ANALISIS ALGORITMA RAREBOOST-1 DALAM KASUS IMBALANCE CLASS

Ary Chandra Irawan[1], Moch. Arif Bijaksana[2], Dhinta Darmantoro[3]

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

Abstrak

Imbalance class merupakan ketidakseimbangan dalam jumlah data training antara dua kelas yang berbeda, salah satu kelasnya merepresentasikan jumlah data yang sangat besar (mayority class) sedangkan kelas yang lainya merepresentasikan jumlah data yang sangat kecil (minority class). Tekhnik klasifikasi tidak dapat memprediksi minority class tersebut sehingga diperlukan suatu cara agar minority class dapat terprediksi dengan baik. Imbalance class dapat ditangani dengan boosting yaitu suatu tekhnik yang menggabungkan tekhnik klasifikasi dengan salah satu algoritma boosting yaitu RareBoost-1. Dalam algoritma RareBoost-1 dilakukan oversampling sederhana terhadap minority class dari data training dan pembobotan terhadap data training. Oversampling dan pembobotan tersebut dilakukan pada setiap iterasi dan pada akhir iterasi akan dilakukan voting terhadap hasil prediksi dari setiap iterasi. Hasil pengujian menunjukkan bahwa algoritma RareBoost-1 dapat memprediksi minority class sehingga performansi semakin baik jika dibandingkan dengan klasifikasi original.

Kata Kunci : imbalance class, boosting, algoritma RareBoost-1, oversampling,

Abstract

Imbalance class represents imbalance in number of training data between two different classes, one of the classes represents mayority class and another classes represents minority class. Classification technique cannot predict minority class so that it needs a way of technique for predict minority class. Imbalance class can handle with boosting that is technique which joining classification technique with one of the boosting algorithm that is RareBoost-1. In Rareboost-1 algorithm, it done simple oversampling to minority class of training data and weighted to training data. Oversampling and the weighted done in each iterasi and by the end of iterasi will be done voting to result of predict from each iterasi. Result of examination indicate that Rareboost-1 algorithm can predict minority class so that good performance progressively in comparison with original classification.

Keywords : imbalance class, boosting, RareBoost-1 algorithm, oversampling,

1. Pendahuluan

1.1 Latar Belakang

Perkembangan ilmu pengetahuan dan teknologi telah mengakibatkan perkembangan data di dunia semakin kompleks. Untuk menangani data-data tersebut diperlukan sistem pengolahan data yang baik. Masalah yang timbul dalam pengolahan data adalah *imbalance class*. *Imbalance class* merupakan ketidakseimbangan dalam jumlah data *training* antara dua kelas yang berbeda, salah satu kelasnya merepresentasikan jumlah data yang sangat besar (*mayority class)* sedangkan kelas yang lainya merepresentasikan jumlah data yang sangat kecil (*minority class*). *Imbalance class* biasanya terjadi pada kasus yang anomali. Contoh-contoh kasus untuk *imbalance class* adalah *detecting fraudulent transaction, network intrusion detection, Web Mining,* dan *direct marketing.*

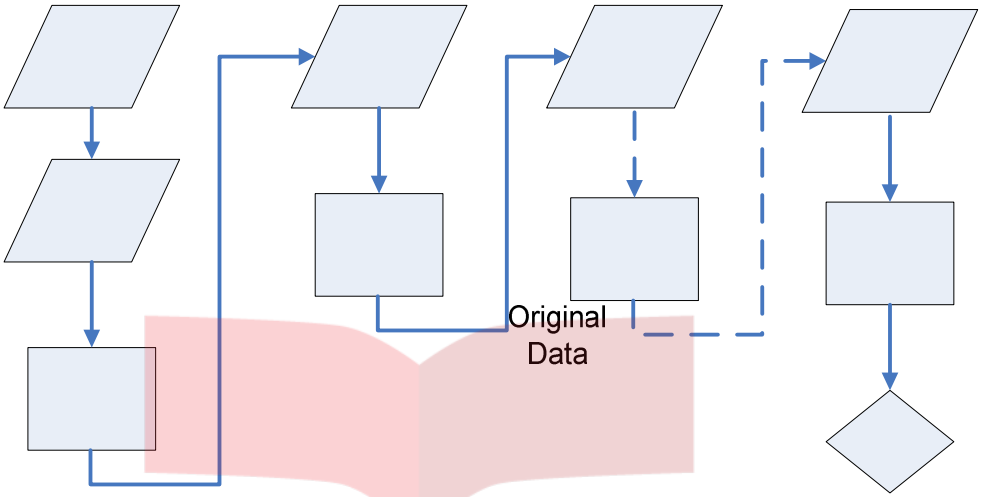
Salah satu cara untuk memprediksi data adalah dengan metode klasifikasi. Klasifikasi adalah suatu metode untuk memprediksi kelas dari suatu data yang belum diketahui sebelumnya dengan mengacu pada model yang telah dibuat dari sekumpulan data *training*. Metode klasifikasi kurang tepat untuk digunakan dalam kasus *imbalance class* karena hanya sebatas untuk kasus yang distribusi kelas labelnya seimbang. Untuk itu perlu dikembangkan suatu pendekatan untuk meningkatkan hasil performansi dari metode klasifikasi tersebut.

