian tentang klasifikasi. Klasifikasi adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memprediksi kelas dari suatu objek yang labelnya tidak diketahui sebelumnya. Model yang ditemukan dapat berupa aturan "jika-maka", *decision tree*, formula matematis atau *neural network, genetic algorithm, fuzzy, case-based reasoning, k-nearest neighbor,* dan *bayesian*. Salah satu teknik klasifikasi *bayesian* yang sedang berkembang yaitu *Bayesian Belief Networks (BBN).*

BBN merupakan graf asiklik berarah yang simpul-simpulnya mewakili variabel-variabel pada *dataset* dan panah-panahnya mewakili relasi ketergantungan antar variabel dan distribusi probabilitas lokal untuk masing-masing variabel yang diberikan oleh orang tuanya. Salah satu keunggulan BBN yaitu *user* dapat mengerti ketergantungan langsung dan hubungan kausal antar variabel dengan mudah.

BBN memiliki lima tipe *classifier*, yaitu : *Naïve Bayes, Tree Augmented Naïve Bayes (TAN), Bayesian network Augmented Naïve Bayes (BAN), Bayesian multi-nets*, dan *General Bayesian networks (GBN)*. Pada Tugas Akhir ini, penulis akan meneliti *Naïve Bayes classifier* dan *Tree Augmented Naïve Bayes (TAN) classifier*. TAN *classifier* merupakan pengembangan *Naive Bayes classifier* dan kedua *classifier* tersebut menggunakan *restricted structure learning*.

1.2 Perumusan masalah

Masalah yang menjadi acuan dalam pembuatan Tugas Akhir ini adalah :

1. Bagaimana membangun model TAN *classifier* berdasarkan algoritma *Forest Augmented Naive Bayes*.
2. Bagaimana kinerja *Naive Bayes classifier* dan TAN *classifier* dalam proses klasifikasi.
3. Bagaimana mengimplementasikan ke dalam perangkat lunak model klasifikasi dengan teknik BBN, khususnya *Naive Bayes classifier* dan TAN *classifier*.

Adapun batasan-batasan dalam Tugas Akhir ini adalah sebagai berikut :

1. Tidak membahas *data mining* secara keseluruhan, hanya membahas klasifikasi dengan BBN menggunakan *Naïve Bayes classifier* dan TAN *classifier*.
2. Inputan data yang dimasukkan ke dalam sistem adalah data yang berupa data *record*.
3. Tidak menangani *data preprocessing*.
4. Data yang digunakan untuk *training* dan *testing* adalah data yang sudah dalam bentuk data diskret dan tidak ada *missing value*.
5. *Dataset* yang akan digunakan sudah tersedia dalam bentuk *tabel* dengan variabel *class* berada pada urutan paling akhir.

1.3 Tujuan

Tujuan yang ingin dicapai pada Tugas Akhir ini adalah :

1. Membuat model klasifikasi *Naïve Bayes classifier* dan TAN *classifier* melalui *learning* pada *dataset*.
2. Mengimplementasikan BBN dan menganalisis perilaku *Naïve Bayes classifier* dan TAN *classifier* terhadap akurasi dan waktu pengklasifikasian data.
3. Membandingkan kinerja *Naive Bayes classifier* dan TAN *classifier*.

1.4 Metodologi penyelesaian masalah

Metodologi yang akan digunakan untuk menyelesaikan Tugas Akhir ini adalah:

1. Studi literatur
   a. Pencarian referensi
      Mencari referensi dan sumber-sumber lain yang berhubungan dengan klasifikasi data, BBN, dan hal-hal lain yang berkaitan dengan Tugas Akhir ini.
   b. Pendalaman materi
      Mempelajari dan memahami materi konsep BBN serta proses pembangunan model klasifikasi dengan BBN.
2. Mempelajari konsep BBN yang akan digunakan dalam implementasi perangkat lunak.
3. Melakukan implementasi perancangan perangkat lunak.
4. Melakukan pengujian perangkat lunak dengan memasukkan data diskret, kemudian mencatat hasil keluaran program.
5. Menganalisis hasil klasifikasi dengan BBN.
6. Pengambilan kesimpulan dan penyusunan laporan Tugas Akhir.

