PENGGUNAAN COST SENSITIVE LEARNING UNTUK KLASIFIKASI DATA IMBALANCE DENGAN METODE NAIVE BAYES
COST SENSITIVE USAGE FOR CLASSIFYING IMBALANCE DATA WITH NAIVE BAYES METHOD

Korano Yudha Prajalaksana[1], Moch. Arif Bijaksanaech[2], Eddy Muntina Dharma[3]

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

Abstrak

Cost Sensitive Learning merupakan pendekatan pada proses pembelajaran klasifikasi dimana dalam proses klasifikasi diperhitungkan juga misclassification cost. Misclassification cost merupakan kerugian (cost) yang timbul apabila suatu data salah diklasifikasikan. Misclassification cost merupakan hal yang sangat penting karena data yang salah diklasifikasikan dapat menimbulkan cost yang berbeda-beda. Untuk penerapan cost sensitive learning dalam klasifikasi data imbalance digunakan metode pembelajaran AdaCost, yaitu metode pembelajaran yang melibatkan perhitungan misclassification cost. Sedangkan untuk klasifikasi data digunakan metode klasifikasi Naive Bayes, yaitu metode klasifikasi yang menggunakan perhitungan probabilitas dalam memecahkan persoalan klasifikasi. Pada Tugas Akhir ini telah dianalisis dan diimplementasikan klasifikasi data imbalance menggunakan AdaCost dan Naive Bayes. Pengujian dilakukan dengan beberapa skenario, antara lain pengujian dengan perbedaan jumlah ronde boosting, pengujian dengan membandingkan performansi AdaCost dengan metode klasifikasi dan metode boosting lain, dan pengujian menggunakan data yang terdapat missing value. Metode klasifikasi yang digunakan sebagai pembanding dalam pengujian adalah Decision Tree, OneR, Neural Network, dan Naive Bayes. Sedangkan metode boosting yang digunakan sebagai pembanding adalah AdaBoost, AdaBoost M1, MultiBoost AB, dan LogitBoost.

Dari hasil pengujian didapatkan kesimpulan bahwa algoritma AdaCost lebih memprioritaskan pencarian pola kelas imbalance sehingga dapat meningkatkan performansi base classifiernya dalam mengklasifikasikan kelas imbalance dan meminimalkan misclassifcation cost. AdaCost juga lebih baik dalam mengklasifikasikan kelas imbalance bila dibandingkan dengan metode AdaBoost, AdaBoost M1, Multiboost AB dan LogitBoost. Namun, performansi dari AdaCost tergantung dari jumlah ronde boostingnya.

Kata Kunci : cost sensitive learning, misclassification cost, data imbalance, adacost, naïve bayes

Abstract
Cost Sensitive Learning is an approach that used in classifier studying process which in classified process, we must also consider misclassification cost. Misclassification cost is a cost that occurs everytime there is a data mislabel. Misclassification cost is very important because when a data is mislabeled it may give different cost. We used AdaCost studying method as cost sensitive learning implementation in classifying imbalance data that using misclassification cost estimation. And as classifier method we used Naïve Bayes, which this method using probabilistic estimation to solve classification problem.
In this Final Project, it has been analysed and implemented imbalance data classification using AdaCost and Naive Bayes. The test was done in different scenarios, such as the test using different number of boosting rounds, the test to compare AdaCost performance with another classification and boosting methods, and the test using data that contains missing value. Classification methods which is used for the comparison are Decision Tree, OneR, Neural Network, and Naive Bayes. While for the boosting methods, we used AdaBoost, AdaBoost M1, MultiBoost AB, andLogitBoost.
From the test result it has been concluded that AdaCost algorithm prioritize searching imbalance class pattern, boosting the performance of the base classifier to classify imbalance class, more than AdaBoost, AdaBoost M1, Multiboost AB or LogitBoost. But, AdaCost's performance depends on its boosting rounds

Keywords : cost sensitive learning, misclassification cost, data imbalance, adacost, naïve bayes

---

BAB I

PENDAHULUAN

1.1 Latar Belakang

Akibat perkembangan teknologi yang demikian cepat, kebutuhan akan data yang ada saat ini sudah berkembang menjadi bagaimana mencari pola tersembunyi yang terdapat di dalam data untuk memperoleh peluang-peluang baru. Pengenalan pola dalam data dipelajari dalam disiplin ilmu *data mining*.

