Pattern Recognition 70 (2017) 12–24

Contents lists available at ScienceDirect

Pattern Recognition

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

Class Switching according to Nearest Enemy Distance for learning from highly imbalanced data-sets

Sergio Gónzalez[a,*], Salvador García[a], Marcelino Lázaro[b], Aníbal R. Figueiras-Vidal[b], Francisco Herrera[a]

[a] Department of Computer Science and Artificial Intelligence, University of Granada, 18071 Granada, Spain
[b] Department of Signal Theory and Communications, Carlos III University of Madrid, 28903 Getafe Madrid, Spain

ARTICLE INFO

Article history:
Received 20 October 2016
Revised 4 April 2017
Accepted 28 April 2017
Available online 29 April 2017

Keywords:
Imbalanced classification
Ensembles
Preprocessing
Class Switching

ABSTRACT

The imbalanced data classification has been deeply studied by the machine learning practitioners over the years and it is one of the most challenging problems in the field. In many real-life situations, the under representation of a class in contrary to the rest commonly produces the tendency to ignore the minority class, this being normally the target of the problem. Consequently, many different techniques have been proposed. Among those, the ensemble approaches have resulted to be very reliable. New ways of generating ensembles have also been studied for standard classification. In particular, Class Switching, as a mechanism to produce training perturbed sets, has been proved to perform well in slightly imbalanced scenarios. In this paper, we analyze its potential to deal with highly imbalanced problems, fighting against its major limitations. We introduce a novel ensemble approach based on Switching with a new technique to select the switched examples based on Nearest Enemy Distance. We compare the resulting SwitchingNED with five distinctive ensemble-based approaches, with different combinations of sampling techniques. With a better performance, SwitchingNED is settled as one of best approaches on the field.

1. Introduction

The classification of datasets with skewed class distributions has drawn the attention of practitioners along the years of machine learning research. The class imbalance problem refers to datasets with a wide disparity in the number of instances for each class. In a binary imbalanced scenario, the minority class has much fewer examples than those of the majority class. These differences in the amount of samples cause a great loss of the minority class accuracy when applying standard classifiers. Real-life classification problems frequently suffer from this difficulty and misclassifying an example of the minority class usually entails greater costs. Therefore, the treatment of this problem is extremely important.

However, the skewed class distribution is not the only problem that has to be handled in order to obtain good behaviors [1,2]. Those other problems are the overlapping between classes [3], the high impact of noise [4,5], the identification of small disjuncts [6], the lack of density in the training data [7,8] and the possible differences in the data distribution between the training and test sets (dataset shift) [9,10]. Even though these are common problems in standard classification, they have a greater impact over the minority class in imbalanced environments.

Consequently, a large number of techniques have been proposed over these years to deal with this problem. Three groups can be defined: Algorithmic based approaches [11], cost-sensitive learning [12,13], and sampling data based approaches [2]. Ensembles have been successfully combined with these earlier methods recalling the best performances towards the problem [14]. EUSBoost was proposed in [15], and it has been found as the best ensemble to deal with highly imbalanced scenarios.

A type of ensemble known as Output Flipping was proposed in [16], based on randomizing the output of the training set, maintaining the input data untouched, differently to Bootstrapping. Output Flipping exchanges the class labels between instances preserving the relative frequencies of the classes. This last characteristic limits its applicability to imbalanced datasets. Later on, this same concept was extended in [17] and named as Class Switching. This algorithm randomly selects the instances to be changed without maintaining the same number of representatives from each class. These approaches have been proved to perform similarly to Bagging. However, the particularities of the Switching algorithm tend to equal the number of representatives from all classes.

[*] Corresponding author.
E-mail addresses: sergiogvz@decsai.ugr.es (S. Gónzalez), salvagl@decsai.ugr.es (S. García), mlazaro@tsc.uc3m.es (M. Lázaro), arfv@tsc.uc3m.es (A.R. Figueiras-Vidal), herrera@decsai.ugr.es (F. Herrera).

http://dx.doi.org/10.1016/j.patcog.2017.04.028
0031-3203/© 2017 Elsevier Ltd. All rights reserved.

Consequently, Switching performs better than other ensembles in slightly imbalanced scenarios.

Due to its effectiveness in slightly imbalanced scenarios, we explore the suitability of Switching for classification in highly imbalanced problems. We set the challenge of using the goodness of Switching approaches to successfully deal with highly imbalanced datasets, expecting to get performance advantages. In particular, we propose a novel Switching-based ensemble with a new technique to select the switched examples based on the Nearest Enemy Distance (NED). We call this approach SwitchingNED. Its new way of selecting the switched instances changes drastically the initial idea of the base approaches and solves their drawbacks for this kind of problems. Class Switching is applied to instances of the majority class according to their proximity to the minority class, measured by the NED. This allows a growth of the minority class population near to the classification boundaries, finding a better balanced representation between classes in highly imbalanced scenarios. As done with other ensembles [14], we combine SwitchingNED with different sampling methods, obtaining very promising results.

