Expert Systems with Applications 42 (2015) 1074–1082

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction

Myoung-Jong Kim [a], Dae-Ki Kang [b,*], Hong Bae Kim [c]

[a] School of Business, Pusan National University, 63 Beon-gil 2, Busandaehag-ro, Geumjeong-gu, Busan 609-735, Republic of Korea
[b] Department of Computer and Information Engineering, Dongseo University, 47, Churye-Ro, Sasang-Gu, Busan 617-716, Republic of Korea
[c] Division of Business, Dongseo University, 47, Churye-Ro, Sasang-Gu, Busan 617-716, Republic of Korea

ARTICLE INFO

Article history:
Available online 16 September 2014

Keywords:
Data imbalance
Bankruptcy prediction
Over-sampling
SMOTE
Cost-sensitive boosting
AdaBoost
GMBoost

ABSTRACT

In classification or prediction tasks, data imbalance problem is frequently observed when most of instances belong to one majority class. Data imbalance problem has received considerable attention in machine learning community because it is one of the main causes that degrade the performance of classifiers or predictors. In this paper, we propose geometric mean based boosting algorithm (GMBoost) to resolve data imbalance problem. GMBoost enables learning with consideration of both majority and minority classes because it uses the geometric mean of both classes in error rate and accuracy calculation. To evaluate the performance of GMBoost, we have applied GMBoost to bankruptcy prediction task. The results and their comparative analysis with AdaBoost and cost-sensitive boosting indicate that GMBoost has the advantages of high prediction power and robust learning capability in imbalanced data as well as balanced data distribution.

1. Introduction

Data imbalance problem is frequently observed in various classification and prediction tasks when most of training instances belong to one majority class. Although most classification algorithms are trained under the assumption that the ratio of the classes is almost equal, this assumption is frequently violated in real world classification tasks. Data imbalance problem is reported in a wide range of classification tasks, such as oil spill detection (Kubat, Holte, & Matwin, 1998), response modeling (Shin & Cho, 2006), remote sensing (Bruzzone & Serpico, 1997), quality assessment of sensor data (Rahman, Smith, & Timms, 2013) and scene classification (Yan, Liu, Jin, & Hauptmann, 2003). It is also pervasive in business applications including card fraud detection (Fawcett & Provost, 1997) and credit rating (Kwon, Han, & Lee, 1997).

Data imbalance problem has received considerable attention in the machine learning community because it is one of the main causes that degrade the performance of machine learning algorithms in classification tasks. There are two main reasons why data imbalance problem causes degradation in performance of machine learning algorithms (Kang & Cho, 2006; Kotsiantis, Tzelepis, Koumanakos, & Tampakas, 2007; Wang & Japkowicz, 2008).

The first reason is associated with the objective function of classification algorithms. One of widely used objective functions for classification algorithms is the arithmetic mean based accuracy (hereafter, arithmetic accuracy) which is a ratio of the number of correctly classified instances over the number of total instances. However, in the presence of data imbalance, arithmetic accuracy might be inappropriate because the accuracy is highly dependent on the classification accuracy of majority class instances. For example, bankruptcy is a very rare event. Credit rating agencies such as Moody's anticipate long term average default rates of Korean audited companies to be about three to five percent. Arithmetic accuracy of the generated classifier will tend to be abnormally high due to the high accuracy for majority class instances (non-bankrupt companies) despite the low accuracy for minority class instances (bankrupt companies) in a configuration when all of audited companies are used as a training data set. To be more specific, in very imbalanced domains, most standard classifiers will tend to learn how to predict the majority class. While these classifiers can obtain higher predictive accuracies than those that also try to consider the minority class more, this seemingly good performance can be argued as being meaningless (Wang & Japkowicz, 2008).

There have been research works to apply receiver operating characteristic (ROC) curve or geometric mean based accuracy (hereafter, geometric accuracy) in measuring performance, because these measures have advantages of reflecting both the

* Corresponding author. Tel.: +82 51 320 1724; fax: +82 51 327 8955.
E-mail addresses: mjongkim@pusan.ac.kr (M.-J. Kim), dkkang@dongseo.ac.kr (D.-K. Kang), rfctogether@gmail.com (H.B. Kim).

http://dx.doi.org/10.1016/j.eswa.2014.08.025
0957-4174/

accuracy on the majority and minority classes at the same time (Fawcett, 2006; Kubat, Holte, & Matwin, 1997).

The second reason for the degradation in performance is the distortion of decision boundaries resulting from imbalanced distribution of the classes. As the imbalance of data is getting severe, the decision (classification) boundary of majority class tends to invade the decision boundary of the minority class, so that the decision boundary of majority class is gradually expanded while the decision boundary of minority class is gradually reduced. This problem eventually causes the decrease in the accuracy for minority class.

