ANALISIS DAN IMPLEMENTASI ALGORITMA CS4 (CASCADING AND SHARING FOR ENSEMBLE OF DECISION TREES) PADA DATA BIOMEDICAL MULTI-ATTRIBUT ANALYSIS AND IMPLEMENTATION CS4 (CASCADING AND SHARING FOR ENSEMBLE OF DECISION TREES) ALGORITHM IN MULTI-ATTRIBUTE BIOMEDIC

Mochammad Ali Fahmi[1], Z.k. Abdurahman Baizal[2], Kiki Maulana[3]

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

Abstrak

Dengan berkembangnya ilmu pengetahuan dan teknologi dalam ilmu kedokteran, gen yang dimiliki oleh manusia dapat direpresentasikan dalam bentuk data dengan jumlah attribute yang banyak. Data tersebut dapat dikatakan sebagai data berdimensi tinggi. Data tersebut perlu diolah untk menggali informasi yang tersimpan, banyak algoritma yang dapat mengolah data tersebut. Salah satunya ialah algoritma CS4 (Cascading and Sharing for Ensemble of Decision Trees) yang merupakan varian dari ensemble method.

Attribut selection berfungsi untuk memilih attribute terbaik berdasarkan nilai information gain. Hal yang membedakan algoritma CS4 dengan algoritma bagging dan boosting adalah penggunaan data dalam pembangunan model. Algoritma CS4 menggunakan data asli dalam setiap pembangunan model, sedangkan algoritma bagging dan boosting menggunakan data bootstrap.

Dari pengujian penggunaan algoritma CS4 yang telah dilakukan, didapatkan bahwa tingkat keakurasian sangat ditentukan oleh penentuan jumlah attribut yang digunakan pada saat pembangunan model.

Kata Kunci : CS4, ensemble method, attribute selection, bagging, boostng.

Abstract

With the improvement of science and technology in medical, gen of human could be presented in a form of data with a large number of attribute. The data is known as high dimensional data. Data need to be processed to mining the information. Many algorithm could process the data. One of the algorithm is CS4 algorithm (Cascading and Sharing for Ensemble of Decision Trees) that is kind of ensemble method.

The function of attribute selection is to find the best attribute based on the value of information gain. The different of CS4 algorithm with bagging and boosting algorithm when construct model, CS4 algorithm use the original data but bagging and boosting algorithm use bootstrap data. The result of implementation CS4 algorithm that has been done, the accuracy level is depend with the number of attribut that is using for constructing tree.

Keywords : CS4, ensemble method, attribute selection, bagging, boostng.

1. Pendahuluan

1.1 Latar Belakang

Seiring dengan berkembangnya ilmu pengetahuan dan teknologi dalam ilmu kedokteran, banyak penyakit yang dapat menyebabkan kematian dapat diketahui lebih dini guna dilakukan pencegahan maupun pengobatan sehingga pasien tersebut dapat ditolong. Pada pengujian atau test yang dilakukan pada seorang pasien dihasilkan data yang memiliki attribut yang sangat banyak (*multi-attribut*). Dari data-data hasil pengujian diatas nantinya digali guna mendapatkan informasi yang bermanfaat, contohnya untuk melakukan prediksi. Proses pencarian atau penggalian informasi pada data tersebut dikenal sebagai data mining.

Dalam melakukan proses mining terhadap suatu data dapat menggunakan salah satu metoda dari data mining yaitu klasifikasi. Metoda klasifikasi ini dapat berupa *decision tree* atau pohon keputusan. Menurut [1], *decision tree* merupakan salah satu metoda klasifikasi yang paling populer karena mudah untuk diinterpretasi oleh manusia. *Tree* yang dihasilkan dari pemodelan suatu data kemudian membentuk suatu *rule* atau aturan ”jika - maka”.

Oleh karena itu, dalam tugas akhir ini akan diimplementasikan algoritma CS4 yang merupakan singkatan dari *Cascading and Sharing for ensemble of decision trees.* Algoritma CS4 ini sangat cocok digunakan untuk mengklasifikasikan data multi attribute, dikarenakan adanya *attribute selection* sehingga attribute yang digunakan hanya attribute penting saja, sedangkan attribute yang mempunyai nilai informasi yang sedikit tidak digunakan.Pada algoritma CS4 ini akan dibangun pohon sebanyak k-pohon, hal ini dilakukan karena suatu pohon dengan attribut terbaik sebagai *root* tidak selamanya merupakan hasil yang terbaik, terdapat kemungkinan bahwa attribut terbaik kedua sebagai root menghasilkan *rule* atau aturan yang *reliable*. Algoritma ini mempertimbangkan *top-ranked* attribut berbeda sebagai *root* pohon [5].

