Ordering-Based Pruning for Improving the Performance of Ensembles of Classifiers in the Framework of Imbalanced Datasets

A. Fernández[1], M. Galar[2], E. Barrenechea[1], H. Bustince[2], and F. Herrera[1]

[1] Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain; {alberto, herrera}@decsai.ugr.es
[2] Departamento de Automática y Computación, Universidad Pública de Navarra, Pamplona, Spain {mikel.galar,edurne.barrenechea,bustince}@unavarra.es

Abstract. This is a summary of our article published in Information Sciences [2], to be considered as part of the Multi- Conference CAEPIA'16 - FINO'16 KeyWorks

Keywords: Imbalanced Datasets, Tree-based Ensembles, Ordering-Based Pruning, Bagging, Boosting

1 Summary

The problem of classification with imbalanced datasets is a challenging task, due to the bias of standard learning models to the most represented classes [4]. Traditionally, solutions for this problem have been divided into three large groups, i.e. rebalancing techniques, ad-hoc adaptation of standard algorithms, and the usage of cost-sensitive learning.

The former approaches can be integrated into an ensemble-type classifier, thus empowering the achieved performance [1]. In spite of their robustness, the choice of the optimal number of classifiers for the ensemble learning is not straightforward. Selecting a low number may cause the ensemble not to reach a high and stable classification accuracy. On the contrary, a high number may imply redundant classifiers with less diversity, or even overfitting 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. 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. Among several approaches, the effectiveness of the ordering-based pruning scheme in standard classification makes it a valuable solution for this task [3]. This methodology 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.

Actas de la XVII Conferencia de la Asociación Española para la Inteligencia Artificial, pp. 937-938

Five heuristic metrics can be highlighted for this task: *Reduce Error* (RE), *Kappa*, *Complementary Measure* (Comp), *Margin distance minimization* (MDM), and *Boosting-based* (BB) pruning. They are aimed at improving the global performance and/or seeking the highest diversity in the outputs among classifiers. However, all were defined for standard classification. In imbalanced domains it can imply a bias in the selection process resulting in sub-optimal models.

Therefore, we proposed to adapt the former, taking the data representation into account, obtaining three novel metrics: "RE-GM" (with the use of the geometric mean), "MDM-Imb" (computing this metrics independently per class), and "BB-Imb" (carrying out a local weighting per class). In this sense, we focused 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 [1]. Second, a posteriori, that is, by selecting the most appropriate classifiers with our novel proposed metrics.

To check the validity of our approach, we developed an exhaustive experimental analysis over a large number of benchmark datasets. For this study we also selected the best bagging- and boosting-based ensemble models that were highlighted in our previous study on the topic [1].

From this complete analysis we extracted several lessons learned. First, the pruning mechanism is positively biased when using the new adapted heuristics metrics for imbalanced classification. Additionally, in all cases, the use of the imbalanced pruning metrics allows the enhancement of the baseline ensemble approaches. Specifically, three heuristic metrics were excelled over the rest: (1) "BB-Imb", (2) "Comp", and (3) "RE-GM". They have shown a clear synergy with every ensemble learning approach. Furthermore, in the case of the UnderBagging algorithm, the use of "BB-Imb" allowed to find statistical differences among the results versus the remaining approaches from the state-of-the-art with pruning. One reason for the good behavior in this case is due to a supervised selection from "randomness" via the "a posteriori" pruning, i.e. only considering those classifiers that present a better cooperation. This is difficult to be achieved in boosting-based ensembles, due to a higher dependency among classifiers.

References

1. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for class imbalance problem: Bagging, boosting and hybrid based approaches. IEEE Transactions on System, Man and Cybernetics Part C: Applications and Reviews 42(4), 463–484 (2012)
2. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: Ordering-based pruning for improving the performance of ensembles of classifiers in the framework of imbalanced datasets. Information Sciences 354, 178–196 (2016)
3. Hernandez-Lobato, D., Martinez-Muñoz, G., Suarez, A.: Statistical instance-based pruning in ensembles of independent classifiers. IEEE Transactions On Pattern Analysis And Machine Intelligence 31(2), 364–369 (2009)
4. Lopez, V., Fernandez, A., Garcia, S., Palade, V., Herrera, F.: An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences 250(20), 113–141 (2013)