Salah satu cara untuk menangani kasus *imbalance class* adalah dengan menggunakan algoritma RareBoost-1. Dibandingkan algoritma lainnya seperti AdaBoost, algoritma ini dapat meningkatkan performansi yaitu keakuratan prediksi *minority class* dalam kasus *imbalance class*[6]. Dalam proses pelatihannya, kelas yang jumlahnya lebih sedikit akan mendapatkan perhatian lebih sehingga diharapkan keseluruhan kelas yang jumlahnya sedikit dapat diprediksikan secara maksimal. Analisa lebih dalam mengenai algoritma RareBoost-1 akan dibahas dalam Tugas Akhir ini.

1.2 Perumusan Masalah

Boosting merupakan salah satu *ensemble-method* yang bisa digunakan untuk meningkatkan performansi klasifikasi dari *classifier*. Salah satu penanganan kasus *imbalance class* adalah dengan menggunakan algoritma RareBoost-1. RareBoost-1 merupakan salah satu algoritma *Boosting* untuk menangani *minority class* sehingga keakuratan hasil prediksi mendekati akurat.

Gambaran umum mengenai Boosting dapat dilihat pada Gambar 1.1 :



Gambar 1-1: Boosting

Penjelasan dari Gambar 1.1 adalah sebagai berikut :

1. *Original* data dibagi menjadi *training* data dan *test* data.
2. Setelah itu metode klasifikasi digunakan untuk memproses *training* data agar terbentuk sebuah model yang nantinya akan digunakan untuk memprediksi *test* data.
3. Untuk memperoleh hasil prediksi yang akurat diperlukan beberapa iterasi dan di dalam setiap iterasi digunakan algoritma RareBoost-1.
4. Setelah itu diadakan voting untuk menentukan hasil prediksi dan hasil dari voting tersebut akan digunakan untuk menghitung performansi dilihat dari parameter *Recall, Precision,* dan *F-Measure.*

Berdasarkan penjelasan tersebut maka dapat dirumuskan beberapa permasalahan diantaranya :

1. Bagaimana mengimplementasikan algoritma RareBoost-1 untuk menangani kasus *imbalance class*.
2. Bagaimana performansi algoritma RareBoost-1 untuk menangani kasus *imbalance class*.

Sedangkan batasan masalah dari pembahasan Tugas Akhir ini adalah :

1. Data yang digunakan untuk analisa adalah data yang *supervised* (memiliki *class* label)
2. Tidak menangani tahap *preprocessing*, data latih dan data uji telah bersih dari *noise*.
3. Data yang ditangani hanya 2 (dua) *class problem*.
4. Data yang dianalisis yaitu data riil yaitu data Churn, data UCI, dan data PAKDD 2006.
5. Algoritma yang digunakan adalah RareBoost-1.

1.3 Tujuan

Tujuan pembuatan Tugas Akhir ini adalah :

1. Membuat sebuah perangkat lunak dengan mengimplementasikan algoritma RareBoost-1 untuk mendeteksi *minority class* pada kasus *imbalance class.*
2. Menganalisis performansi algoritma RareBoost-1 pada kasus *imbalance class* dengan menggunakan parameter *Recall, Presicion dan F-Measure.*

1.4 Metodologi Penyelesaian Masalah

Metodeologi yang akan digunakan untuk menyelesaikan tugas akhir ini adalah:

1. Studi Literatur.
   Studi Literatur dengan mempelajari literatur-literatur yang relevan dengan permasalahan yang meliputi : melakukan studi pustaka dan referensi mengenai data mining, metode klasifikasi, *imbalance class*,, algoritma RareBoost-1, dan Boosting secara umum.
2. Pengumpulan Data.
   Mencari data yang akan digunakan sebagai studi kasus.
3. Analisis dan Perancangan Perangkat Lunak.
   Menganalisis permasalahan yang akan ditangani, menganalisis metode yang akan digunakan untuk menyelesaikan permasalahan.
4. Implementasi Sistem.
   Melakukan *coding* dengan membangun perangkat lunak dengan menggunakan algoritma RareBoost-1.
5. Pengujian Sistem dan Analisis Hasil.
   Melakukan proses pengujian terhadap data-data yang digunakan sebagai studi kasus sehingga keakuratan hasil dan performansi algoritma RareBoost-1 dapat terlihat, serta melakukan analisis terhadap kelebihan dan keterbatasan algoritma RareBoost-1 pada permasalahan *imbalance class*.
6. Penyusunan laporan tugas akhir dan kesimpulan akhir.

