Instance Selection for Imbalanced Data

Sarah Vluymans[1], **Nele Verbiest**[1], **Chris Cornelis**[1,3], **Yvan Saeys**[2]
{Sarah.Vluymans, Nele.Verbiest, Chris.Cornelis, Yvan.Saeys}@UGent.be
Chris.Cornelis@decsai.ugr.es
[1] Department of Applied Mathematics, Computer Science and Statistics, Ghent University
[2] Department of Plant Systems Biology, VIB, Ghent University
[3] Department of Computer Science and Artificial Intelligence, University of Granada

Keywords: imbalanced data, instance selection, classification

Imbalanced data exhibit an unequal distribution with respect to the class labels. In a two-class imbalanced problem, elements of the *majority class* can vastly outnumber those belonging to the *minority class*. Several standard learning methods are hindered by such skewness present in the training set and fail to recognize minority instances in a posterior classification process. In real-world applications, e.g. in the medical domain or in the context of fraud detection, the minority class will usually be the class of interest. This motivates the development of techniques overcoming the challenges posed by data imbalance and ensuring an improvement of the classification performance. A considerable body of research (He & Garcia, 2009) has recently been done in this area. One prominent family of solutions are the *resampling methods*, which balance the dataset by introducing additional minority elements (*oversampling*), removing certain majority elements (*undersampling*) or a combination of both (*hybrid methods*).

On the other hand, Instance Selection (IS), a procedure that selects a subset of the instances available in the training set to learn a classifier, has been shown to yield advantageous results, by both boosting the classification performance and reducing the storage requirements (Garcia et al., 2012). IS methods may not be directly applicable in the context of imbalanced data, as they may incorrectly favor majority elements in their selection criterion and result in a large or even complete removal of the minority class.

In our work, we propose a new set of IS algorithms called IS_{Imb} methods that are tailored specifically to imbalanced data. In particular, IS_{Imb} methods are modified versions of existing IS methods, respecting their intrinsic characteristics, but ensuring that the imbalance between classes is taken into account, such that they can once again prove their uses in improving the classification performance of several learning algorithms. Our approach is related to undersampling, but does not coincide with it. IS_{Imb} will allow for the reduction of both classes in the dataset, whereas undersampling is inherently limited to the majority class. Our new methods deviate from the hybrid approach as well, as IS_{Imb} applies no oversampling whatsoever.

We are working with 33 IS methods. In an extensive experimental study including 102 datasets with varying sizes and degrees of imbalance, we compare the IS_{Imb} methods with their original forms as well as the baseline classification by several standard classifiers (kNN, C4.5 and SVM). We have also included 21 state-of-the-art methods handling imbalance, to provide a clear picture of the competitiveness of our proposal.

Our experiments show that our new methods generally yield better results compared to the original IS methods. Furthermore, the performance of the baseline classifiers can be significantly improved by preprocessing the imbalanced training sets by IS_{Imb}. Finally, several widely-used resampling techniques are also outperformed by our new methods, indicating that IS_{Imb} certainly deserves its place among the set of valid solutions to resort to when dealing with data imbalance.

References

Garcia, S., Derrac, J., Cano, J. R., & Herrera, F. (2012). Prototype selection for nearest neighbor classification: Taxonomy and empirical study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*, 417–435.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*, 1263–1284.