[1.]Elena GRESOVA, [2.]Jozef SVETLIK

COMPREHENSIVE USAGE OF DECISION TREES: THE COST-SENSITIVE MODELING

[1.]Institute of Control and Informatization of Production Processes, Faculty of Mining, Ecology, Process Control and Geotechnologies, Technical University of Kosice, Kosice, SLOVAKIA
[2.]Department of Manufacturing Machinery, Faculty of Mechanical Engineering, Technical University of Kosice, Kosice, SLOVAKIA

Abstract: The rate at which the societies globally evolve and change is immense. It entails a lot of difficult challenges in many areas. Decision trees are one of the tools that are suitable for offering the solutions for many of them. The issue of decision tree algorithms is quite extensive. Their possible implementation is extensive, as well. The present paper highlights importance, usefulness, high informative value, huge practical contribution, good interpretability and wide range of usage discussing decision trees topic with emphasis on the cost-sensitive modeling fraction. Some selected areas of their use have been specified in more detail.
Keywords: algorithm, classification, cost-sensitive, decision tree, modeling

INTRODUCTION

— Urgency of the research

Nowadays, when society develops rapidly day by day, many new pressing issues occur. Likewise a lot of long-time existing and well known matters become increasingly emergent. There are urgent topics as an information technologies, software engineering, fraudulent transactions or cost saving. Decision tree algorithms are absolutely supportive and frequently used in these domains. Thus, their research as well as applications appear to be highly actual.

— Target setting

An implementation of the cost-sensitive approaches gains benefits in solving of the numerous real-world tasks. Therefore, their modeling accompanied with proper improvements has a great potential. That is the reason why it is very important to raise awareness of the opportunities they offer.

— Actual scientific researches and issues analysis

The open source publications were reviewed for the purposes of the paper with emphasis on their thematic and empirical aspects jointly with temporal feature.

— Uninvestigated parts of general matters defining

There is still certain gap for increasing the efficiency of decision tree algorithms.

THE RESEARCH OBJECTIVE

The principal aim of this paper is to point out the broad spectrum utilization of decision trees. Specifically, the focus is on cost-sensitive decision trees. The research sphere along with practice prove a significant popularity of this tool. The dominant areas of the usage are for instance cloud computing, data mining, information security, fraud detection, direct marketing or medical science. What is more, another goals are raising awareness and stimulating progress and development in this field with link to the real applications.

THE STATEMENT OF BASIC MATERIALS

Decision trees are greatly employed in the connection with machine learning due to their understandable cognition mapping. Designing of the minimal cost decision tree signifies a key part in the context of cost-sensitive learning. Minimal cost classification is the eminent question in data mining, as well. The practice showed that it is necessary to deal with, inter alia, issues as imbalanced datasets, algorithms' performance efficiency or datasets extent.

— Wide utilization of decision trees

Decision trees represent probably the most favorite algorithms for the purposes of classification. Some instances of their employment are declared in Figure 1 (own interpretation based on [4]).

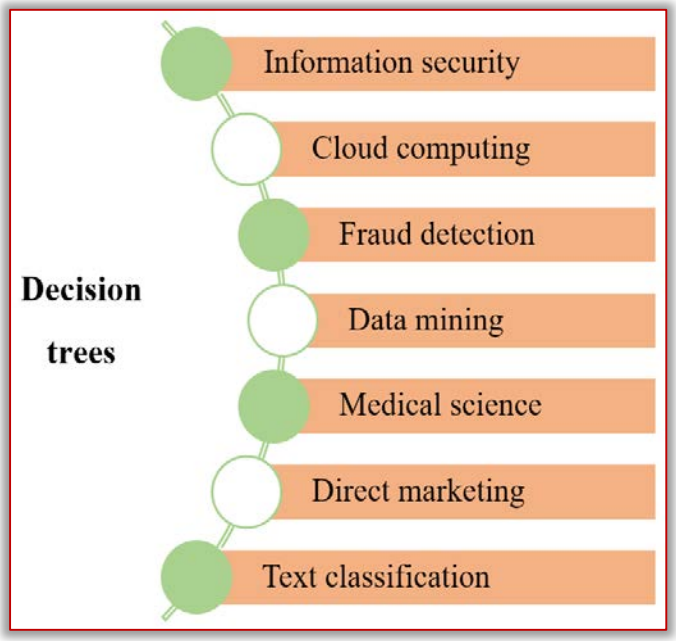

Figure 1. Selected fields of decision trees utilization

Naturally, there are additional areas of decision trees usage, as well. However, the areas above are characterized by very high incidence in scientific studies. The fact which enhances significance is the sort of these scientific studies - they do not deal only with theoretical analysis but also with real-world applications.

The great potential of modeling under decision tree algorithms highlight [2]. The connection between mathematics and economics is presented there on the trees basis. A matter of fact, the tree constitutes one of the elementary terms in a part of mathematics - combinatorics called graph theory (see e. g. [9] or [11]). Likewise the role of decision theory is pointed in this association.

