Neural Networks 96 (2017) 101–114

Contents lists available at ScienceDirect

Neural Networks

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

Post-boosting of classification boundary for imbalanced data using geometric mean

Jie Du [a], Chi-Man Vong [a,*], Chi-Man Pun [a], Pak-Kin Wong [b], Weng-Fai Ip [c]

[a] Department of Computer and Information Science, University of Macau, Macau
[b] Department of Electromechanical Engineering, University of Macau, Macau
[c] Faculty of Science and Technology, University of Macau, Macau

ARTICLE INFO

Article history:
Received 27 April 2016
Received in revised form 21 June 2017
Accepted 5 September 2017
Available online 14 September 2017

Keywords:
Imbalance learning
Boosting
Weighted ELM
SMOTE

ABSTRACT

In this paper, a novel imbalance learning method for binary classes is proposed, named as *Post-Boosting of classification boundary for Imbalanced data* (PBI), which can significantly improve the performance of any trained neural networks (NN) classification boundary. The procedure of PBI simply consists of two steps: an (imbalanced) NN learning method is first applied to produce a classification boundary, which is then adjusted by PBI under the *geometric mean* (G-mean). For data imbalance, the *geometric mean* of the accuracies of both minority and majority classes is considered, that is statistically more suitable than the common metric *accuracy*. PBI also has the following advantages over traditional imbalance methods: (i) PBI can significantly improve the classification accuracy on minority class while improving or keeping that on majority class as well; (ii) PBI is suitable for large data even with high imbalance ratio (up to 0.001). For evaluation of (i), a new metric called *Majority loss/Minority advance ratio* (MMR) is proposed that evaluates the loss ratio of majority class to minority class. Experiments have been conducted for PBI and several imbalance learning methods over benchmark datasets of different sizes, different imbalance ratios, and different dimensionalities. By analyzing the experimental results, PBI is shown to outperform other imbalance learning methods on almost all datasets.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

DATA imbalance is a classical and challenging problem in many practical applications in which the rarely occurring instances are critical and interested. For example, in fraud detection (Phua, Alahakoon, & Lee, 2004), the interested class (called minority) are "fraudulent" cases which appear far less frequently than "non-fraudulent" cases (called majority). In medical diagnosis (Mazurowski et al., 2008), the number of cancerous cases is much smaller than that of normal ones. Other examples are also shown in the literature (Wang & Japkowicz, 2004) including helicopter gearbox fault monitoring, discrimination between earthquakes and nuclear explosions, document filtering, and detection of oil spills. The common issue of these applications is that the distribution of data is *highly* imbalanced. Consequently, the detection or classification of rarely occurring instances in imbalanced data is of primary interest but non-trivial.

Imbalance learning is then proposed for this kind of data imbalance problem, which refers to the procedure to construct a model, mostly with machine learning techniques, that accurately classifies the minority in imbalanced data. Although many practical applications only require binary classification, some exceptions such as document filtering belong to multiclass classification. Nevertheless, there are many strategies to decompose a multiclass classification into multiple binary classifications, such as one-against-all (Kumar & Gopal, 2011), one-against-one (Kang, Cho, & Kang, 2015), and their combined version: all-and-one (A&O) approach (Ghanem, Venkatesh, & West, 2010). From this viewpoint, only binary classification for imbalanced data is considered in this paper.

In the literature (Han, Kamber, & Pei, 2011), there are many machine learning methods for binary classification: decision tree induction, Bayes method, rule-based classification, and Neural Networks (NN). Among these methods, NN is widely adopted because of its high effectiveness and efficiency. However, NN is not natively designed for imbalanced data and in fact it is very sensitive to the distribution of data (Zhou & Liu, 2006). For highly imbalanced data, NN will misclassify almost all minority instances to majority because the objective function of NN is only designed to correctly classify as many instances as possible, regardless the instance is minority or majority (Zhou & Liu, 2006). As a result, almost all majority instances are correctly classified while almost all interested minority instances are ignored. Therefore, the objective of

* Corresponding author.
E-mail addresses: yb57415@umac.mo (J. Du), cmvong@umac.mo (C.-M. Vong), cmpun@umac.mo (C.-M. Pun), fstpkw@umac.mo (P.-K. Wong), andyip@umac.mo (W.-F. Ip).

http://dx.doi.org/10.1016/j.neunet.2017.09.004
0893-6080/© 2017 Elsevier Ltd. All rights reserved.

imbalance learning is to improve the classification accuracy for minority (i.e., positive class) while maintaining that for majority (i.e., negative class). ***Remark*:** In the remainder of the paper, we will use "*positive class*" (+) to refer to minority and "*negative class*" (−) to majority, because minority is of primary interest and hence positive.

There are many methods (Cano, Zafra, & Ventura, 2013; Chawla, Bowyer, & Hall, 2002; Gao, Chen, Tang, Zhang, & Li, 2016; Hong, Chen, & Harris, 2007; Li, Kong, Lu, Wenyin, & Yin, 2014; Nanni, Fantozzi, & Lazzarini, 2015; Pang et al., 2013; Sharma & Kiet, 2015; Zheng, Zhang, Chen, Liu, Lu, & Sun, 2013; Zong, Huang, & Chen, 2013) to address the imbalance problem, but can be generally classified into two ways:

(i) modifying the objective function;
(ii) resampling the training data.

