Information Sciences 354 (2016) 178–196

Contents lists available at ScienceDirect

Information Sciences

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

Ordering-based pruning for improving the performance of ensembles of classifiers in the framework of imbalanced datasets

Mikel Galar[a], Alberto Fernández[b,*], Edurne Barrenechea[a,c], Humberto Bustince[a,c], Francisco Herrera[d,e]

[a] Departamento de Automática y Computación, Universidad Pública de Navarra, Pamplona, Spain
[b] Department of Computer Science, University of Jaén, Jaén, Spain
[c] Institute of Smart Cities (ISC), Universidad Pública de Navarra, Pamplona, Spain
[d] Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain
[e] Faculty of Computing and Information Technology - North Jeddah, King Abdulaziz University (KAU), Jeddah, Saudi Arabia

ARTICLE INFO

Article history:
Received 23 May 2015
Revised 17 February 2016
Accepted 25 February 2016
Available online 18 March 2016

Keywords:
Imbalanced datasets
Tree-based ensembles
Ordering-based pruning
Bagging
Boosting

ABSTRACT

The scenario of classification with imbalanced datasets has gained a notorious significance in the last years. This is due to the fact that a large number of problems where classes are highly skewed may be found, affecting the global performance of the system. A great number of approaches have been developed to address this problem. These techniques have been traditionally proposed under three different perspectives: data treatment, adaptation of algorithms, and cost-sensitive learning.

Ensemble-based models for classifiers are an extension over the former solutions. They consider a pool of classifiers, and they can in turn integrate any of these proposals. The quality and performance of this type of methodology over baseline solutions have been shown in several studies of the specialized literature.

The goal of this work is to improve the capabilities of tree-based ensemble-based solutions that were specifically designed for imbalanced classification, focusing on the best behaving bagging- and boosting-based ensembles in this scenario. In order to do so, this paper proposes several new metrics for ordering-based pruning, which are properly adapted to address the skewed-class distribution. From our experimental study we show two main results: on the one hand, the use of the new metrics allows pruning to become a very successful approach in this scenario; on the other hand, the behavior of Under-Bagging model excels, achieving the highest gain with the usage of pruning, since the random undersampled sets that best complement each other can be selected. Accordingly, this scheme is capable of outperforming previous ensemble models selected from the state-of-the-art.

[*] Corresponding author. Tel.: +34 948166048; fax: +34 948168924.
E-mail addresses: mikel.galar@unavarra.es (M. Galar), ahilario@ujaen.es, alberto.fernandez@ujaen.es (A. Fernández), edurne.barrenechea@unavarra.es (E. Barrenechea), bustince@unavarra.es (H. Bustince), herrera@decsai.ugr.es (F. Herrera).

http://dx.doi.org/10.1016/j.ins.2016.02.056
0020-0255/© 2016 Elsevier Inc. All rights reserved.

1. Introduction

When working with classification tasks, it may be observed that datasets frequently present a very different distribution of examples within their classes. This issue is known as the problem of imbalanced classes [30,66], and it has been addressed throughout the last ten years [13]. Even so, the development of algorithms for learning classifiers in this scenario is still a hot topic of research [46,60]. This is mainly due to the high number of real applications that are affected by this condition. Enumerating some examples we may refer to bankruptcy prediction [37], medical data analysis [9,38] and bioinformatics [7,29], among others.

The presence of classes with few data can generate sub-optimal classification models, since there is a bias towards the majority class. This is due to the fact that, when the standard accuracy metric is considered, predicting the class with a higher number of examples is preferred during the learning process; therefore, the discrimination functions computed by the algorithm will be positively weighted towards the majority class [46]. Hence, there is an undeniable need for developing more precise approaches in order to reach the maximum precision in every class, independently of its representation or distribution. Furthermore, recent studies have shown that additional data intrinsic characteristics have a strong influence on the correct identification of the minority class examples [46,64].

Traditionally, solutions for this problem have been divided into three large groups [46,48], i.e. preprocessing [4] (to balance the example distribution per class), ad-hoc adaptation of standard algorithms [78], and the usage of cost-sensitive learning [19]. Any of the former approaches can be integrated into an ensemble-type classifier, thus empowering the achieved performance, as it has been shown in the specialized literature [22,23,46,65].

In summary, an ensemble is a collection of classifiers aimed at increasing the generalization capability of a single classifier, since classifiers in the ensemble are supposed to complement each other [59,62,75]. These classifiers are then jointly applied in order to obtain a single solution in agreement. Reader might guess that the more elements the ensemble has, the more reliable the solution will be, but there is a limit from which the accuracy does not improve or even worse, it could be degraded [83]. There are two main reasons for this behavior: (1) the difficulty in the decision process regarding possible contradictions or even redundancy among the components of the ensemble; and (2) the overfitting problem when adjusting the weights in a boosting-based ensemble.

In accordance with these issues, several proposals have been developed to carry out a selection of classifiers within the ensemble [5,82], which are named as *pruning* methods. The goal is to obtain a subset of the ensemble that solves the classification problem in an optimal way, i.e., maintaining or improving the accuracy of the system. In this paper we focus on ordering-based pruning, whose working procedure is based on a greedy approach and whose effectiveness in standard classification has been already proved [31,51]. This scheme starts from a trained ensemble composed of a large number of classifiers. Then, classifiers are iteratively selected one by one from the pool according to the maximization of a given metric and added to the final ensemble. This process is usually carried out until a pre-established number of classifiers are selected.

The heuristic metrics used for the ensemble pruning methodology were originally defined for standard classification tasks. In the scenario of imbalanced datasets, the effect of each classifier in the recognition of both classes must be analyzed in detail in order to obtain valid results. Therefore, ordering-based pruning metrics must be adapted this specific scenario, taking the data representation into account. Our objective is to focus on the class imbalance of the problem during the whole learning process. First, in the ensemble learning stage, via the use of those learning methods inherently adapted to this context [23]. Second, a posteriori, that is, in the classifier pruning step by selecting the most appropriate classifiers with our novel proposed metrics. As we will show in the experimental study, this positive synergy will allow us to boost the final performance of the system.

Specifically, the contributions of this paper can be summarized as follows:

- To use the ensemble pruning methodology in the context of imbalanced classification for improving the behavior of ensemble-based solutions in this framework.
- To develop novel ordering-based pruning metrics taking the properties of the class imbalance problem into account. In particular, we focus on the adaptation of five of the most popular schemes for ordering-based pruning [51].
- To carry out a thorough experimental study in order to analyze the usefulness of this methodology in the imbalanced scenario. More specifically, we carry out an exhaustive comparison of all the adaptation of the five metrics so as to verify their results with the state-of-the-art ensembles on the topic, which were those previously stressed in [23].
- To study the true benefits of the application of these new metrics both with respect to the baseline methodologies and the state-of-the-art models. It will be shown that incorporating ensemble-pruning allows one to go a step further into the performance of ensemble-based solution.

For a fair evaluation of the ordering-based pruning in imbalanced classification, we have selected the best bagging- and boosting-based ensemble models that were highlighted in our previous study on the topic [23]. Finally, the validation of the novel imbalanced pruning methodology will be carried out using a wide benchmark of 66 different problems commonly used in this area of research [46], and supported by means of the statistical analysis of the results [24].

The rest of this paper is organized as follows. Section 2.1 introduces classification with ensembles for the problem of imbalanced data, as well as the ordering-based pruning approach with the metrics considered to perform this process. Then, Section 3 contains the core part of the manuscript, in which we present our adaptations to imbalanced classification for