IMPLEMENTASI ALGORITMA SMOTEBOOST PADA KASUS IMBALANCE CLASS

Ifa Saptina Rani[1], Moch. Arif Bijaksana[2], Rimba Widhiana Ciptasari[3]

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

Abstrak

Imbalance class adalah ketidakseimbangan distribusi class label pada suatu data set. Dalam data mining berbagai penelitian telah dilakukan untuk mengatasi permasalahan imbalance class tersebut, salah satunya adalah algoritma SMOTEBoost. SMOTEBoost merupakan kombinasi dari algoritma SMOTE (Synthetic Minority Over-sampling Technique) dengan teknik boosting. Algoritma SMOTE mengcreate sejumlah instant synthetic dari minority class yang menyebabkan wilayah minority class semakin besar. Boosting adalah suatu teknik yang mengombinasikan hasil prediksi dari beberapa classifier yang berbeda.

Pada algoritma SMOTEBoost dilakukan oversampling dengan menggunakan algoritma SMOTE 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 akhir dari pengujian menunjukkan bahwa penggunaan algoritma SMOTE dapat meningkatkan prediksi True Positif (kenaikan Recall) dan teknik boosting dapat meminimalkan nilai False Positif (kenaikan precision)

Kata Kunci : Imbalance class, SMOTEBoost, voting, pembobotan, oversampling

Abstract

Imbalance class represents imbalance class label distribution in number of dataset. In data mining, many research have been done to handle imbalance class problems, one of that is SMOTEBoost algorithm. SMOTEBoost is combines of SMOTE (Synthetic Minority Over-sampling Technique) algorithm with boosting technique. SMOTE algorithm create some of synthetic instant from minority class that caused minority class region larger. Boosting is one of technique that combine result of predict some different classifier.

In SMOTEBOost algorithm, it done oversampling with SMOTE algorithm 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 SMOTEBoost algorithm can increase predict true positif (increase recall) and boosting can minimalize false positif (increase precision) .

Keywords : Imbalance class, SMOTEBoost, voting, weighted, oversampling

1. Pendahuluan

1.1 Latar belakang

Imbalance class merupakan ketidakseimbangan distribusi *class label* pada suatu data latih. Karakteristik dari *imbalance class* adalah pada salah satu classnya direpresentasikan dengan jumlah data yang sangat besar (*mayority class)* sedangkan class yang lainya direpresentasikan dengan jumlah yang sangat kecil (*minority class*). *Imbalance class* biasanya ditemukan pada kasus-kasus anomali misalnya *fraud detection, network intrusion detection, curn prediction,* dan *thyroid desease.* Kasus-kasus seperti itu kalau tidak segera diatasi bisa menyebabkan kerugian finansial, seperti pada industri telekomunikasi Amerika Serikat kerugian yang diakibatkan fraud antara 4%-6% dari pendapatan [2].

Imbalance class dalam konteks data mining perlu dipelajari karena *minority class* lebih sulit untuk diprediksi daripada *mayority class*. Padahal, terkadang *class* yang minoritas inilah yang mempunyai informasi yang sangat berharga. Oleh karena itu, bagaimana memprediksi *class label* yang tepat sangat diperlukan agar kita memperoleh hasil dengan nilai *true positif* yang tinggi.

Pada tugas akhir ini, permasalahan imbalance class akan coba ditangani dengan algoritma SMOTEBoost (*Synthetic Minority Over-sampling TEchnique* yang dikombinasikan dengan teknik boosting) oleh karena algoritma SMOTE dapat meningkatkan akurasi dari *minority class* dan penggunaan boosting tidak mengorbankan akurasi dari data set secara keseluruhan [2].

1.2 Perumusan masalah

Dari uraian latar belakang diatas maka dapat dirumuskan beberapa permasalahan yang harus diselesaikan yaitu :

1. Bagaimana memprediksi *minority class* pada kasus *imbalace class*.
2. Bagaimana menentukan prediksi *class label* yang tepat sehingga didapatkan nilai *True Positif* yang tinggi dan *False Positif* yang kecil.
3. Bagaimanakah performansi algoritma SMOTEBoost pada kasus *imbalance class*.

