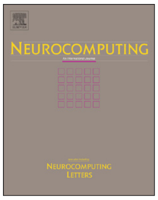
Contents lists available at ScienceDirect

Neurocomputing

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



Neighbourhood sampling in bagging for imbalanced data

Jerzy Błaszczyński *, Jerzy Stefanowski **

Institute of Computing Sciences, Poznań University of Technology, 60-965 Poznań, Poland

ARTICLE INFO

Article history:
Received 21 December 2013
Received in revised form
27 April 2014
Accepted 26 July 2014
Available online 2 October 2014

Keywords:
Class imbalance
Ensemble classifiers
Bagging

ABSTRACT

Various approaches to extend bagging ensembles for class imbalanced data are considered. First, we review known extensions and compare them in a comprehensive experimental study. The results show that integrating bagging with under-sampling is more powerful than over-sampling. They also allow to distinguish Roughly Balanced Bagging as the most accurate extension. Then, we point out that complex and difficult distribution of the minority class can be handled by analyzing the content of a neighbourhood of examples. In our study we show that taking into account such local characteristics of the minority class distribution can be useful both for analyzing performance of ensembles with respect to data difficulty factors and for proposing new generalizations of bagging. We demonstrate it by proposing Neighbourhood Balanced Bagging, where sampling probabilities of examples are modified according to the class distribution in their neighbourhood. Two of its versions are considered: the first one keeping a larger size of bootstrap samples by hybrid over-sampling and the other reducing this size with stronger under-sampling. Experiments prove that the first version is significantly better than existing over-sampling bagging extensions while the other version is competitive to Roughly Balanced Bagging. Finally, we demonstrate that detecting types of minority examples depending on their neighbourhood may help explain why some ensembles work better for imbalanced data than others.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

An analysis of challenging real-world classification problems still reveals difficulties in finding accurate classifiers. One of the sources of these difficulties is class imbalance in data, where at least one of the target classes contains a much smaller number of examples than the other classes. For instance, in medical problems the number of patients requiring special attention (e.g., therapy or treatment) is usually much smaller than the number of patients who do not need it. Similar situations occur in other problems, such as fraud detection, risk management, technical diagnostics, image recognition, text categorization or information filtering. In all those problems, the correct recognition of the minority class is of key importance. Nevertheless, class imbalance constitutes a great difficulty for most learning algorithms. Often the resulting classifiers are biased toward the majority classes and fail to recognize examples from the minority class. As it turns out, even ensemble methods, where multiple classifiers are trained to deal with complex classification tasks are not particularly well suited to this problem.

Although the difficulty with learning classifiers from imbalanced data has been known earlier from applications, this challenging problem has received a growing research interest in the last decade and a number of specialized methods have already been proposed, for their review see, e.g., [11,17,18,40]. In general, they may be categorized into *data level* and *algorithm level* ones. Methods within the first category try to re-balance the class distribution inside the training data by either adding examples to the minority class (*over-sampling*) or removing examples from the majority class (*under-sampling*). They also include informed pre-processing methods as, e.g., SMOTE [10] or SPIDER [37].

The other category of algorithm level methods involves specific solutions dedicated to improving a given classifier. They usually include modifications of the learning algorithm, its classification strategy or adaptation to the cost sensitive framework. Within the algorithm level approaches, *ensembles* are also quite often applied. However, as the standard techniques for constructing ensembles are rather too overall accuracy oriented they do not sufficiently recognize the minority class and new extensions of standard techniques have been introduced. These new proposed solutions usually either employ pre-processing methods before learning component classifiers or embed the cost-sensitive framework in the ensemble learning process; see their review in [13,29]. Most of these ensembles are based on known strategies from bagging, boosting or random forests.

* Principal corresponding author.
** Corresponding author.
E-mail addresses: jerzy.blaszczynski@cs.put.poznan.pl (J. Błaszczyński), jerzy.stefanowski@cs.put.poznan.pl (J. Stefanowski).

http://dx.doi.org/10.1016/j.neucom.2014.07.064
0925-2312/© 2014 Elsevier B.V. All rights reserved.

Although the ensemble classifiers are recognized as a remedy to imbalanced problems, there is still a lack of a wider study of their properties. Authors often compare their proposals against the basic versions of other methods or compare over a too limited collection of data sets. Up to now, only two quite comprehensive studies were carried out in different experimental frameworks [13,24]. The first study [13] covers comparison of 20 different ensembles from simple modifications of bagging or boosting to complex cost or hybrid approaches. The main conclusion from this study is that simple versions of under-sampling or SMOTE re-sampling combined with bagging works better than more complex solutions. In the second study [24], two best boosting and bagging ensembles are compared over noisy and imbalanced data. The experimental results show that bagging significantly outperforms boosting. The difference is more significant when data are more noisy. The similar observations on good performance of under-sampling generalizations of bagging vs. cost like generalization of boosting have been recently reported in [2]. Furthermore, the most recent chapter of [29] includes a limited experimental study showing that new ensembles specialized for class imbalance should work better than an approach consisting of first pre-processing data and then using standard ensembles.