Dalam *data mining* terdapat berbagai metode untuk menemukan pola tertentu dalam data. Salah satu metode yang digunakan ialah klasifikasi. Dalam klasifikasi data dianggap sebagai kumpulan data tunggal, dimana setiap data tunggal terdiri dari beberapa atribut dan satu atribut kelas. Selanjutnya dari kumpulan data tersebut akan dicari sebuah model untuk atribut kelas sebagi fungsi dari nilai atribut lainnya. Tujuan akhirnya ialah agar data yang sebelumnya *invisible* dapat diberikan kelas yang sesuai dengan seakurat mungkin. Permasalahan mulai timbul ketika data pada dunia nyata berkembang menjadi bersifat *imbalance*, yaitu suatu kondisi dimana terdapat ketidakseimbangan atau terjadi ketimpangan dalam jumlah data dimana salah satu kelas dalam data tersebut merepresentasikan kasus yang bersifat jarang terjadi sehingga jumlah datanya relatif sedikit dibandingkan dengan jumlah data yang merepresentasikan kasus normal. Beberapa metode klasifikasi yang ada sekarang ini lebih ditujukan untuk kasus yang jumlah data tiap kelasnya seimbang, dengan tujuan akhir memaksimalkan akurasi keseluruhan, sehingga metode tersebut tidak dapat memprediksi data *imbalance* secara maksimal. Salah satu pendekatan yang dapat digunakan untuk memprediksi data *imbalance* ialah menggunakan *cost sensitive learning*.

Cost sensitive learning merupakan pendekatan pembelajaran dalam *data mining.* Ide dasar dari *cost sensitive learning* berasal dari sifat dasar manusia, dimana dalam mengambil suatu keputusan, keputusan yang diambil yang mengandung resiko seminimal mungkin. Dalam *cost sensitive learning* yang menjadi perhatian bukanlah mencapai klasifikasi yang seakurat mungkin atau tingkat kesalahan seminimal mungkin dalam mengelompokkan data, melainkan bagaimana meminimalkan *misclassification cost*.

Dalam Tugas akhir ini, masalah yang diambil adalah bagaimana penerapan *AdaCost* sebagai metode *cost sensitive learning* dan Naive Bayes sebagai metode klasifikasi untuk memprediksi data *imbalance*. Dalam *AdaCost*, kelas yang jumlahnya lebih kecil akan diberi *misclassification cost* yang lebih besar dibanding data yang jumlahnya lebih besar sehingga kelas tersebut mendapatkan perhatian lebih walaupun jumlahnya sangat sedikit sehingga diharapkan keseluruhan kelas yang jumlahnya kecil dapat diprediksikan secara maksimal.

2.1 Perumusan Masalah

Berdasarkan latar belakang yang dikemukakan di atas, maka masalah yang akan dibahas dalam tugas akhir ini adalah bagaimana mengimplementasikan *cost sensitive learning* dalam mengklasifikasikan data *imbalance* menggunakan

metode pembelajaran *AdaCost* yang diharapkan dapat meningkatkan predikasi klasifikasi oleh metode klasifikasi *Naive Bayes*.

3.1 Tujuan

Tujuan dari penulisan Tugas Akhir dengan judul " **Penggunaan Cost Sensitive Learning Untuk Klasifikasi Data Dengan Metode Naive Bayes**" yaitu :

1. Mengimplementasikan pendekatan *cost sensitive learning* dalam pembuatan klasifikasi menggunakan metode *AdaCost*.

2. Menganalisa hasil penerapan metode *AdaCost* pada prediksi klasifikasi data *imbalance* dengan metode klasifikasi *Naïve Bayes* menggunakan parameter *recall*, *presicion*, *F-Measure* dan *total misclassification cost*.

4.1 Batasan Masalah

Dalam penyusunan tugas akhir ini permasalahan dibatasi dalam beberapa hal yaitu:

1. Data yang digunakan telah melalui tahap *preprocessing* data sehingga bebas dari noise dan data yang tidak konsisten.

2. Sistem menerima masukan berupa sekumpulan data yang kelasnya bertipe nominal.

3. Data yang digunakan terutama yang bersifat *imbalance* dan sudah tersedia dalam bentuk tabel dengan atribut kelas berada pada urutan paling akhir.

4. *Misclassification cost* untuk kelas *imbalance* selalu lebih besar daripada *misclassification cost* untuk kelas bukan *imbalance*.