In order to test the proposed method under the given hypotheses, we perform several experiments which support that SwitchingNED successfully deals with highly imbalanced datasets and is one of the best approaches on the field. We have checked the improvement of our scheme against the basic Switching algorithm and its behavior in combination of preprocessing techniques. SwitchingNED is compared with five different representative approaches of each combination of sampling techniques and ensemble schemes [14]: EUSBoost [15], UnderBagging [18], SMOTEBoost [19], SMOTEBagging [20] and EasyEnsemble [21]. Among these methods, the best method of the literature is included, which is EUSBoost [15]. This experimental framework includes a total of 33 highly imbalanced datasets, where the majority class is at least 9 times bigger than the minority one. Furthermore, the empirical study has been validated using non-parametric statistical testing [22].

This paper is organized as follows. In Section 2 we present the class imbalance problem, within its drawbacks, its measures and the possible solutions to it. Section 3 introduces the randomizing output approaches. Section 4 is devoted to introducing SwitchingNED with its novel technique of selecting the instances to be switched based on NED. Section 5 describes the experimental framework followed by the empirical results and compares them to those offered by the original Switching approach. In Section 6 the use of sampling techniques in SwitchingNED is presented and analyzed with different experiments. Section 7 closes the paper presenting its main conclusions.

2. The problem of imbalanced class distributions in classification

In this section, we first present the class imbalance problem in detail, specifying its characteristics, the difficulties that it creates for standard classifiers, and the particular metrics used in this problem. Afterwards, we briefly survey the different approaches to deal with this problem.

2.1. The imbalanced class problem

This problem appears when one of the classes, known as minority or positive class, has fewer representing examples than the other, majority or negative class. It is common in many real applications, such as biometric identification [23,24] and bioinformatics [25,26], that have brought a growth of attention by researchers. Furthermore, the minority class normally merits more interest from a learning point of view, implying a greater cost when it is not well classified [27].

The main concern with imbalanced datasets is that the standard classification learning methods are usually biased toward the majority class. Standard classifiers use commonly global performance measures, like the accuracy rate, to guide the learning process, which benefits the majority class. Even more, they are normally designed to discard very highly specialized classification rules, those needed to predict the positive class, in favor of more general ones. Consequently, there is a higher misclassification rate for the minority class.

This loss of performance is mainly caused by the differences in the amounts of samples between classes, increasing with the Imbalance Ratio (IR) [1] of the datasets. This ratio is computed as the number of the negative class examples divided by the number of positive class examples. The datasets with IR greater than 9 are considered as highly imbalanced [15]. There are several studies [1,14] pointing out that there are other problems to consider in order to obtain good behaviors. The presence of small disjuncts, overlapping, lack of density or noisy have been proved to greatly reduce the accuracy of the minority class prediction [1].

The evaluation of the algorithms applied to this problem is of great relevance in order to properly guide their learning phase. Standard measures like accuracy cannot be used, because they assess the methods with the overall performance on the entire datasets and not the performance on each class independently. There are other more suitable measures for this environment, such as the Area Under the ROC Curve or Geometric Mean. The Area Under the ROC Curve (AUC) [28] is a widely used measure in imbalanced domains [1,2,14]. The Receiver Operating Characteristic (ROC) curve [29] represents graphically the trade-off between the percentage of well classified positive instances (TP_{rate}) and the percentage of incorrectly classified negative instances (FP_{rate}). In our experiments, we use the simplified expression of the AUC computed with the TP_{rate} and FP_{rate} [15].

2.2. Solutions to the class imbalance problem

As previously mentioned, the approaches for dealing with imbalanced dataset classifications could be divided in the three groups: Cost-sensitive learning solutions [13,30], algorithmic approaches [11,31] or data level methods [2,19,32].

The cost-sensitive learning considers the costs of the errors by minimizing a cost function that includes them. In imbalanced classification, these methods take into account the higher cost of misclassification of the minority class with respect to its alternative. These costs are defined by domain experts or learned with other approaches. The algorithmic approaches are modifications of base learning algorithms that achieve good performance with imbalanced datasets.

Finally, the data level methods focus on pre-processing the original dataset in order to obtain a more balanced one. These techniques allow to use standard learning algorithms after the preprocessing. The equal representation between classes is reached by generating more examples for the positive class (oversampling), removing examples from the negative class (undersampling) or both (hybrid methods). The most elementary techniques in this field are Random Under/Over- sampling, which randomly remove or replicate instances from the majority/minority class. Its first modality has the disadvantage of being able to eliminate real valuable data, while the oversampling can come out with over-fitting. Several sophisticated approaches have been proposed, such as SMOTE (Synthetic Minority Oversampling TEchnique) [33] and others based on complex heuristics like genetic algorithms [34].

Another category can be defined when ensemble classifiers are considered. Recently, these types of approaches have become more