ANALISIS PERBANDINGAN ALGORITMA DECISION TREE PADA PERMASALAHAN IMBALANCE CLASS

Thahira Kemala Dewi[1], Arie Ardiyanti Suryani[2], Angelina Prima Kurniati[3]

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

Abstrak

Imbalance class adalah salah satu permasalahan yang ada dalam data mining. Permasalahan ini diakibatkan karena adanya jumlah data yang tidak seimbang pada kelas yang berbeda, dimana kelas yang satu memiliki jumlah data yang sangat banyak (mayoritas), sedangkan kelas lainnya memiliki jumlah data yang sangat sedikit (minoritas). Pada pengklasifikasian biasa, kelas minoritas tersebut tidak dapat terprediksi. Hal ini disebabkan karena bila jumlah data pada satu kelas sangat kecil, kelas tersebut akan salah diprediksi sebagai kelas mayoritas.

Salah satu metode klasifikasi yang biasa digunakan adalah metode decision tree. Metode decision tree ini akan membentuk model seperti flow chart dengan struktur tree. Pada Tugas Akhir ini diperbandingkan tiga algoritma pada decision tree yaitu C4.5, CART, dan HDDT. Setiap algoritma tersebut memiliki splitting criterion atau attribute selection measure yang berbeda dengan tingkat sensitivitas yang berbeda pula terhadap kecondongan data.

Analisis yang dilakukan pada Tugas Akhir ini adalah mengetahui bagaimana performansi dari algoritma C4.5, CART, dan HDDT yang memiliki splitting criterion yang berbeda pada permasalahan imbalance class serta bagaimana perbandingan performansi dari ketiga algoritma decision tree tersebut berdasarkan nilai precision, recall, dan f-measure.

Hasil yang didapat dari penelitian menunjukkan bahwa algoritma HDDT memiliki performansi yang lebih baik pada tingkat imbalance yang rendah dibandingkan algoritma CART dan C4.5. Hal ini disebabkan karena algoritma HDDT memiliki tingkat sensitivitas yang rendah terhadap kecondongan data yang menjadi masalah dalam imbalance class.

Kata Kunci : imbalance class, algoritma C4.5, CART, HDDT, performansi.

Abstract

Imbalance class is one of existing problems in data mining. This problem is caused by number of data in which is uneven in the different class, where the one of the class own the number of a lot of data (majority), while the other classes own a little the number of data (minority). In the ordinary classification, the minority class cannot classified. This is caused by when the number of data in the very small class, the class will be wrong classified as majority class.

One of classification method which commonly use is decision tree method. This method will form model like flow chart with structure tree. In this final project is comparing three algorithm of decision tree that is C4.5, CART, and HDDT. Every algorithm own splitting criterion or different attribute selection measure with level of different sensitivities also to data skewnes.

The analysis in final project is knowing how performance from the C4.5, CART, and HDDT algorithm owning different splitting criterion in problems of imbalance class and also how comparison of performance from third of the algorithm decision tree based on value of precision, recall, and f-measure.

The result from this research present that HDDT algorithm own better performance at low level of imbalance compared to CART and C4.5 algorithm. This is caused by algorithm of HDDT own level of low sensitivities to skewness of data becoming internal issue of imbalance class.

Keywords : imbalance class, algorithm C4.5, CART, HDDT, performance.

1. Pendahuluan

1.1. Latar Belakang

Cepatnya kemajuan dalam pengumpulan data dan teknologi penyimpanan data memungkinkan suatu organisasi untuk menghimpun data dalam jumlah yang sangat besar tanpa dapat dimanfaatkan dengan baik. Akibatnya pengambilan keputusan seringkali berdasarkan intuisi saja, sebab para pengambil keputusan tidak mempunyai alat untuk mengekstrak informasi barharga dari database yang sangat besar. Hal inilah yang melatarbelakangi lahirnya sebuah bidang ilmu yang berupaya menemukan pola ataupun informasi yang belum diketahui sebelumnya dari sekumpulan besar data, yaitu data mining.

Dalam data mining juga terdapat berbagai macam permasalahan, salah satunya yaitu permasalahan *imbalance class* atau kelas *imbalance.* Permasalahan ini diakibatkan karena adanya jumlah data yang tidak seimbang pada kelas yang berbeda, dimana kelas yang satu memiliki jumlah data yang sangat banyak (mayoritas), sedangkan kelas lainnya memiliki jumlah data yang sangat sedikit (minoritas). Pada pengklasifikasian biasa, kelas minoritas tersebut tidak dapat terprediksi. Hal ini disebabkan karena bila jumlah data pada satu kelas sangat kecil, kelas tersebut akan salah diprediksi sebagai kelas mayoritas. Salah satu metode klasifikasi yang biasa digunakan adalah metode *decision tree.* Metode *decision tree* ini akan membentuk model seperti *flow chart* dengan struktur *tree*. *Node* internal menandakan *test* terhadap sebuah atribut. Cabang merepresentasikan keluaran dari *test*. *Node leaf* merepresentasikan label kelas atau distribusi kelas.