Following these related works which show good performance of bagging extensions for class imbalance vs. other boosting like or cost sensitive proposals, we have decided to focus our interest in this paper on studying more deeply bagging ensembles and to look for possible other directions of their generalizations. First, we want to study behaviour of bagging extensions more thoroughly than it was done in [13,24]. In particular, Roughly Balanced Bagging [19] was missed in [13], although it is appreciated in the literature. On the other hand, the study presented in [24] was too much oriented on the noise level and only two versions of random under-sampling in bagging were considered. Therefore, we will consider a larger family of known extensions of bagging. Our comparison will include Exactly Balanced Bagging, Roughly Balanced Bagging, and more variants of using over-sampling in bagging, in particular, a new type of integrating SMOTE.

While analyzing existing extensions of bagging one can also notice that most of them employ the simplest random re-sampling technique and, what is even more important, they modify bootstraps to simply balance the cardinalities of minority and majority classes. So, they represent a kind of a *global* point of view on handling the *imbalance ratio* between classes.

Recent studies on class imbalances have shown that this global ratio between imbalanced classes is not a problem itself. For some data sets with high imbalance ratio, the minority class can still be sufficiently recognized even by standard classifiers. The degradation of classification performance is often linked to other *difficulty factors* related to data distribution, such as decomposition of the minority class into many rare sub-concepts [23], the effect of too strong overlapping between the classes [36,16] or the presence of too many minority examples inside the majority class regions [32]. When these factors occur together with class imbalance, they seriously hinder the recognition of the minority class. In earlier research of Napierala and Stefanowski on single classifiers [33] it has been shown that these data difficulty factors could be at least partly approximated by analyzing the *local characteristics* of learning examples from the minority class. Depending on the distribution of examples from the majority class in the local neighbourhood of the given minority example, we can evaluate whether this example could be safe or unsafe (difficult) to be learned. This local view on distributions of imbalanced classes leads us to main aims of this paper.

The main aim of our paper is to study usefulness of incorporating the information about the results of analyzing the local neighbourhood of minority examples into two directions: proposing new generalizations of bagging for class imbalance and extending analysis of classifier performance over different imbalanced data sets.

Following the first direction our aim is to propose extensions of bagging specialized for imbalanced data, which are based on a different principle than the existing ones. Our new approach is to resign from simple integration of pre-processing with unchanged bootstrap sampling technique. Unlike standard bootstrap sampling, we want to change probability of drawing different types of examples. We would like to focus the sampling toward the minority class and even more to the examples located in the most difficult sub-regions of the minority class. The probability of each minority example to be drawn will depend on the class distribution in the neighbourhood of the example [33]. We plan to consider this modification of sampling in two versions of generalizing bagging: (1) over-sampling one, which replicates the minority examples and filters some majority examples to keep the size of a bootstrap sample larger, similar to the size of the original data set; (2) under-sampling one, which is following the idea of explored in Rough Balanced Bagging, and Exactly Balanced Bagging. The under-sampling modification constructs a smaller bootstrap with the size equal to the double the size of the minority class. We plan to evaluate usefulness of both versions in comparative experiments.

The next aim is to better explain differences in performance of various generalizations of bagging ensemble. Current, related studies on this subject are based on a global view on selected evaluations measures over many imbalanced data sets. We hypothesize that it could be beneficial to differentiate between groups of data sets with respect to their underlying data difficulty factors and to study differences in performance of classifiers within these groups. We will show that it could be done by analyzing contents of the neighbourhood of the examples as it leads to an identification of dominating types of difficulty for minority examples. Furthermore, we plan to study more thoroughly contents of bootstrap samples generated by the best performing extensions of bagging. This examination will also be based on analyzing neighbourhood of the minority examples. We will identify differences between bootstrap samples and the original data, and we will try to find a new view on learning of these generalized ensembles.

To sum up, the main contributions of our study are the following. The first one is to study more closely the best known extensions of bagging over a representative collection of imbalanced data sets. Then, we will present a method for analyzing contents of the neighbourhood of the examples and to discuss its consequences. The next methodological contribution is to introduce a new extension of bagging for imbalanced data based on this analysis of a neighbourhood of each example, which affects the probability of its selection into a bootstrap sample. The new proposal will be compared against the best identified extensions. Finally, we will use the same type of the local analysis to explain differences in performance of bagging classifier and to answer a question why contents of bootstrap samples in particular extension of bagging may lead to its good performance.

2. Related works on ensembles for imbalanced data

Several studies have already investigated the problem of class imbalance. The reader is referred to the recent book [18] for a comprehensive overview of several methods and the current state of the art in the literature. Below we very briefly summarize these methods only, which are most relevant to our paper.

First, we describe data *pre-processing methods* as they are often integrated with many ensembles. The simplest data pre-processing re-sampling techniques are *random over-sampling*, which replicates examples from the minority class, and *random under-sampling*, which randomly eliminates examples from the majority classes until a

Download English Version:

https://daneshyari.com/en/article/407544

Download Persian Version:

https://daneshyari.com/article/407544

Daneshyari.com