Sedangkan batasan masalah dari pembahasan Tugas Akhir ini diantaranya :

1. Data yang dianalisis yaitu data *Thyroid Desease (sick)*, data PAKDD 2006, data operator ilegal Telkom pada perioda tertentu, dan data prediksi *churn* PT. Telkom.
2. Tidak membahas tahap *preprocessing* secara detail.

1.3 Tujuan

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

1. Membangun sebuah perangkat lunak dengan mengimplementasikan algoritma SMOTEBoost untuk memprediksi *minority class* pada kasus *imbalanced class.*
2. Menerapkan algoritma SMOTE pada pendekatan Boosting, untuk mendapatkan hasil *True Positif* (*recall*) yang tinggi dan nilai *False Positif* yang kecil.
3. Menganalisis perfomansi algoritma SMOTEBoost dengan menggunakan parameter *recall, precision,* 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 : studi pustaka dan referensi mengenai data mining, klasifikasi, *imbalanced class,* SMOTE, dan Boosting.
2. Pengumpulan Data.
   Mencari data yang akan digunakan sebagai studi kasus.
3. Analisis dan Perancangan Perangkat Lunak.
   Menganalisis dan merancang perangkat lunak yang akan digunakan.
4. Implementasi system.
   Mengimplementasikan perangkat lunak yang telah didesain.
5. Pengujian sistem dan analisis hasil.
   Melakukan proses pengujian terhadap keakuratan hasil dan performansi algoritma SMOTEBoost pada permasalahan *imbalanced class* berdasarkan parameter *recall, precision,* dan *F-Measure.*
6. Penyusunan laporan tugas akhir dan kesimpulan akhir.

5. Kesimpulan dan Saran

5.1 Kesimpulan

1. Algoritma SMOTEBoost mempunyai performansi paling baik dibandingkan dengan algoritma C5 , SMOTE+C5, dan AdaBoost.M1.
2. Penggunaan algoritma SMOTE dapat meningkatkan prediksi *True Positif* (kenaikan *Recall*) sedangkan teknik boosting digunakan untuk meminimalkan nilai *False Positif* (kenaikan *precision).*
3. Perbandingan *imbalance class* dari setiap data sangat mempengaruhi hasil performansi. Semakin besar *imbalance* (persentase *minority class* semakin kecil) maka hasil performansi akan semakin kecil.
4. Tidak ada parameter SMOTE (N) dan jumlah iterasi (T) yang specific untuk menentukan nilai performansi yang optimal. Semakin besar parameter SMOTE (*N*) maka nilai *recall* akan semakin naik, tetapi nilai *Precision* yang didapatkan semakin turun. Semakin banyak jumlah iterasi (T) maka nilai recall akan cenderung naik dan nilai *precision* akan cenderung turun.

5.2 Saran

1. Untuk penelitian selanjutnya, lebih difokuskan pada analisis besarnya prosentase inputan jumlah *minor class* dan *mayor class* yang terbaik.
2. Untuk penelitian selanjutnya, lebih difokuskan pada pengaruh nilai-nilai atribut pada besarnya performansi algoritma SMOTEBoost.

Daftar Pustaka

[1] A. Joshi. *Applying the Wrapper Approach for Auto Discovery of Under-Sampling and Over-Sampling Percentages on Skewed Datasets*. Dept. of Computer Science and Eng., University of South Florida MS Thesis, 2004.

[2] Chawla, N, A. Lazarevic, L. O. Hall, and K. W. Bowyer. SMOTEBoost: Improving Prediction of the Minority Class in Boosting. *7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), pp. 107 to 119, Dubrovnik, Croatia,2003.*

[3] Domingos, P. (1999). "*MetaCost: A general method for making classifiers cost-sensitive*.Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining, pp. 155-164. ACM Press.

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

[5] Sotiris, K, K Dimitris, P. Panayotis. *Handling Imbalanced Dataset : A review*. Educational Software Development Laboratory, Departement of Mathematic, Univercity of Patras, Grece.

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

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

[8] Z. Zheng, X. Wu, and R. Srihari. *Feature selection for text categorization on imbalanceddata.* SIGKDD Explorations, 6(1):80-89, 2004.