Neurocomputing 158 (2015) 48–61

Contents lists available at ScienceDirect

Neurocomputing

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

Coupling different methods for overcoming the class imbalance problem

Loris Nanni [a,*], Carlo Fantozzi [a], Nicola Lazzarini [b]

[a] DEI, University of Padua, Via Gradenigo, 6 - 35131- Padova, Italy
[b] School of Computing Science Newcastle University Newcastle, NE1 7RU, UK

ARTICLE INFO

Article history:
Received 25 September 2013
Received in revised form
23 January 2015
Accepted 27 January 2015
Available online 13 February 2015

Keywords:
Imbalanced dataset
Ensemble of classifiers
Support vector machine
Downsampling/oversampling

ABSTRACT

Many classification problems must deal with imbalanced datasets where one class – the majority class – outnumbers the other classes. Standard classification methods do not provide accurate predictions in this setting since classification is generally biased towards the majority class. The minority classes are oftentimes the ones of interest (e.g., when they are associated with pathological conditions in patients), so methods for handling imbalanced datasets are critical.

Using several different datasets, this paper evaluates the performance of state-of-the-art classification methods for handling the imbalance problem in both binary and multi-class datasets. Different strategies are considered, including the one-class and dimension reduction approaches, as well as their fusions. Moreover, some ensembles of classifiers are tested, in addition to stand-alone classifiers, to assess the effectiveness of ensembles in the presence of imbalance. Finally, a novel ensemble of ensembles is designed specifically to tackle the problem of class imbalance: the proposed ensemble does not need to be tuned separately for each dataset and outperforms all the other tested approaches.

To validate our classifiers we resort to the KEEL-dataset repository, whose data partitions (training/test) are publicly available and have already been used in the open literature: as a consequence, it is possible to report a fair comparison among different approaches in the literature.

Our best approach (MATLAB code and datasets not easily accessible elsewhere) will be available at https://www.dei.unipd.it/node/2357.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Highly imbalanced datasets are not uncommon in many pattern recognition tasks [3,4]. For example, in medical datasets instances of diseased patients are typically rarer than instances of sane individuals. Yet, it is the rare cases that attract the most interest, as detecting them enables patients to be diagnosed and treated. Similar needs also appear in other real-world applications such as anomaly detection, fault diagnosis, email foldering, face recognition, fraud detection.

In binary classification, the under-represented class is called the *minority class* or *positive class*. The other class, which contains the vast majority of the members, is referred to as the *majority class* or *negative class*. When the class distribution is asymmetric, regular classifiers – such as support vector machines (SVMs) – tend to ignore data in the minority class and treat them as noise, resulting in a class boundary that unduly benefits the majority class. In turn, this produces a drop in precision when classifying the minority class [5].

A common approach in m-class learning, with m greater than 2, is the one-against-all method: the original problem is decomposed into m binary classification instances, where each class is in turn labeled as positive and the remaining m–1 as negative. Unfortunately, this method aggravates the issue of imbalance in each of the m instances. As a consequence, ad-hoc systems for handling multi-class imbalanced problems must be developed [4]. In the rich literature on imbalanced classification, the most common methods employed [6–13] are **undersampling** of the majority class, **oversampling** of the minority classes, **ensemble methods**, **cost-sensitive learning**, **asymmetric classification**, **dimension reduction**.

The simplest approach to **undersampling** is to randomly select a fraction of records from the majority class. Unfortunately, this may lead to a loss of useful information. An interesting undersampling procedure is proposed in two methods [13] called *EasyEnsemble* and *BalanceCascade*. In *EasyEnsemble*, the majority class is sampled into several independent subsets that are used to train separate classifiers, whose outputs are finally combined to produce the classification decision. In *BalanceCascade*, trained models are used to guide the sampling process for succeeding classifiers. The system is more focused on training patterns that are hard to classify. The main drawback of these preprocessing

* Corresponding author.
E-mail addresses: nanni@dei.unipd.it (L. Nanni), ncl.lazzarini@gmail.com (N. Lazzarini).

http://dx.doi.org/10.1016/j.neucom.2015.01.068
0925-2312/© 2015 Elsevier B.V. All rights reserved.

algorithms is that, again, potentially useful data from the majority class may not be considered.

As far as **oversampling** is concerned, the basic approach is to randomly duplicate the records in the minority classes to increase the cardinality of the classes themselves. A very popular approach is SMOTE (Synthetic Minority Oversampling TEchnique), which increases diversity by generating pseudo minority class data [14]. Several variants of SMOTE have been proposed: among them, we cite Borderline-SMOTE [57], MSMOTE [15], and the recent MWMOTE [53]. In [52] the imbalance problem is tackled by generating artificial instances, using an evolutionary framework, in order to modify the class ratio in the original dataset. A further interesting approach is RAMOBoost, introduced in [31] for binary classification. This technique oversamples the minority class using an adaptive weight adjustment procedure that shifts the decision boundary towards the difficult-to-learn examples from both the minority and majority classes.

