Pattern Recognition 48 (2015) 1623–1637

Contents lists available at ScienceDirect

Pattern Recognition

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

A novel ensemble method for classifying imbalanced data

Zhongbin Sun [a], Qinbao Song [a,*], Xiaoyan Zhu [a], Heli Sun [a], Baowen Xu [b], Yuming Zhou [b]

[a] *Dept. of Computer Science & Technology, Xi'an Jiaotong University, China 710049*
[b] *Dept. of Computer Science & Technology, Nanjing University, China 210093*

ARTICLE INFO

Article history:
Received 2 January 2014
Received in revised form
18 November 2014
Accepted 22 November 2014
Available online 29 November 2014

Keywords:
Imbalanced data
Classification
Ensemble learning

ABSTRACT

The class imbalance problems have been reported to severely hinder classification performance of many standard learning algorithms, and have attracted a great deal of attention from researchers of different fields. Therefore, a number of methods, such as sampling methods, cost-sensitive learning methods, and bagging and boosting based ensemble methods, have been proposed to solve these problems. However, these conventional class imbalance handling methods might suffer from the loss of potentially useful information, unexpected mistakes or increasing the likelihood of overfitting because they may alter the original data distribution. Thus we propose a novel ensemble method, which firstly converts an imbalanced data set into multiple balanced ones and then builds a number of classifiers on these multiple data with a specific classification algorithm. Finally, the classification results of these classifiers for new data are combined by a specific ensemble rule. In the empirical study, different class imbalance data handling methods including three conventional sampling methods, one cost-sensitive learning method, six Bagging and Boosting based ensemble methods, our previous method EM1vs1 and two fuzzy-rule based classification methods were compared with our method. The experimental results on 46 imbalanced data sets show that our proposed method is usually superior to the conventional imbalance data handling methods when solving the highly imbalanced problems.

1. Introduction

Class imbalance data refers to at least one of its classes is usually outnumbered by the other classes. The class imbalance problems have been reported to occur in a wide variety of real-world domains, such as facial age estimation [1], detecting oil spills from satellite images [2], anomaly detection [3], identifying fraudulent credit card transactions [4], software defect prediction [5], and image annotation [6]. Therefore, researchers have paid much attention to the class imbalance problems, and several thematic workshops and conferences were held, such as the Association for the Advancement of Artificial Intelligence (AAAI) 2000 [7], the International Conference on Machine Learning (ICML) 2003 [8], and the ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) Explorations 2004 [9].

For a binary class imbalance problem, the instances are usually categorized into majority and minority classes. Generally speaking, the minority class usually represents a concept with greater interest than the majority class. However, it is often outnumbered by the majority class, and sometimes this scenario may be very severe. As most traditional classification algorithms, such as decision trees [10–12], *k*-nearest neighbors [13,14], and RIPPER [15,16], tend to generate models which maximize the overall classification accuracy, and the minority class is usually ignored [17–19]. For example, for a data set where only 1% of the instances belong to the minority class, even if a model classifies all instances as the majority class, it still achieves an overall accuracy of 99%. However, the minority class instances, which we want to accurately classify, are all misclassified by this model though it achieves a very high accuracy. Therefore, a number of methods have been proposed to deal with the imbalanced binary classification problems.

One of the most popular methods for solving the class imbalance problems is sampling [20–23]. However, the most used under-sampling and over-sampling methods alter the original class distribution of imbalanced data by eliminating the majority class instances or increasing the minority class instances. In addition, cost-sensitive learning is also employed to solve the class imbalance problems [24–26]. This method assigns different cost of misclassification errors for different classes, generally high cost for the minority class and low cost for the majority class as the minority class is usually more interesting. Moreover, Bagging [27] and Boosting [28] based ensemble methods are another widely used methods to deal with imbalanced problems [15,16,29,30].

We argue that the above mentioned methods might encounter some unexpected problems when employed to solve the class imbalance problems. For instance, under-sampling methods might

[*] Corresponding author. Tel.: +86(0)29 82668645; fax: +86(0)29 82668971.
E-mail address: qbsong@mail.xjtu.edu.cn (Q. Song).

http://dx.doi.org/10.1016/j.patcog.2014.11.014
0031-3203/© 2014 Elsevier Ltd. All rights reserved.

abandon some potentially useful data which could be very important for a learning process, while over-sampling methods may increase the likelihood of overfitting in the induction process. Furthermore, Bagging and Boosting based ensemble methods may eliminate some useful data as they use sampling methods to obtain balanced data in each of their iteration procedure, and they may suffer from overfitting as well. In a word, these two kinds of methods both may alter the original class distribution of imbalanced data by adding minority class instances or deleting majority class instances. Moreover, for the cost-sensitive learning methods, it is difficult to obtain the accurate misclassification cost, and the different misclassification cost might result in different induction results. So the classification results are not stable.

This paper introduces a novel ensemble method for addressing binary-class imbalance problems. Our proposed method handles the class imbalance problems by converting a imbalanced binary learning process into multiple balanced learning processes, which does not make use of introducing minority class instances or removing majority class instances to get away from the imbalanced data. The proposed method firstly converts the imbalanced data into multiple balanced data using the data balancing methods random splitting (SplitBal) or clustering (ClusterBal). Then multiple classifiers could be built from these balanced data with a specific classification algorithm. Finally, we use a specific ensemble rule to combine the classification results of these classifiers for the new data. Five ensemble rules Max, Min, Product, Majority Vote, and Sum from [31] and another five improved ensemble rules MaxDistance, MinDistance, ProDistance, MajDistance, and SumDistance proposed by us are studied.