5.1 Metodologi Penyelesaian Masalah

Metode yang akan digunakan untuk menyelesaikan tugas akhir ini adalah :

a. Studi Literatur

   Mempelajari dasar teori dan literatur-literatur tentang data mining, *imbalance class*, *Cost Sensitive Learning*, klasifikasi Naive Bayes , dan *AdaCost*.

b. Studi Pengembangan Aplikasi

   Bertujuan untuk menentukan metodologi pengembangan perangkat lunak yang digunakan dengan pendekatan terstruktur dan melakukan analisa perancangan

c. Desain dan Implementasi Perangkat Lunak

   Bertujuan untuk melakukan implementasi metode pada perangkat lunak sesuai dengan analisa perancangan yang telah dilakukan.

d. Pengujian dan Analisa Perangkat Lunak

Dalam tahap ini akan diuji dan dianalisis hasil penerapan metode *AdaCost* terhadap prediksi klasifikasi *Naive Bayes* yang dibuat untuk berbagai kasus yang disediakan dan mencari kesalahan-kesalahan yang masih muncul dalam pengimplementasian.

e. Pengambilan kesimpulan dan penyusunan laporan

6.1 Sistematika Penulisan

Sistematika penulisan tugas akhir ini adalah sebagai berikut :

BAB I PENDAHULUAN

Berisi latar belakang pembuatan tugas akhir, perumusan masalah, tujuan, batasan masalah, metodologi penyelesaian masalah dan sistematika penulisan

BAB II LANDASAN TEORI

Pada bab ini dikemukakan berbagai teori yang mendukung penyusunan tugas akhir, antara lain meliputi teori tentang data mining, *imbalance class*, *Cost Sensitive Learning*, klasifikasi *Naive Bayes*, dan *boosting* (meliputi *AdaBoost* dan *AdaCost*).

BAB III ANALISIS KEBUTUHAN DAN PERANCANGAN SISTEM

Menjelaskan tentang proses analisa masalah dan kebutuhan perangkat lunak. Berisi rancangan menggunakan notasi-notasi Diagram Aliran Data (DAD).

BAB IV IMPLEMENTASI DAN ANALISA HASIL PENGUJIAN

Implementasi yang detail mengenai sistem yang dibangun berdasarkan rancangan pada Bab III serta hasil analisis pengujian terhadap beberapa skenario tes.

BAB V KESIMPULAN DAN SARAN

Berisi kesimpulan akhir dan saran pengembangan dari tugas akhir.

BAB V

KESIMPULAN

5.1. Kesimpulan

Kesimpulan yang dapat diambil dari penelitian tugas akhir ini antara lain :

1. Karakteristik kelas *imbalance* yang jumlahnya sangat sedikit (berkisar antara 0.01% sampai 10%) dibandingkan kelas lainnya menjadikan kelas ini seringkali dianggap sebagai *noise*. Hal ini mengakibatkan penurunan performansi suatu *classifier* dalam mengklasifikasikan kelas *imbalance* ini.

2. Jumlah ronde *boosting* mempengaruhi performansi AdaCost dalam meminimalkan nilai total *misclassification cost*. Untuk memperoleh performansi yang optimal, maka kita perlu menentukan jumlah ronde *boosting* yang tepat. Pada kasus hepatitis, performansi yang optimal didapat pada ronde ke 4 dengan nilai total *misclassification cost* 340.

3. AdaCost sangat baik dalam mengklasifikasi kelas *imbalance* sekaligus dapat meningkatkan performansi nilai recall dan nilai total *misclassification cost base classifier*nya.

4. AdaCost cukup baik dalam memprediksi kelas *imbalance* bila dibandingkan dengan algoritma klasifikasi lainnya seperti decision tree J.48, OneR, Neural Network, C.5, dan Naive Bayes.

5. Metode klasifikasi Naive Bayes lebih baik daripada One Rule, C.5 dan Neural Network dalam mengklasifikasikan data dengan *missing value*.

6. Algoritma AdaCost dapat meningkatkan performansi metode klasifikasi Naïve Bayes dalam mengklasifikasikan data dengan *missing value*.

