Information Sciences xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Information Sciences

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

Diversity techniques improve the performance of the best imbalance learning ensembles

Q1 José F. Díez-Pastor [a],*, Juan J. Rodríguez [a], César I. García-Osorio [a], Ludmila I. Kuncheva [b]

[a] *University of Burgos, Spain*
[b] *University of Bangor, UK*

ARTICLE INFO

Article history:
Received 6 October 2014
Revised 15 May 2015
Accepted 4 July 2015
Available online xxx

Keywords:
Classifier ensembles
Imbalanced data sets
SMOTE
Undersampling
Rotation forest
Diversity

ABSTRACT

Many real-life problems can be described as unbalanced, where the number of instances belonging to one of the classes is much larger than the numbers in other classes. Examples are spam detection, credit card fraud detection or medical diagnosis. Ensembles of classifiers have acquired popularity in this kind of problems for their ability to obtain better results than individual classifiers. The most commonly used techniques by those ensembles especially designed to deal with imbalanced problems are for example Re-weighting, Oversampling and Undersampling. Other techniques, originally intended to increase the ensemble diversity, have not been systematically studied for their effect on imbalanced problems. Among these are Random Oracles, Disturbing Neighbors, Random Feature Weights or Rotation Forest. This paper presents an overview and an experimental study of various ensemble-based methods for imbalanced problems, the methods have been tested in its original form and in conjunction with several diversity-increasing techniques, using 84 imbalanced data sets from two well known repositories. This paper shows that these diversity-increasing techniques significantly improve the performance of ensemble methods for imbalanced problems and provides some ideas about when it is more convenient to use these diversifying techniques.

1. Introduction

The class imbalance problem[1] arises when one class has much more examples than the others [11].

Imbalance learning has attracted much attention because imbalanced data sets are common in real world problems like those related to security: spam detection [29], fraud detection [17], software defect detection [65]; biomedical data: finding the transition between coding and non-coding DNA in genes [28], mining cancer gene expression [70]; or financial data, for example, risk predictions in credit data [25].

Classification of imbalanced data is difficult because standard classifiers are driven by accuracy, hence the minority class may simply be ignored [58], besides generally all classifiers present some performance loss when the data is unbalanced [50]. In addition, many imbalanced datasets suffer problems related to its intrinsic characteristic. According to [44] there are at least six

* Corresponding author. Tel.: +34653030301.
E-mail address: jfdpastor@ubu.es (J.F. Díez-Pastor).

[1] Being aware of the terminological debate about "*unbalanced*" vs "*imbalanced*", we will use both words interchangeably. Our reason is that a keyword look-up should be able to retrieve this study, whichever word has been picked.

http://dx.doi.org/10.1016/j.ins.2015.07.025
0020-0255/© 2015 Published by Elsevier Inc.

of these problems: overlapping [59], lack of density and information [66], noisy examples [8], small disjuncts [68], the significance of borderline instances to discriminate between positive and negative classes [47] and differences in the data distributions between training and test stages [54].

In [23], the approaches to dealing with unbalanced datasets are sorted into four categories[2]:

- **The algorithm-level** category encompasses modifications of existing general learning algorithms which bias the learning toward the minority class. Examples of this category are Hellinger Distance Decision Trees (HDDT) [14], Class Confidence Proportion Decision Tree (CCPDT) [42] and Significant, Positively Associated and Relatively Class Correlated Classification Trees (SPARCCC) [63], as well as other class-size insensitive decision trees. In other occasions misclassification costs are different for different examples, [71] presents decision tree and Naïve Bayesian learning methods that learns with unknown costs.
- **The data-level** category includes pre-processing algorithms that change the prior distribution of the classes either by increasing the number of minority class examples or by reducing the size of the majority class. In the first category of algorithms the simplest technique is to randomly add examples, without caring about neighbors from other class or the overlap between classes, some examples are Oversampling [4], SMOTE [10]. Other methods creates artificial instances taking into account these issues: Borderline-SMOTE [31], Safe-level SMOTE [9], ADASYN [32] or Cluster Based Oversampling [38]. In the second category Random Undersampling [3] removes random examples from the majority class and other methods like Edited Nearest Neighbor (ENN) [69] and Tomek Links [61] are based on data cleaning techniques.
- **The cost-sensitive** category contains methods that assign different costs for each class. Examples include AdaCost [20], AdaC1, AdaC2, and AdaC3 [60].
- **Classifier ensembles** [40,49] are combinations of several classifiers which are called base classifiers or member classifiers. Ensembles often give better results than individual classifiers. Although ensembles were not designed to work with imbalanced data, they have been successfully applied to this task through combination with processing techniques from the data-level category.

According to [23], the algorithm level and cost-sensitive approaches are more problem-dependent, whereas data level and ensemble learning approaches based on data processing are more versatile.

Ensemble methods for imbalanced learning tackle the imbalance problem using techniques like re-weighting, Oversampling and Undersampling. These preprocessing techniques attempt to train base classifiers with a less unbalanced dataset. These preprocessing techniques not only address the problem of imbalance, but add diversity, since each base classifier is trained on a different version of the data set. Diversity is one of the cornerstones of ensembles. An ideal ensemble system should have accurate individual classifiers and at the same time their errors should be in different instances. Several techniques have been developed to increase the diversity of an ensemble (see Section 2.4). In this paper we argue that techniques especially designed to increase diversity impact the performance of imbalance learning, significantly improving even the specific techniques.

To prove this claim, we conducted an experimental study where both classifiers especially designed for unbalanced sets and standard classifiers were tested in its original form and in conjunction with several diversity-increasing techniques. Finding that ensembles combined with diversity-increasing techniques ranked better than their original counterpart, even though the original version was specifically designed to work for imbalanced data. We also try to provide some clues when it is more appropriate to use diversity techniques using meta-learning, and evaluate the performance of the techniques in the presence of noisy and borderline examples.

The rest of the paper is structured as follows: Section 2 presents some background of ensemble learning, state-of-the-art techniques for imbalanced data, and our research hypothesis concerning diversity enhancing techniques. Section 3 shows the experimental study and results. Section 4 enumerates the findings extracted in the experimental study. And finally, in Sections 5 and 6 the conclusions and several future lines of research are presented.

2. Ensemble learning for imbalanced problems

Q2 In this section, the concept of ensemble and the importance of diversity will be introduced, then the preprocessing techniques and the ensembles methods for imbalanced problems used in this paper will be described. Finally, several techniques to increase the diversity in ensembles will be explained

2.1. Ensembles of classifiers

Ensemble of classifiers is combinations of multiple classifiers, referred as base classifiers. Ensembles usually achieve better performance than any of the single classifiers [40]. In order to build a good ensemble, it is necessary not only to build good base classifiers, also the base classifiers must be diverse, this means that for the same instance, the base classifiers return different outputs and their errors should be in different instances. Ensemble methods differ in the way they induce diversity between the base classifiers. The most common approach is modifying the training set for each member of the ensemble. In Bagging [6], each base classifier is obtained from a random sample of the training data. In the resampling version of AdaBoost [22], the data set for each subsequent ensemble member is drawn according to a distribution of weights over the data. The weights are modified

[2] Notice that these categories are not mutually exclusive, for example some cost-sensitive methods can be included in the classifier ensembles category.

Please cite this article as: J.F. Díez-Pastor et al., Diversity techniques improve the performance of the best imbalance learning ensembles, Information Sciences (2015), http://dx.doi.org/10.1016/j.ins.2015.07.025