Review of the Doctoral Thesis

Classification on Unbalanced Data

submitted by Ing. Martin Hlosta

a) Is the treated topic up to date?

During recent 30 years machine learning (ML) proved its competence to learn useful models from the training data that represent most diverse classification tasks specified by industrial companies, banks, health services, etc. Nowadays, ML is expected to help in solving important practical problems that seem to resist solution attempts based on more traditional approaches. Such real-life problems are often bound by specific constraints that have to be respected both by the applied problem solving methodology and by the resulting solution. Handling unbalanced data represents an important challenge per se in this context, because such type of data occur rather often (e.g. in medical research as or in security domain) and no standard approach is ready to solve all their problems - many open questions still remain unanswered. The thesis is devoted to the following research questions:

1. How to train the classification model for the data with imbalanced classes with constraints posed on the classification performance measures?
2. How to predict achieving the goal within the specified deadline from a population of objects, in the presence of class imbalance, which is decreasing in time?

Both these questions address the problem from a truly novel perspective when setting additional constraints on classification performance measures of the designed model or when considering dynamic problems related to learning analytics. *It is no doubt that the treated topic is most actual, relevant and important.*

b) What methods have been selected?

The questions posed by the thesis generalize some of the machine learning (ML) concepts to fit them for novel applications. In the search for solutions the author tries to reuse some standard machine learning methods that he modifies and improves. The basic notions are explained in the Chapter 2. The Chapter 3 provides a solid review of the present state of art in machine learning from data represented by vectors of features – this review is based on respectful list of 112 well selected references. Further, it identifies those ML methods that seem to support the intended tasks.

c) Did the thesis succeed to achieve its target?

The Chapter 4 introduces two original approaches to the optimisation of binary classification on imbalanced data with respect to performance constraints – while one applies genetic algorithms, the other uses particle swarm approach. The initial models to be subjected to optimization are designed by cost sensitive logistic regression. The sections 4.2 and 4.3 explain the design of the corresponding novel algorithms, namely of the CC-LRGA algorithm applying genetic algorithms and two algorithms using particle swarm approach: PSO1 based on well-designed penalty function and PSO2 relying on updates that maintain constraint satisfaction. It is confirmed by extensive experiments on data from security domain that all the three methods lead to significant improvement of the performance of the resulting classifier and provide an answer to the research question 1. The most promising results seem to be provided by PSO1.

The rest of the thesis analyses the second research question. The Chapter 5 introduces the goal achieving problem with a fixed deadline, it explains in detail what is expected from the solution and selects the appropriate evaluation measure that will allow comparison of the results obtained by

diverse algorithms. The Chapter 6 applies the Self-Learning approach introduced in the section 5.2 to a real life problem of predicting at-risk students assuming absence of any legacy data. The author uses following classification algorithms to construct the requested predictive classifiers: Logistic Regression, SVM, Random Forest, Naïve Bayes and eXtreme Gradient Boosting. Performance of these predictive models on a rich dataset of Open University is carefully documented and compared. Performance of Random Forest as the winning classifier has been compared in the Section 6.6.5 to that of the Prev-Pres classifier designed using the experience from the last run of the same course. The classifier designed using Self-Learning performs reasonably well and towards the end of semester it approaches performance of the Prev-Pres classifier. This proves that the Self-Learning approach offers a good solution to the problem identified as the research question 2.

The Chapter 7 shows that performance of the Self-learning approach can be further improved if some domain knowledge is taken into account e.g. during sampling and when selecting the size of labelling window. This indicates promising direction of future research.

d) Evaluation of the presented results and their originality

The main results of the thesis have been published in three prestigious journals with IF – M. Hlosta is the first author of one of these papers and the second author of the other two papers. Moreover, he is the first author of 3 contributions and a co-author of additional 7 contributions appearing in proceedings of several international conferences, e.g. Proceedings of the 7th Int. Learning Analytics & Knowledge Conference (ACM, Vancouver 2017). This is a clear proof of importance of the treated topic as well as of high quality and originality of the achieved results and documented research.

e) What are the merits for practical applications and for further advance of science?

Distance learning and MOOCs seem to be a hope for current knowledge society where life-long is becoming a must. The suggested method of early identification of students in danger of drop-out can significantly improve efficiency of these didactic tools and contribute to their acceptance by wider group of users.

f) Additional comments

The Thesis is not an easy reading – the inclusion of an Index with abbreviations and names of the used approaches/methods would undoubtedly improve clarity of the text.

g) Can the thesis be classified as an original creative research of its author? Does it include new scientific results published by the author?

The submitted thesis describes author's original creative research, presents valuable scientific results and proves his high research competence. The chosen topic is systematically treated building on extensive amount of preliminary knowledge described in the cited literature. The Thesis meets all requirements expected by the Czech law.

I recommend accepting this thesis for the defence procedure.

Prague 11th February 2018

Prof. RNDr. Olga Štěpánková, CSc.

Katedra kybernetiky, FEL ČVUT