Nilai *k* merupakan jumlah *top-ranked* attribut yang digunakan untuk membangun *k*-pohon. Pada masing-masing pohon akan dihasilkan *rule* atau aturan-aturan, untuk memprediksi atau mengklasifikasikan data seorang pasien termasuk dalam kelas apa maka dapat dilihat aturan-aturan dari masing-masing pohon. Dari aturan tersebut didapatkan kelas yang paling dominan dari *k*-pohon, maka kelas yang dominan tersebut merupakan jawabannya.

Perbedaan antara algoritma CS4 dengan *ensemble method* lain yaitu algoritma bagging dan boosting adalah pada penggunaan data pada proses pembangunan model. Pada algoritma CS4 data yang digunakan dalam setiap pembangunan modelnya ialah data asli sedangkan pada algoritma bagging dan boosting memakai data *bootsrap* yaitu data random yang diambil dari data asli. Dengan menggunakan data asli pada pembangunan model maka model yang dihasilkan dapat mewakili data atau attribute yang tidak terpakai. Oleh karena itu algoritma CS4 lebih baik daripada algoritma bagging dan boosting untuk menangani klasifikasi pada data berdimensi tinggi.

1.2 Perumusan Masalah

Dengan mengacu latar belakang di atas, maka permasalahan yang dibahas dan diteliti adalah :

1. Bagaimana melakukan klasifikasi pada data *multi-attribut* menggunakan metoda *decision tree* dengan algoritma CS4.
2. Bagaimana melakukan pengujian dan analisis dari implementasi algoritma CS4.

Dalam pembahasan dan penelitian terhadap permasalahan di atas digunakan beberapa asumsi, yaitu :

1. Perbandingan hasil pengujian algoritma CS4 dengan algoritma lain (menggunakan *tools* WEKA).
2. Penggunaan DBMS Microsoft Access untuk penyimpanan data, baik *data training* maupun *data testing*.

Sedangkan ruang lingkup yang menjadi batasan dari pembahasan tugas akhir ini adalah :

1. Data Set yang digunakan merupakan data biomedik *multi-attribut*.
2. Data yang digunakan untuk analisis merupakan data yang *supervised* (memiliki *class label*).
3. *Data training* dan *data testing* bersih dari *noise* dan *missing value*.
4. Fase dikretisasi menghasilkan dua buah kategori yaitu data yang kurang dari sama dengan *cutpoint* dan data yang lebih dari *cutpoint* pada setiap attributnya.

1.3 Tujuan Pembahasan

Berdasarkan pada masalah yang telah didefinisikan di atas, maka tujuan tugas akhir ini adalah :

1. Membangun aplikasi yang mengimplementasikan algoritma CS4 untuk menyelesaikan klasifikasi pada data multi-attribut.
2. Menganalisis pengaruh nilai *k* (jumlah attribut yang digunakan dalam pembangunan model) yang bervariasi terhadap akurasi model yang dihasilkan oleh algoritma CS4.
3. Membandingkan tingkat *accuracy, precision, recall* dan *F-measure* algoritma CS4 dengan algoritma bagging dan boosting menggunakan *tools* Weka.

1.4 Metodologi Penyelesaian Masalah

Pendekatan sistematis atau metodologi yang akan digunakan dalam merealisasikan tujuan dan pemecahan masalah di atas adalah dengan menggunakan langkah-langkah sebagai berikut:

1. Studi literatur
   Membaca serta mempelajari literatur-literatur tentang data mining, klasifikasi, *decision tree*, algoritma ID3, algoritma CS4, *Information Gain*.

2. Pengumpulan dan pemahaman data

Mengumpulkan data set yang diperlukan yaitu berupa *data training* dan *data testing*, memahami maksud dari data-data maupun attribut tersebut serta melakukan *preprocessing* terhadap data sehingga data siap untuk di mining.

3. Analisis kebutuhan dan perancangan perangkat lunak
   Menentukan kebutuhan perangkat lunak serta perancangan perangkat lunak.
4. Implementasi
   Implementasi rancangan dengan membangun suatu perangkat lunak untuk melakukan klasifikasi objek-objek ke dalam *class* tertentu dengan menggunakan algoritma CS4.
5. Pengujian dan analisis hasil
   Melakukan pengujian perangkat lunak dengan data training sebagai data pembangun model dan data testing sebagai pengukur tingkat akurasi model yang dihasilkan oleh algoritma CS4. Membandingkan tingkat akurasi algoritma CS4 dengan algoritma lain.
6. Pengambilan kesimpulan dan penyusunan laporan Tugas Akhir.

1.5 Sistematika Penulisan

Tugas Akhir ini akan disusun berdasarkan sistematika penulisan sebagai berikut :

BAB 1 PENDAHULUAN
Bab ini memaparkan latar belakang, perumusan masalah yang dibahas, tujuan pembahasan, metode penyelesaian masalah dan sistematika penulisan.