5. Kesimpulan dan Saran

5.1 Kesimpulan

1. Algoritma klasifikasi biasa bisa menangani kasus *imbalance class* jika besar dat yang *imbalance* tidak terlalu besar (jumlah kelas minor dan mayor tidak terlalu jauh) .
2. Permasalahan *imbalance class* dapat dipecahkan dengan algoritma RareBoost-1 yang digabungkan dengan algoritma klasifikasi biasa.
3. Akurasi yang dihasilkan dengan menggunakan algoritma RareBoost-1 terhadap pengujian data yang imbalance menurun dibandingkan dengan akurasi yang dihasilkan oleh algoritma klasifikasi biasa. Hal ini dikarenakan kemampuan algoritma RareBoost-1 dalam memprediksi *minority class* tetapi sedikit mengorbankan prediksi *mayority class* (Recall semakin tinggi tetapi Precision turun).
4. Perbandingan *imbalance class* dari setiap data sangat mempengaruhi hasil performansi. Semakin besar *imbalance* (persentase *minority class* semakin kecil) maka hasil performansi akan semakin kecil.
5. Banyaknya iterasi mempengaruhi hasil performansi. Semakin banyak jumlah iterasi maka semakin baik performansi yang dihasilkan meskipun fluktuasi tetapi masih cenderung naik.
6. Banyaknya *oversampling* mempengaruhi hasil performansi, khususnya nilai *Recall*. Semakin banyak *oversampling* maka semakin tinggi nilai *Recall,* tetapi akan mengorbankan nilai *Precision* dan *F-Measure.*
7. Nilai αtp dan αtn yang dihasilkan oleh algoritma RareBoost-1 sangat berperan dalam hasil voting yang nantinya akan mempengaruhi hasil performansi. Semakin besar nilai αtp dan nilai αtn kecil maka *minority class* akan terprediksi dengan baik. Sebaliknya, semakin besar nilai αtn dan nilai αtp kecil maka *minority class* tidak akan terprediksi dengan baik.

5.2 Saran

1. Untuk penelitian selanjutnya, aplikasi yang dihasilkan dari Tugas Akhir ini lebih disempurnakan sehingga dapat menangani data yang memiliki lebih dari 2 (dua) class problem.

Daftar Pustaka

[1] Buckland, M, F. Gey, The Relationship Between Recall and Precision, *Journal of the American Society for Information Science*, 45(1):12--19, 1994.

[2] Jo, Taeho dan N. Japkowicz. *Class Imbalance versus Small Disjuncts.* University of Otawa.

[3] Kumar, Vipin, Pang Nim Tan, Michael Steinbach. *Introduction to Data Mining.* Michigan State University, University of Minnesota.

[4] Kumar, Vipin, Aleksandar Lazarevic, Jaideep Srivastava, *Data Mining for Analysis Rare Event : A Case Study in Security, Financial, and Medical Applications.* Department Computer Science University of Minnesota, 2004.

[5] Kumar, Vipin, Aleksandar Lazarevic, Jaideep Srivastava. *Data Mining for Rare Class Analysis.* Department Computer Science University of Minnesota.

[6] Kumar, Vipin, M. Joshi, R. Agarwal. *Evaluating Boosting Algorithms to Classify Rare Classes: Comparison and Improvements*. First IEEE International Conference on Data Mining, San Jose, CA, 2001.

[7] Kumar, Vipin, M. Joshi, R. Agarwal. *Predicting Rare Classes: Can Boosting Make Any Weak Learner Strong?.* University of Minnesota, San Jose, CA.2002.

[8] Novitasari, Dewi. Tugas Akhir. *Analisis Klasifikasi Algoritma Credos Pada Kasus Imbalanced Class*. Jurusan Teknik Informatika STTTelkom Bandung, 2006.

[9] Weiss, Gary M. *Mining with Rare Cases*. Department of Computer and Information Science Fordham University.

[10] Weiss, Garry M. Mining with Rarity : A Unifying Framework. AT&T Laboratories.

[11] Yun Jung, Alexander. Thesis. The Effect of Oversampling and Undersampling on Classifying Imbalanced Text Dataset. The University of Texas at Austin. 2004.