The first way is to assign different costs or weights to different classes in the objective function such that the weight associated with the positive class is relatively larger than that of negative class. A very intuitive method (Zong et al., 2013) is to use the reciprocal of the quantity of instances in one class as the weight of this class. However, this weighting scheme is simple but slightly ineffective because the search space is restricted for finding optimal weights (Sharma & Kiet, 2015). In the literature (Sharma & Kiet, 2015; Zheng et al., 2013), other ways are proposed to optimize the weights, all of which however still employs *accuracy* (which is unsuitable in imbalance data) as the objective function, resulting to only little improvement upon the intuitive method in Zong et al. (2013).

The second way is to preprocess the imbalanced training data such that the numbers of positive and negative instances become approximately the same so that a balance data ratio is achieved. In general, there are three schemes of resampling techniques including the following:

(i) over-sampling the minority class;
(ii) under-sampling the majority class, and
(iii) synthetic minority over-sampling technique (SMOTE) (Chawla et al., 2002).

In simple terms, over-sampling directly duplicates the minority instances to balance the dataset while under-sampling removes some redundant majority instances from the dataset. However, over-sampling may cause over-fitting because the classifier learns a very specific model from many copies of the same instance. On the other hand, the problem of under-sampling is relatively obvious that it may cause significant information loss from the removed negative instances. In other words, the classification accuracy for the majority class is degenerated for these two schemes. SMOTE is also an over-sampling scheme which *artificially* generates some synthetic minority instances along lines in feature space between the minority instance and its selected k nearest minority neighbors, rather than simply duplication. In these three resampling schemes, SMOTE currently yields the best result (Chawla et al., 2002). Nevertheless, SMOTE may still suffer from over-generalization since the synthetic minority instances are blindly generated without considering if the neighboring instances belong to minority or majority, which might cause overlapping between classes.

To effectively resolve the imbalance classification, its nature and fundamental issues must be addressed. In general, *accuracy* is the most popular metric for classification effectiveness, which is simply calculated as the ratio of *the number of correctly classified instances* to *the total number of instances*. However, accuracy is unsuitable to evaluate the effectiveness of imbalance learning as shown in the following example. In a test set of 100 instances (5 positive, 95 negative), assume there are 5 instances misclassified in both classes respectively. The accuracy is calculated as $\frac{(5-5)+(95-5)}{5+95} = 90\%$. Obviously, even though the instances in positive class are totally misclassified, an excellent accuracy of 90% is still achieved. Alternatively, a common choice of evaluation metric in imbalance problem is the *Geometric mean* (G-mean) (He & Ma, 2013), which is the geometric mean of accuracy for each class. The G-mean considers both positive and negative accuracies and is calculated as the square root of the product of positive and negative accuracies. Under the G-mean, the evaluation for this example is $\sqrt{\frac{5-5}{5} \times \frac{100-5}{100}} = 0\%$. In other words, the G-mean reveals a more realistic situation as long as there is one class mostly misclassified. Therefore, G-mean is more suitable than accuracy to evaluate the classification effectiveness in imbalanced data. By maximizing the G-mean of a classifier, an optimal classification boundary for imbalanced data can be obtained.

Clearly, the G-mean metric is an effective utility to learn the classification boundary of a NN classifier so that *both* positive and negative instances can be classified as accurately as possible. The first intuition is to simply include the G-mean metric in the objective function and derive a "new" NN classifier. Unfortunately, the G-mean consists of a piecewise function *sign* which is non-differentiable. For this reason, an estimated differentiable G-mean (i.e., with *sign* removed) can be included in the objective function. However, the optimized boundary based on this estimated G-mean is with lower classification effectiveness (detailed in Section 3.2). From this inspiration, we propose a way to *post-boost* the classification boundary under the G-mean to improve the effectiveness over *both* positive and negative instances. This proposed post-processing method is called *Post-Boosting of classification boundary for Imbalanced data (PBI)*. The main contributions of PBI are enumerated as follows:

(i) PBI can be generically applied on all kinds of boundary-based NN for binary imbalanced data to improve their classification effectiveness under the G-mean.
(ii) Different from traditional classification methods whose objectives are to minimize the training error regardless of positive and negatives classes, the G-mean based PBI treats both positive and negative instances equally, even though the number of positive instances is much smaller than that of negative ones.
(iii) Different from traditional imbalanced methods, PBI *automatically* learns the classification boundary over training data instead of *artificially* adjusting the weights in weighting scheme and make up the shortcomings of resampling methods.

In this paper, the proposed PBI was applied to both weighting and resampling schemes for imbalanced data. Experiments under datasets with different sizes, imbalance ratios, and dimensionalities were conducted to demonstrate the effectiveness of PBI. Among various NN methods, random projection algorithms (Chen & Wan, 1999) such as *extreme learning machine* (ELM) (Huang, Zhou, Ding, & Zhang, 2012; Huang, Zhu, & Siew, 2006) are employed because of its high efficiency and accuracy for different size of data. Therefore, ELM was selected as the demonstrative NN classifier in our work. In addition, PBI can be integrated with other resampling schemes such as SMOTE to post-boost the classification boundary. In order to have a better and fair comparison on the effectiveness of PBI, a novel evaluation metric called *MMR* (Majority loss/Minority advance Ratio) or its general form *gMMR* is proposed. *MMR*/*gMMR* indicates the loss of classification effectiveness for negative class when increasing the classification effectiveness for positive class, which actually reflects a common and severe issue suffered by all imbalance learning methods.

Download English Version:

https://daneshyari.com/en/article/4946561

Download Persian Version:

https://daneshyari.com/article/4946561

Daneshyari.com