7. AdaCost lebih baik dalam mengklasifikasikan data *imbalance* secara keseluruhan dan meminimalkan *misclassification cost* daripada AdaBoost, AdaBoost M1, Multiboost AB dan LogitBoost

5.2. Saran

Pengembangan yang dapat dilakukan pada tugas akhir ini adalah :

1. Implementasi AdaCost dapat dikembangkan dengan menggunakan *classifier* lain sebagai base classifiernya. Hal ini dikarenakan performansi algoritma *boosting* (AdaCost) sangat bergantung pada pemilihan base classifiernya

2. Pengembangan dilakukan sehingga dapat menangani kelas bertipe numerik.

3. Menerapkan algoritma yang berdasarkan *cost sensitive learning* lainnya.

Daftar Pustaka

[1] Blake, C. & Merz, C., 1998, Uci repository of machine learning databases, http://www.ics.uci.edu/~mlearn/~MLRepository.html.

[2] Bianca Z, John L, Naoki Abe. **"**Cost-Sensitive Learning by Cost-Proportionate ExampleWeighting", Mathematical Sciences Department, IBM T. J. Watson Research Center, Yorktown Heights, NY.

[3] Elkan, Charles, "The Foundations of Cost-Sensitive Learning", Department of Science and Engineering, University of California, San Diego.

[4] Fan, Stolfo, Zhang J., and Chan, 1999 "AdaCost: Misclassification Cost-sensitive Boosting", *Proceedings of the Sixteenth International Conference on Machine Learning,* San Fransisco, Morgan Kaufmann.

[5] Han, J., Kamber, M., 2001, "Data Mining : Concept and Techniques", Simon Fraser University, Morgan Kaufmann.

[6] Jo, Taeho , Nathalie J, "Class Imbalance versus Small Disjuncts", University of Otawa.

[7] Mahesh V. J, Ramesh C. A, Vipin K., "Predicting Rare Classes :Can Boosting Make Any Weak Learner Strong?", SIGKDD '02 Edmonton, Alberta, Canada.

[8] Margineantu, Dragos, 2001, "Methods for Cost-Sensitive Learning", Oregon State University.

[9] Nathalie J, "Inductive Learning from Imbalanced Data Sets" , University of Ottawa.

[10] P. Roger S, 2002, "Rekayasa Perangkat Lunak : Pendekatan Praktisi (diterjemahkan oleh LN. Harnaningrum)", buku I, Andi, Yogyakarta

[11] Schapire, R., Singer Y., "Improved Boosting Algorithms Using Confidence-rated Prediction", AT&T Labs, Shannon Laboratory, New York..

[12] Schapire, Robert, "A Brief Introduction to Boosting", AT&T Labs, Shannon Laboratory, New York.

[13] Tan, Steinbach, M., Kumar, V., 2006, "Introduction to Data Mining", Pearson, Addison Weasley.

Telkom University

LAMPIRAN A

ILUSTRASI ALGORITMA ADABOOST

Sebuah tabel berikut merepresentasikan kasus *imbalance class*. Tabel berikut merupakan contoh sebuah data latih yang digunakan untuk mengklasifikasi evade:

Tabel A- 1: Data Latih Evade

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	No
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	No

Dalam penelitian ini kita akan mencoba mengklasifikasi data evade. Telah dibangun sekumpulan data latih dengan jumlah kelas evade=*yes* adalah 1 dan kelas evade=*no* adalah 9. Terlebih dahulu kita sepakati bahwa kelas *no* notasinya -1, kelas *yes* notasinya 1, dan jumlah ronde *boosting* sebanyak 3. Kita akan mengikuti langkah demi langkah pada algoritma *AdaBoost*.

1. Inisialisasi $D1(i) = 1/m$

   m ialah jumlah data pada data latih, dalam hal ini 10, sehingga $D1(i)$=1/10=0.1

Tabel A- 2 : Tabel distribusi ronde 1

round	x=1	x=2	x=3	x=4	x=5	x=6	x=7	x=8	x=9	x=10
1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

2. Untuk iterasi t=1, kita gunakan distribusi di atas untuk proses pelatihan. Asumsi dengan metode klasifikasi sembarang, v, kita dapatkan hipotesis ht

Tabel A- 3 : Prediksi ronde 1

1	2	3	4	5	6	7	8	9	10
no	yes	yes	no	no	no	no	no	no	no

3. Mencari nilai $\alpha t \in R$

   Sebelum mencari nilai αt, terlebih dahulu kita harus mencari nilai $ui = yiht(xi)$ dan $r = \sum_i D(i)ui$

   $u1$=-1.-1=1 $u2$=-1.1=-1 $u3$=-1.1=-1 $u4$=-1.-1=1 $u5$=-1.-1=1