Classification on Unbalanced Data

Martin Hlosta

Doctoral Thesis Report

This work focuses on a challenging area of machine learning – supervised learning from data where one or more classes are minority classes having extremely small number of instances. Concentrating to the case of binary classification, the thesis approaches to that filed from two different directions. The first case study solve malicious/benign file detection in the computer security domain. The second one deals with a success prediction in e-learning.

Those two studies are preceded with ***theoretical background and a state-of-the-art chapter***. The author demonstrates a deep insight into relevant areas, (2.5. and 2.5 are excellent), particularly into learning from imbalanced data and also into learning with constraints. Only small correction would be welcome: p. 11 and Fig. 2.1 - when talking about ROC, appropriate to say which of the two classifiers performs better; algorithms for learning decision trees (p. 14) - ID3 and C4.5 are actually of similar nature, however, main difference between them is not mentioned. CART differs in that it employs only binary split. For boosting (p. 15) it is worth mentioning AdaBoost. In Section 3.1, possibly a reference to the link below could be exploited http://www.site.uottawa.ca/~nat/Research/class_imbalance_bibli.html .

In Chapter 4, ***malicious file detection*** is solved as ***binary classification of imbalanced data with (performance) constraints***. As the author argues, a formulation of this task is novel and the problem itself has not been solved yet. The proposed solution combines a simple meta-learning for finding a set of initial solutions - the best initial models of cost-sensitive logistic regression (CS-LF) - with genetic algorithms (GA) to find the global maximum. In Section 4.3 particle swarm optimization (PSO) instead of GA has been employed and results of those two optimization methods are compared. With no doubts this solution is novel and looks like promising research direction in imbalanced data classification with constraints. Moreover, it works well for a particular case of malicious data recognition, as demonstrated in the thesis. A bit weaker is validation of the method. Experiments with a single dataset were performed, and two optimization techniques were evaluated on different datasets. For the version with PSO, data has been enriched - 70000 instances of minority class were added. How PSO behaves on the original data? Not even acceptance by scientific community is impressive. How PSO behaves on the original data?

The second part of this thesis finds ***imbalanced data in a e-learning***. This work aims at ***student sucess prediction early from data about un/success of other student***. In general, it is the task when a value of class is changing in time, from false (not yet successful, negative examples) to true (succesfull, positive examples). Clearly, at early stages, a number of positive examples must be very small and thus this class is a minority. As a general problem it is very interesting, and up to my knowldge, not solved yet. This work is elaborated more deeply if compared with the first part. Even the dataset is much richer and allows to accept experimental result with a confidence. Publication record of this case study is now adequate for PhD level. Two observations are introduced, a trivial one (imbalance decreases to towards the deadline) and a very interesting one - there are label data in the training set that are better candidates for removal. Domain-driven (or model-driven?) sampling method could be a real contribution to learning from imbalanced data. It would be welcome, and could result in

excellent research work, if author performed a step ahead to generalisation of this problem, not only domain-driven sampling method, as imbalanced classes appear not only in e-learning data.

Concerning a thesis structure, the author does not facilitate reading too much. To read text easily, it would need to describe main decision points (and also main motivation) first, and to postpone details to sections that come later leaving discussion to the very end (and not e.g. as a part of method description 4.2.3). The first part (Chap. 5) that describes learning analytics view, and also introductory parts of Chap.6, could be omitted as their relevance (not only to the thesis topic) is insignificant. Towards a title ot the thesis, although *unbalanced* and *imbalanced* are synonyms in general, I recommend to use an adjective *imbalanced* that is used in a machine learning community (see also your list of references). English is acceptable as well as graphic level. Typesettings is not perfect, see e.g. G-means, a comma not immediately after a formula (p. 12, p.30, p.50 etc.), parentheses (p. 17) etc. References need correction, see e.g. [37] (incomplete, wrong year as well as pp.), incomplete [14,84,98] etc.

Despite of some criticism, I can declare that the author in his dissertation thesis, mainly in its second part, has demonstrated the ability to work independently and creatively in the specified field. With respect to publication report for the learning analytics part, the author meets the standard requirements imposed on a doctoral research and I recommend to accept this text as PhD thesis.

Brno, January 22, 2018 Luboš Popelínský, FI MU