In the empirical study, the performance of six different types of classification algorithms Naive Bayes [32], C4.5 [33], RIPPER [34], Random Forest [35], SMO [36], and IBK [37] were evaluated over 46 highly imbalanced data sets. We first studied which ensemble rule performs better for the two data balancing methods ClusterBal and SplitBal. The experimental results show that for both of the two data balancing methods, ensemble rule MaxDistance performs better than other nine ensemble rules. Then we studied which combination of data balancing method and ensemble rule performs better with these six classification algorithms, and found that the two best combinations ClusterBal+MaxDistance and SplitBal+MaxDistance perform differently for different classification algorithms. After that, we compared our method with the conventional external imbalance data handling methods, including random under-sampling [11], random over-sampling [11], SMOTE [22], MetaCost [38], Bagging [27], Boosting [28], EasyEnsemble [19], SMOTEBoost [15], RUSBoost [16], UnderBagging [39] and our previous method EM1vs1 [40] which deals with the class imbalance problems in software defect prediction. Finally we compared our method with two representative internal imbalance data handling methods Chi3-GTS and Chi5-GTS [41]. The experimental results show that our method is usually more effective than these conventional external and internal imbalance data handling methods over the 46 highly imbalanced data sets.

The remainder part of this paper is organized as follows: Section 2 introduces the related work. In Section 3, we present our proposed method. Section 4 is devoted to the experiments, discusses the detailed experimental setup and analyzes the corresponding results. Finally, in Section 5 we make our concluding remarks.

2. Related work

The imbalance problem is one of the top 10 challenging problems in data mining [42]. It occurs in many real-world domains [40,43,44], and has drawn a significant amount of attention in the fields of data mining and machine learning. Apart from the class imbalance characteristic, there might be other data characteristics that could affect the performance of traditional classification algorithms for imbalanced problems [45–47], such as dataset shift [48], class overlapping [49], and small disjuncts [50]. However, in this study we focus on handling the class imbalance characteristic as it is the primary and intuitive data characteristic in imbalanced data, and other unintuitive data characteristics that may affect the performance of imbalanced data learning will be our future study.

Up to now, a variety of methods has been proposed for dealing with class imbalance problems. These methods can be broadly divided into two categories, namely external methods and internal methods [51]. Internal methods modify existing classification algorithms for reducing their sensitiveness to class imbalance [41,52–54], while external methods preprocess the training data to make them balanced. As the external methods have the advantage of independence on the underlying classification algorithms and our method could be regarded as a kind of external method, in this section we will pay more attention to the external methods, including the sampling methods [11,13,21,55–59], Bagging and Boosting based ensemble methods [16,19,60–64], and cost-sensitive learning methods [18,24–26,65,66].

Sampling methods, which are employed to balance class distribution in imbalanced data sets, can be divided into two groups: under-sampling and over-sampling. The under-sampling methods eliminate the majority class instances while the over-sampling methods increase the minority class instances for the purpose of obtaining a desirable rate of class distribution. Japkowicz [55] mainly discussed two strategies: under-sampling and resampling. She noted that both the sampling approaches were effective, and furthermore she also observed that using the sophisticated sampling techniques does not provide any clear advantage in solving the class imbalance problem. Moreover, Mani [13] also indicated that the random under-sampling strategy usually outperformed some other complicated under-sampling strategies. Kubat et al. [56] proposed a heuristic under-sampling strategy named One-Sided Selection for eliminating the majority class instances that are either borderline or noisy. In addition to the under-sampling strategies, the over-sampling strategies are also widely used for dealing with the class imbalance problem. Chawla et al. [22] proposed an over-sampling approach in which the minority class is over-sampled by creating “synthetic” instances rather than by over-sampling with replacement. Han et al. [67] proposed the borderline-SMOTE to over-sample the minority class instances near the borderline. Batista et al. [11] and Xie at al. [68] both showed that over-sampling usually perform better than under-sampling. Estabrooks et al. [58] and Barandela et al. [69] both suggested that a combination of over-sampling and under-sampling might be more effective to solve the class imbalance problem. However, we argue that the sampling methods alter the original data class distribution of imbalanced data and then lead to some unexpected mistakes. For example, over-sampling might lead to overfitting and under-sampling may drop some potentially useful information.

Apart from the sampling strategies, Bagging and Boosting based ensemble methods have also been widely applied to the class imbalance problem. Seiffert et al. [63] conducted a comprehensive study comparing sampling methods with boosting for improving the performance of decision trees model built for identifying the software defective modules. Their results showed that sampling methods were effective in improving the performance of such models while boosting outperformed even the best data sampling methods. Chawla et al. [15] proposed a novel approach SMOTEBoost for learning from imbalanced datasets on the basis of the SMOTE algorithm and the boosting procedure. Seiffert et al. [16] presented a different hybrid ensemble methods named RUSBoost, which combined the random under-sampling strategy with the boosting procedure. Their empirical results showed that RUSBoost performed comparably to SMOTEBoost while being a faster and simpler technique. Liu et al. [19] proposed the