Pada Tugas Akhir ini diperbandingkan tiga algoritma pada *decision tree* yaitu C4.5, CART, dan HDDT. Setiap algoritma tersebut memiliki *splitting criterion* atau *attribute selection measure* yang berbeda dengan tingkat *sensitivitas* yang berbeda pula terhadap kecondongan data. Algoritma C4.5 merupakan algoritma yang paling populer digunakan. Algoritma ini menggunakan *information gain* sebagai *splitting criterion-*

nya. *Information gain* adalah salah satu a*ttribute selection measure* yang digunakan untuk memilih *test* atribut tiap *node* pada *tree*. Atribut dengan *information gain* tertinggi dipilih sebagai *test* atribut dari suatu *node*. CART merupakan kepanjangan dari *classification and regression tree*. Algoritma CART menggunakan *gini index* sebagai *splitting criterion*-nya. Dan algoritma yang ketiga yaitu algoritma *hellinger distance decision tree* (HDDT). Algoritma ini menggunakan suatu *measure* yaitu *hellinger distance* sebagai *splitting criterion*-nya.

Pada Tugas Akhir ini, dianalisis bagaimana performansi dari ketiga algoritma *decision tree* tersebut terhadap data yang bersifat *imbalance* serta akan diperbandingkan algoritma mana yang memiliki performansi yang paling baik. Performansi yang diukur adalah nilai *precession, recall*, dan *f-measure*.

1.2. Perumusan Masalah

Permasalahan yang dijadikan objek penelitian dalam Tugas Akhir ini antara lain :

1. Bagaimana performansi dari algoritma C4.5, CART, dan HDDT yang memiliki *splitting criterion* yang berbeda pada permasalahan kelas *imbalance*.
2. Bagaimana perbandingan performansi dari ketiga algoritma *decision tree* tersebut berdasarkan nilai *precision, recall*, dan *f-measure*.

1.3. Batasan Masalah

Untuk menghindari meluasnya materi pembahasan Tugas Akhir ini, maka permasalahan yang dibahas dalam Tugas Akhir ini mencakup hal-hal berikut :

1. Dataset yang digunakan adalah dataset dari UCI Machine Learning Repository dan dataset buatan yang dihasilkan dari data generator.
2. Dataset yang digunakan merupakan data dengan kelas *binary* tanpa *missing value*.
3. Implementasi algoritma HDDT menggunakan NetBeans IDE 6.0 dengan JDK 1.6.

4. Klasifikasi algoritma C4.5 dan CART menggunakan *classifier* yang ada di WEKA, yaitu J48 yang menerapkan algoritma C4.5 dan SimpleCart yang menerapkan algoritma CART.

1.4. Tujuan Pembahasan

Dalam Tugas Akhir ini, hal-hal yang diharapkan untuk dicapai adalah sebagai berikut :

1. Menganalisis performansi dari algoritma C4.5, CART, dan HDDT terhadap permasalahan kelas *imbalance* yang diterapkan pada beberapa dataset dengan tingkat *imbalance* yang berbeda.
2. Menganalisis dan mengetahui performansi mana yang lebih baik dari ketiga algoritma *decision tree* tersebut pada dataset yang bersifat *imbalance* dilihat dari nilai *precision, recall,* dan *f-measure*. Semakin besar nilai *precision, recall,* dan *f-measure* dari suatu algoritma, semakin baik performansi algoritma tersebut.

1.5. Metodologi Penyelesaian Masalah

Metode yang akan digunakan untuk menyelesaikan Tugas Akhir ini adalah :

1. Studi literatur

   Berupa pencarian sumber-sumber bacaan yang dapat menunjang topik Tugas Akhir ini. Yaitu dengan mencari referensi dan mendalami seluruh materi yang berhubungan dengan data mining, klasifikasi dengan data *imbalance*, algoritma C4.5, CART, HDDT, *information gain, gini index*, dan *hellinger distance measure*.

2. Pengumpulan data-data penunjang Tugas Akhir

   Berupa pengumpulan data penunjang yang dapat membantu perancangan sistem. Data penunjang tersebut akan digunakan untuk pengujian dan analisis, maupun data lain yang membantu terselesainya Tugas Akhir ini.