Boosting (see also [19,20,34,51]) and other **ensemble methods**, such as Bagging [16,17,18,21], have proved to be particularly robust at handling imbalanced data. A recent review on ensemble methods applied to handle the class imbalance problem is [48]. AdaBoost [58], for instance, is designed to reduce the bias towards the majority class by focusing on misclassified training patterns [19], while Bagging introduces the concept of *bootstrap aggregating* [18] that consists in training several classifiers with bootstrapped copies of the original training set. Ensemble classifiers by themselves do not ameliorate the issue of imbalance if they are directly applied on data: this is due to their accuracy-oriented design. However, their combination with other techniques leads to positive results. Some examples are: SMOTEBoost [20], SMOTEBagging [21], IIVotes [22,44], RUSBoost [34]. In these approaches, a data-preprocessing algorithm is applied before bagging/boosting, hence a 3-step process can be identified: resampling, ensemble building, and voting for the final classification. IRUS [23] is a method that couples random undersampling and Bagging. The main idea behind it is to create bags where the majority class is so severely under-sampled that the imbalance situation is reversed with respect to the original one. Each bag contains all the positive patterns but only a few negatives: in this way, the focus of classification is on the minority class, which can be successfully separated from the majority class.

Another group of classifiers is based on **cost-sensitive learning**. In this approach, a different cost is assigned to false negative and false positive patterns. SVM-WEIGHT implements cost-sensitive learning for SVM modeling [32]. It is implemented in LIBSVM[1], so it is a very interesting baseline. The cost-sensitive principle is also applied in [37] to the ELM classifier.

A number of recent studies have focused on the development of **asymmetric classifiers** [23]. The main difference with cost-sensitive classification is that asymmetric classifiers are not exclusively focused on assigning a different weight to false negative and false positive patterns. Chew et al. [24] propose an unbalanced SVM (called UnSVMs in their paper) to adjust the error penalties of each class. Granular SVM is proposed in [25]: it resorts to a repetitive under-sampling method where information loss is minimized and the undersampling process maximizes the positive effect of data cleaning.

Some researchers have focused on improving **dimension reduction** methods, such as principal component analysis (PCA) and linear discriminant analysis (LDA), as a way of handling imbalanced data sets [26,27,28]. The key idea behind many of these methods is the eigen decomposition problem, which is tightly bound with data structure and class distribution, where the latter is asymmetric when data are skewed. To offset the effects of imbalanced data when applying PCA, an asymmetric principal component and discriminant analysis (APCDA) method was successfully employed in [26]. It is also worth to mention the method proposed in [30], where the authors implement an asymmetric classifier based on partial least squares (PLS) [29] to generate a new classification hyperplane and tackle the imbalance.

In recent years, some methods based on plain SVM have been proposed as well. For instance, in [32] a method called VQSVM is introduced. It is well known that SVM selects a subset of training patterns and uses them as the set of support vectors within the decision function. In VQSVM, vector quantization replaces the original set of support vectors with a subset, so that the number of instances belonging to the majority class is reduced.

In this paper, our aim is twofold.

1. We study the fusion among different approaches for handling the imbalance in the datasets following the "ensemble of ensembles" strategy.
2. We propose a new ensemble method, called HardEnsemble, to overcome the problems of existing approaches (e.g. parameters tuning, removal of potentially informative patterns, generation of new outliers, etc.). HardEnsemble provides good performances with both 2-class and multi-class datasets. Note that previous works were commonly focused only one of the two types of datasets, while we cover both.

To validate our results we test many state-of-the-art approaches using more than 40 datasets. In order to properly evaluate the results we employ a statistical test, the Wilcoxon signed-rank test, as commonplace in the open literature when it is necessary compare algorithms over multiple datasets [2].

2. Tested approaches

The aim of this paper is to find a set of approaches that works well with several datasets: to be precise, we want to determine whether some fusion of different classifiers for handling the imbalance problem consistently outperforms each of the classifiers when taken in isolation. We have tested methods based on different approaches, e.g. cost-sensitive learning, oversampling of the minority classes, and undersampling of the majority class. In all cases, an SVM is used as the base classifier unless differently specified.

In this section we give a technical summary of the approaches that have been previously presented in the open literature and included in our investigation; a separate subsection is devoted to each approach.

2.1. Asymmetric PLS classifier (APLSC)

APLSC is an asymmetric partial least squares (PLS) classifier [30] which complexly researches into the skewed distribution between classes and is prone to give high accuracy to the minority class in the cost of poor performances on the majority class. It can be summarized into two steps:

- feature extraction performed using PLS method [29] on normalized feature vectors;
- classification of compressed vectors by translated hyperplane, that is influenced by the variance of low dimensional data.

2.2. One-class SVM (OCSVM)

One-class SVM is an adaptation, proposed by Scholkpof [33], of SVM to one-class classification problem. SVM is usually

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Download English Version:

https://daneshyari.com/en/article/411946

Download Persian Version:

https://daneshyari.com/article/411946

Daneshyari.com