Impact of Local Data Characteristics on Learning Rules from Imbalanced Data

Jerzy Stefanowski

Institute of Computing Science, Poznań University of Technology,
60-965 Poznań, Poland

In this paper we discus improving rule based classifiers learned from class imbalanced data. Standard learning methods often do not work properly with imbalanced data as they are biased to focus on the majority classes while "disregarding" examples from the minority class. The class imbalance affects various types of classifiers, including the rule-based ones.

These difficulties include two groups of reasons – algorithmic and data level ones. The algorithmic factors include the following issues. First, most algorithms induce rules using the top-down technique, which hinders finding rules for smaller sets of learning examples, especially from the minority class. It is also connected with using improper evaluation measures to guide the search for best conditions in the induced rule and also for further rule pruning. Secondly, most algorithms use a greedy sequential covering approach, which may increase the data fragmentation and results in weaker rules, i.e., supported by a small number of learning examples. The "weakness" of the minority class rules influences classification strategies, where minority rules have a smaller chance to contribute to the final classification decision.

The other difficulties concern characteristics of imbalanced data distributions. Learning classifiers becomes particularly difficult when other data characteristics occur together with imbalanced distribution of classes, such as decomposition of the minority class into many rare sub-concepts, too extensive overlapping of decision classes or presence of minority class examples inside the majority class regions. In our previous study [2], these data difficulty factors have been associated with different types of examples from the minority class: safe (located in the homogeneous regions populated by the examples from one class only), borderline, rare cases and outliers.

The aim of this study is to present two different, recent algorithms proposed by K.Napierla and J.Stefanowski for inducing classification rules from imbalanced data and to show the usefulness of studying local data characteristics for two sub-tasks: (1) improving rule classifiers by incorporating types of examples into the induction strategy; (2) applying the analysis of minority class examples to identify differences in performance of these algorithms and establishing their area of competence.

The BRACID (Bottom-up induction of Rules And Cases for Imbalanced Data) algorithm is constructed following the critical analysis of limitations of current rule algorithms [3]. Its main features include: the bottom-up generalization of the most specific rules representing single examples in order to overcome the problems of data fragmentation; using two-fold rule-based and instance-based knowledge representations to better deal with borderline examples; evaluating rule candidates with measures dedicated to class imbalanced; constructing a nearest rule strategy to classify new coming examples. An important component of BRACID is also using information about the nature of the neighbouring examples. Unsafe examples from the majority classes are removed from the learning set while unsafe minority examples are treated as seeds for checking possibility of inducing additional rules. It allows us to create more rules for the minority class in unsafe regions and to diminish the possibility of overwhelming the minority class with the majority class rules. Furthermore, using the categorization of data sets obtained with analysing their local characteristics, shows that the best improvements of BRACID are observed for unsafe data sets containing many borderline examples from the minority class.

The ABMODLEM algorithm has been introduced in [1]. It adapts an idea of argument based learning, where an expert annotates some of learning examples to describe reasons for assigning them to specific classes. Using local arguments improves the consistency of rule with the domain knowledge, and also results in a better recognition of the minority class without deteriorating the majority class (which is not offered by most of current improvements of rule classifiers). Furthermore, authors introduced a strategy of active learning with the query by an ensemble to identify the examples which should be explained by an expert.

The new contribution of this paper is an experimental study, which points out that ranking of examples indicated by this ensemble strategy is consistent with results of identifying types of minority examples (based on analysing class labels inside the neighbourhood of these examples).

Furthermore, results of other comparative experiments show that ABMODLEM improves the minority class recognition, especially for difficult data distributions with rare examples and outliers while BRACID and other extensions of rule classifier work better for datasets containing mostly safe and borderline examples.

Ack.: The research was supported by NCN grant DEC-2013/11/B/ST6/00963.

References

1. Napierala, K., Stefanowski, J.: Argument Based Generalization of MODLEM Rule Induction Algorithm. In Proc. of 7th Int. Conf. RSCTC 2010, Springer, LNAI vol. 6086, 138-147 (2010).
2. Napierala, K., Stefanowski, J.: The influence of minority class distribution on learning from imbalance data. In. Proc. 7th Conf. HAIS 2012, LNAI vol. 7209, Springer, 139-150 (2012).
3. Napierala, K., Stefanowski, J.: BRACID: a comprehensive approach to learning rules from imbalanced data. Journal of Intelligent Information Systems, vol. 39 (2), 335-373 (2012).