Random Balance: Ensembles of Variable Priors Classifiers for Imbalanced Data

Jose F Diez-Pastor, Juan J. Rodriguez, César I Garcia-Osorio and Ludmila I Kuncheva

University of Burgos, Spain.
{jfdpastor,jjrodriguez,cgosorio}@ubu.es
Bangor University, United Kingdom.
{l.i.kuncheva}@bangor.ac.uk

Abstract. This is a summary of our article published in Knowledge Based Systems [1], presented to the Multi Conference CAEPIA'16 KeyWorks.

Keywords: Classifier ensembles, imbalanced data sets, Bagging, AdaBoost, SMOTE, Undersampling

1 Motivations and objectives

There are several preprocessing strategies for dealing with imbalance datasets: one is to decrease the size of the majority class and another to increase the size of the minority class. The problem is that the performance is very dependent on the parameters of these techniques and it is difficult to find out the optimal values for a particular dataset. This motivates us to study whether it is possible or not to rely on randomness and repetition to address this problem. Eliminating the need to choose the proper technique and adjust their optimal parameters.

2 Discussion of results

In this work it is proposed a new approach for building ensembles for two-class imbalance datasets. It is based on a simple randomisation heuristic: The dataset obtained after applying the preprocessing technique will have random class proportions (The classes are either reduced, using Random Undersampling or augmented with artificial examples using SMOTE) In the paper, it is also described a new ensemble method for imbalance learning which combines Random Balance with AdaBoost.M2 and is called RB-Boost (Random Balance Boost). The highlights of the paper are:

– The paper explains the intuition behind the method in a graphical way. Representing base classifiers as a point in the True Positive Rate - False Positive Rate space.

Actas de la XVII Conferencia de la Asociación Española para la Inteligencia Artificial, pp. 823-824

– The paper includes an exhaustive experimental study comparing the proposed technique with state of the art methods like SMOTEBoost, RUSBoost, SMOTEBagging and so on with 86 datasets from known repositories.
  • The ensembles using the proposed method obtains the top positions in the average ranks according to AUC, F-Measure and Geometric Mean.

3 Conclusions

The proposed technique is based on the idea of creating preprocessed output datasets in which the ratio between classes varies randomly. With this idea and taking inspiration from other methods such as SMOTEBoost and RUSBoost, this technique has been combined with boosting creating a new ensemble method: RB-Boost. This intuitive heuristic avoids the need to tune the proportion parameter and at the same time outperforms other state-of-the-art ensembles for imbalance.

4 Future lines

Imbalance dataset presents some problem that have a strong influence on imbalance classification. Some of these problems like overlapping, noisy examples or small disjuncts are addressed using preprocessing techniques. It can be very difficult to determine when a set of data suffers from these problems, unless the dataset is artificially generated. And much more difficult to adjust the value of the parameters of each technique for each dataset. One interesting future line is to use the ideas of Random Balance to combine preprocessing techniques that have been designed to address the aforementioned problems.

Acknowledgments

This work was partially supported by the project TIN2011-24046 of the Spanish Ministry of Economy and Competitiveness.

References

1. José F Díez-Pastor, Juan J Rodríguez, César García-Osorio, and Ludmila I Kuncheva. Random balance: ensembles of variable priors classifiers for imbalanced data. *Knowledge-Based Systems*, 85:96–111, 2015.