BAB 2 LANDASAN TEORI
Bab ini memuat berbagai dasar teori yang mendukung dan mendasari penulisan tugas akhir ini, yaitu mengenai konsep dari *data mining, discretization, decision tree,* dan algoritma CS4 *(Cascading and Sharing for Ensemble of Decision Trees).*

BAB 3 ANALISIS DAN PERANCANGAN SISTEM
Bab ini menguraikan tentang tahapan yang dilakukan untuk membangun perangkat lunak sebagai pembantu dalam mendapatkan data untuk proses analisis, alur kerja (*work flow*) dari perangkat lunak yang dibuat, bagaimana keterhubungan antar objek dan kelas-kelas yang terbentuk.

BAB 4 IMPLEMENTASI DAN ANALISIS PENGUJIAN
Bab ini menyajikan implementasi hasil analisis dan perancangan sistem ke dalam bentuk pemrograman aplikasi serta melakukan pengujian terhadap aplikasi menggunakan data sample dengan merubah nilai konstanta-*k*, untuk mencari nilai konstanta-*k* yang ideal sehingga memperoleh prediksi yang optimum. Mengevaluasi pengaruh variabel *k* pada tingkat akurasi.

BAB 5 KESIMPULAN DAN SARAN
Berisi kesimpulan dari hasil penelitian tugas akhir ini serta saran-saran untuk pengembangan lebih lanjut.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Kesimpulan yang dapat diambil dari Tugas Akhir judul "Analisis dan Implementasi Algoritma CS4 (*Cascading and Sharing for Ensemble of Decision Trees*) pada Data Biomedical Multi-Attribut" adalah :

1. Penentuan nilai *k* terhadap tingkat akurasi yang dihasilkan tergantung pada *data training*-nya. Tidak selamanya penggunaan seluruh attribute menghasilkan tingkat akurasi yang terbaik.
2. Proses *attribute selection* sangat penting dalam algoritma CS4 untuk menentukan attribute yang terlibat dalam pembangunan model algoritma CS4.
3. Jumlah rule yang dihasilkan dari algoritma CS4 lebih sedikit daripada jumlah rule yang dihasilkan algoritma lain dengan data yang sama.
4. Tingkat *accuracy* dan *F-measure* yang dihasilkan algoritma CS4 lebih baik daripada menggunakan algoritma bagging dan boosting dengan *base classifier decision stump, randomtree* dan *adtree*.

5.2 Saran

Saran yang dapat diajukan untuk pengembangan Tugas Akhir judul "Analisis dan Implementasi Algoritma CS4 (Cascading and Sharing for Ensemble of Decision Trees) pada Data Biomedical Multi-Attribut" adalah :

1. Dataset yang digunakan pada pengujian hanya terbatas dengan jumlah attribute 1000, 1500 dan 2000. Untuk pengembangan selanjutnya, data yang digunakan memiliki jumlah attribute yang lebih banyak.
2. Penggunaan *pruning* dalam pembangunan *tree* sehingga dapat memungkinkan rule yang dihasilkan lebih sedikit dan model yang dihasilkan lebih sederhana.

Daftar Pustaka

[1] Han, Jiawei dan Kamber, Micheline. 2006. *Data Mining : Concepts and Techniques 2nd Editio.* San Fransisco : Morgan Kaufmann Publishers.

[2] Tan, P., Steinbach, M. dan Kumar, V.. 2004. *Introduction to Data Mining*. Addison-Wesley.

[3] Larose, Daniel T.. 2005. *Discovering Knowledge in Data : An Introduction to Data Mining*. New Jersey : John Wiley & Sons, Inc.

[4] Michael, Hans.. 2004. *ID3 : Induksi Decision Tree.* http://www.hansmichael.com/download/DiktatID3.pdf didownload pada tanggal 28 Juni 2008.

[5] Li, Jinyan dan Liu, Huiqing. 2003. *Ensembles of Cascading Trees.* http://sdmc.i2r.a-star.edu.sg/jinyan/publications/lij_cascadingtrees.pdf. didownload pada tanggal 21 September 2007.

[6] 2007. *Decision Tree*. http://en.wikipedia.org/wiki/Decision_tree. didownload pada tanggal 02 November 2007.

[7] 2007. *Kent Ridge Data Set Repository*. http://sdmc.i2r.a-star.edu.sg/GEDatasets/Datasets.html. diakses pada tanggal 01 November 2007.

[8] Moore, Andrew W.. 2003. *Information Gain.* http://www.autonlab.org/tutorials/infogain11.pdf. didownload pada tanggal 07 November 2007.

[9] 2007. *Information Gain in Decision Tree.* http://en.wikipedia.org/wiki/Information_gain_in_decision_trees. didownload pada tanggal 07 November 2007.