For the alternatives to solve this problem, various methods have been proposed including under-sampling, over-sampling, cost-sensitive algorithms, and boosting algorithms. Under-sampling method decreases the number of samples from majority class to that of minority class in order to make the number of both classes the same. Over-sampling method, which is opposite to under-sampling method, increases the number of samples in minority class to meet the number of samples in majority class. In cost-sensitive algorithms, the penalty is assigned to misclassified instances from minority class (Elkan, 2001; Provost & Fawcett, 2001). Boosting algorithm is one of recently proposed techniques that can be used to resolve data imbalance problem because it provides more learning opportunities for minority class samples which is more likely to be misclassified than majority class sample. Recently, hybrid algorithms of data sampling and boosting have been proposed as alternatives for data imbalance problem including SMOTEBoost (Chawla, Lazarevic, Hall, & Bowyer, 2003), RUSBoost (Seiffert, Khoshgoftaar, Hulse, & Napolitano, 2008) and cost-sensitive boosting (Sun, Kamel, Wong, & Yang Wang, 2007; Ting, 2000; Zhang & Wang, 2013;Zheng, 2010). In particular, SMOTEBoost, which is used in the experiments of this research, is an application of boosting techniques to over-sampled data generated by synthetic minority over-sampling technique (SMOTE) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). SMOTEBoost has shown promising results in resolving data imbalance problem. However, it might suffer from over-fitting problem. New minority class samples, which are generated from SMOTE, are likely to have the higher similarity than majority data samples. Most standard learning algorithms, including boosting algorithm, will tend to generate classifiers focusing on samples with higher similarity because that strategy is helpful to maximize the objective function (i.e. arithmetic accuracy). This drawback might increase generalization errors when classifiers are applied to new validation data set which is not trained (Drummond et al., 2003; Seiffert et al., 2008).

This paper proposes a novel boosting algorithm called geometric mean based boosting (GMBoost) to resolve data imbalance problem. GMBoost is a modification of AdaBoost algorithm (Freund & Schapire, 1997) that replace arithmetic error and accuracy calculation of AdaBoost with the concept of geometric error and accuracy calculation. It has the advantage of enabling balanced learning against both majority and minority classes. The proposed GMBoost algorithm is applied to bankruptcy prediction task which is one of the typical data imbalance problems in business domains. Two different data samples are constructed to verify the performance of GMBoost algorithm. At the first stage, five sample groups are constructed according to different data balance rates (1:1(denoted as A), 1:3(B), 1:5(C), 1:10(D), and 1:20(E)) and perform classification experiments using AdaBoost, cost-sensitive boosting and GMBoost for the performance comparison in imbalanced data.

At the second stage, SMOTE algorithm is applied to generate new bankrupt company data sets for B, C, D, and E of the first stage, and thus bankrupt companies to normal companies are in the ratio of 1:1. We apply the newly sampled sets to AdaBoost, cost-sensitive boosting and GMBoost for the performance comparison in balanced data.

Experimental results show that GMBoost has the advantages of high prediction power and robust learning capability over AdaBoost and cost-sensitive boosting in imbalanced data distribution as well as in balanced data distribution. These results on actual bankruptcy data mean that the usefulness and effectiveness of GMBoost are promising on broader range of real-world problems of decision making.

This paper is organized as follows. The problems of data imbalance and the previous methods to solve these problems are briefly described in Section 2. Four algorithms including SMOTE, cost-sensitive support vector machines(SVM), AdaBoost and GMBoost, which are used in this research, are explained in Section 3. In Section 4, we explain the processes of data collection and experimental design. Experimental results are presented in Section 5. We conclude with future research directions in Section 6.

2. Data imbalance problem in binary classification problems

In this section, we will describe data imbalance problems and then existing methods to resolve data imbalance problem.

2.1. Data imbalance problem

Kang and Cho (2006) constructed six sample groups according to different data balance rates (1:1, 1:3, 1:5, 1:10, 1:30, and 1:50) in order to analyze the effects of data imbalance on classification accuracy of SVM. Their experimental results show that for the two sample groups with little or no data imbalance problem (1:1, and 1:3), the sizes of decision boundary areas of the two classes are similar to each other. However, for the sample groups with serious data imbalance problems (1:5, and 1:10), the area of minority class is reduced because the area of the majority class invades the area of minority class, and thus the classification accuracy for minority class samples gets degraded. In particular, for the sample groups with extreme data imbalance (1:30, and 1:50), the decision boundary area for minority class is excessively small, which makes the classification for minority class meaningless. Also, they reported that, as data imbalance is getting severe, arithmetic accuracy over total samples steadily increases due to the high accuracy over samples of majority class, while the arithmetic accuracy for minority class is dramatically reduced, and thereby geometric accuracy over total samples gradually decreases. They argued that these results demonstrate that arithmetic accuracy is not a suitable objective function for imbalanced data.

Wu and Chang (2003) asserted that data imbalance leads to skewing the boundaries of SVM. One reason is that the decision boundary area of majority class is expanded and the decision boundary area of minority class is reduced, as data imbalance is getting severe. This problem causes the distortion of decision boundary area. Another reason is that data imbalance induces that samples of minority class do not reside in the decision boundary area of minority class, as the decision boundary area of minority class is getting small. Consequently, the possibility becomes very high that the classifier will classify a sample as a majority class.

2.2. The approaches to resolve performance measure problem

Let us assign minority class as positive and majority class as negative in order to explain the concept of accuracy. Simple arithmetic mean based accuracy which is widely used is calculated as $\frac{(TP+TN)}{(TP+FN+FP+TN)}$ as shown in the confusion matrix as of Table 1. As described before, arithmetic accuracy is a proper performance measure for classifiers in balanced data set. However, under data imbalance, it is an improper performance measure because it is highly influenced by the classification accuracy of majority class