5. Penutup

5.1 Kesimpulan

1. TAN *classifier* menunjukkan performansi yang lebih baik daripada *Naive Bayes classifier* dalam hal nilai akurasi. Penerapan ketergantungan antar atribut pada TAN *classifier* mampu meningkatkan nilai akurasi lebih dari 5%.
2. Dalam pembangunan model klasifikasi, TAN *classifier* membutuhkan waktu rata-rata yang lebih lama sepuluh kali daripada *Naive Bayes.*
3. Faktor lain yang mempengaruhi waktu klasifikasi yaitu jumlah *record,* jumlah *class,* dan jumlah atribut. Semakin besar jumlah *record*, jumlah *class*, dan jumlah atribut, maka semakin lama waktu klasifikasi.
4. Peningkatan jumlah baris *data testing* tidak menjamin akan meningkatkan nilai akurasi.
5. Nilai akurasi dipengaruhi oleh *data training* yang digunakan untuk membangun model klasifikasi. Semakin merata persebaran data pada *data training*, maka tingkat kebenaran model klasifikasi yang dihasilkan semakin tinggi, dan nilai akurasi semakin meningkat.

5.2 Saran

1. Implementasi BBN dapat dikembangkan dengan implementasi *classifier* yang menggunakan *unrestricted structure learning*.
2. Pengembangan dilakukan sehingga dapat menangani atribut yang bertipe *continuous*.

Daftar Pustaka

[1]	Agustina Ratna Puspitasari, 2005, "Klasifikasi Pada Data Mining Menggunakan Naive Bayesian Classifier", STT Telkom Bandung.
[2]	Bart Baesens, dkk, 2002, "Bayesian Network Classifiers for Identifying the Slope of The Customer Lifecycle of Long-Life Customers", www.feb.ugent.be/fac/research/WP/Papers/wp_02_154.pdf, didownload pada tanggal 26 Juli 2007.
[3]	"Bayesian Book" , http://www.cs.colorado.edu/~grudic/teaching/CSCI4202_2004/Bayesian_Book.pdf, didownload pada tanggal 24 Januari 2007
[4]	Charles River Analytics, Inc, 2004, "About Bayesian Belief Networks", https://www.cra.com/pdf/BNetBuilderBackground.pdf, didownload pada tanggal 22 Januari 2007.
[5]	Chia-Ping Chen, "Entropy and Mutual Information Notes on Information Theory", http://www. slpl.cse.nsysu.edu.tw/cpchen/courses/ita/l1_entropy.pdf, didownload pada tanggal 17 Juli 2007.
[6]	Harry Zhang, Liangxiao Jiang, Jiang Su, "Augmenting Naïve Bayes for Ranking", www.ai.mit.edu/projects/jmlr/papers/volume3/ling02a/top.pdf, didownload pada tanggal 09 Juli 2007.
[7]	Jiawei Han, Micheline Kamber, 2001, "Data Mining : Concepts and Techniques", Simon Fraser University.
[8]	Jie Cheng, Russel Greiner, "Learning Bayesian Belief Network Classifiers: Algorithms and System", www.ee.bgu.ac.il/~boaz/LPGM/ChengGreinerA101Algorithms.pdf, didownload pada tanggal 24 Januari 2007.
[9]	Liangxiao Jiang, Harry Zhang, Zhihua Cai, Jiang Su, "Learning Tree Augmented Naive Bayes for Ranking", China University, University of New Brunswick, didownload pada tanggal 03 Maret 2007.
[10]	M. Agus J. Alam, 2002, "Belajar Sendiri Borland Delphi 6.0", Jakarta.
[11]	Pang-Ning Tan, Vipin Kumar, Michael Steinbach, 2004, "Introduction to Data Mining", Michigan State University, University of Minnesota.
[12]	Rinaldi Munir, 2003, "Matematika Diskrit", Bandung.
[13]	Roger S. Pressman, 1997, "Software Engineering : A Practitioner's Approach Fourth Edition", New York.
[14]	"Teori Informasi", http://id.wikipedia.org/wiki/Teori_informasi, didownload pada tanggal 25 Juni 2007.