3. Analisis dan perancangan sistem

   Menganalisis masalah dan perancangan perangkat dengan menggunakan metode yang telah dipilih sebagai batasan masalah. Output dari perangkat lunak yang akan dianalisis, di-*training* dan di-*testing* kemudian dibandingkan satu sama lain sesuai variasi penggunaan algoritmanya.

4. Implementasi

   Berupa realisasi sistem dari rancangan yang dikembangkan. Sistem direalisasikan dengan menggunakan program aplikasi NetBeans IDE 6.0.

5. Evaluasi unjuk kerja sistem

   Melakukan pengujian perangkat lunak dengan memperbandingkan hasil performansi dari setiap algoritma.

6. Penulisan dokumentasi dan laporan

   Mengambil kesimpulan dari analisis hasil yang telah dilakukan, kemudian menyusun laporan dari analisis hasil yang telah dilakukan.

5. Kesimpulan dan Saran

5.1. Kesimpulan

1. Algoritma HDDT memiliki performansi yang paling baik untuk dataset dengan tingkat *imbalance* yang rendah (dibawah 30%) baik pada dataset dengan tipe atribut kategorikal ataupun kontinu.
2. Algoritma HDDT memiliki tingkat sensitivitas yang paling kecil terhadap kecondongan data dibandingkan algoritma CART dan C4.5.
3. Algoritma C4.5 memiliki performansi yang paling baik secara keseluruhan yang diujikan pada beberapa dataset real UCI *Machine Learning Repository.*
4. Algoritma C4.5 dan CART dengan *pruning* memiliki performansi yang lebih buruk dilihat dari nilai *f-measure* daripada tanpa *pruning* pada dataset dengan tingkat i*mbalance* yang rendah.
5. Dalam permasalahan *imbalance class* jumlah data tidak mempengaruhi performansi dari setiap algoritma.

5.2. Saran

1. Perlu dilakukan penelitian yang lebih lanjut untuk dataset yang *multi class.* Untuk mengetahui bagaimana performansi dari ketiga algoritma tersebut pada dataset selain *binary class*.
2. Untuk memperbaiki tingkat keakurasian, maka hendaknya pemilihan data yang digunakan untuk *training* dapat lebih tersebar merata. Agar menghasilkan model yang mampu menangani berbagai kemungkinan pada saat *testing*.

Daftar Pustaka

[1] Arie Yanuar. 2008. Tugas Akhir. *Analisis Perbandingan Metode Boosting untuk Klasifikasi Kelas Imbalance*. Jurusan Teknik Informatika Institut Telkom Bandung.

[2] C. Apte and S.M. Weiss. "*Data Mining with Decision Trees and Decision Rules*". Future Generation Computer Systems. November 1997.

[3] David A. Cieslak and Nitesh V. Chawla. "*Increasing Skew Insensitivity of Decision Trees with Hellinger Distance*". University of Notre Dame, Notre Dame IN 46556. USA.

[4] David A. Cieslak and Nitesh V. Chawla. "*Learning Decision Trees for Unbalanced Data*". University of Notre Dame, Notre Dame IN 46556. USA.

[5] Floriana E, Donato M, and Giovanni S. *"A Comparative Analysis of Methods for Pruning Decision Trees"*. Volume 19, No.5. IEEE Transactions On Pattern Anlysis and Machine Intelligence. May 1997

[6] Jiawei Han and Micheline Kamber. *"Data Mining: Concepts and Techniques"*. Department of Computer Science University of Illinois at Urbana-Champaign. USA

[7] Laura E. Raileanu and Kilian Stoffel. "*Theoretical Comparison between the Gini Index and Information Gain Criteria*". University of Neuchˆatel, Computer Science Departement. Switzerland.

[8] Nitesh V. Chawla , Nathalie Japkowicz and Aleksander Ko lcz . "*Editorial: Special Issue on Learning from Imbalanced Data Sets*". Volume 6, Issue 1 - Page 6. Sigkdd Explorations.

[9] Pang-Ning Tan, Michael Steinbach and Vipin Kumar. "*Introduction to Data Mining*". Pearson Education, Inc. 2006.

[10] Rafi K. and Ming Dong. "*Decision Trees For Classification : A Review and Some New Result*". University of Cincinnati. USA.

[11] Roger J. Lewis. "*An Introduction to Classification and Regression Tree (CART) Analysis*". Department of Emergency Medicine. California. 2000.

[12] Ross Quinlan. 2008. *"C4.5 algorithm"*. http://en.wikipedia.org/wiki/C4.5_algorithm. Didownload pada 30 November 2008.

[13] Witten I. H. and Frank E. *"Data Mining : Practical Machine Learning Tools and Techniques"*. 1996.