— Cost-sensitive modeling

Nowadays, design of the minimal cost decision trees constitutes a decisive challenge in the context of the cost-sensitive learning. There exist numerous algorithms that deal with stated issue.

However, pursuant to [6], the algorithms proposed so far did not report adequate efficiency with regard on large datasets. Thus, the authors suggested a cost-sensitive decision tree algorithm accompanied by two adaptive mechanisms. They declared that such algorithm considerably enhances the problematical efficiency.

The subsequent research on this topic conducted by [14] provides a cost-sensitive decision tree algorithm formed by weighted class distribution with a batch deleting attribute mechanism. Based on realized experiments, the mean overall costs were diminished through using constructed algorithm in compare with using the existing ones. Therefore, same as by previous exploration, the efficiency rises.

Another sweeping question is imbalanced dataset. The real-world situations show imbalanced character of datasets in many cases. Mentioned task was undertaken for solving by [5]. Standard fixing integrates undersampling, oversampling and cost-sensitive classification. In the given paper, an effective group of cost-sensitive decision trees intended for imbalanced classification was presented. Core classifiers were built on the precise cost matrix and offered algorithm was rated on manifold benchmark datasets. Invented cost-sensitive ensemble grounded on decision trees supports improvement of the minority class cognizance plus brings handy solution for imbalanced datasets.

In connection with frequent imbalanced nature of datasets, the categorization of usually used techniques for such data is stated in Figure 2 (own interpretation based on [4]). The cost-sensitive learning with respective practices is shown there, as well.

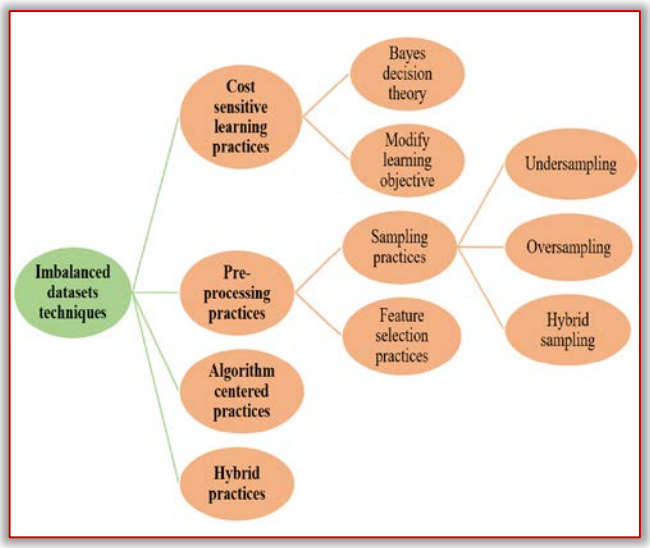

Figure 2. Categorization of the common practices applied on imbalanced datasets

— Cost-sensitive decision tree algorithms in computer technologies.

The extensive sphere where cost-sensitive decision trees can find their justification is software engineering. In the concrete, software development was discussed in the review of [7]. Vendors of the business software regularly stand face to face the situation when releasing the software outcome is needed regardless of missing some fixations of the announced defects.

The reasons are prosaic - constrained funds and deadlines, too. However, the clients escalate a minimum of proclaimed errors. The software vendors are thus forced to tackle them straightaway which entails very high expenses. The review brought original Escalation Prediction system. Its main aim was to find a solution for the maximum net profit dilemma. The cost-sensitive learning was used in relationship with meeting the aim. The conclusions of the exploration pointed that the cost-sensitive decision tree algorithm came out as the best choice. As a great success can be considered the flourishing implementation of the developed Escalation Prediction system in the product group of a business software vendor.

The software development was deliberated in the work of [12], as well. The software defect prediction was brought out there. One of the techniques suggested and consequently utilized for answering to this question was a cost-sensitive classification technique. It was represented by a collection of decision trees, so-called decision forest. The empirical verification ran and signaled the pros of all offered approaches.

The theme of software defect prediction also resonated in the research of [13] where the cost-sensitive learning was implemented. The goal was to mitigate the costs caused by forecasts. The algorithm

was based on decision forest principal what enabled taking out the information from each and every decision tree separately. What is very interesting, the execution was realized to NASA software defects.

— **Cost-sensitive decision tree algorithms in financial information technologies.**

From the financial point of view, damages are noted in relation with various kinds of frauds. Mainly information technologies expansion caused frequent and novel malpractice, inter alia, taking into consideration credit card systems. Fraud detection thus appears as highly alarming topic. Following [10] it is the key instrument for avoiding fraudulent behavior. Described current issue served as subject of interest for these researchers who developed yet unknown cost-sensitive decision tree approach. Its fulfillment was matched with the multiple popular models aimed on classification. The comparison was realized through using a feasible credit card dataset obtained from practice. Such selection underlines the relevancy hand in hand with benefits of the study. As a result of investigation and consecutive construction of the particular cost-sensitive decision tree the financial losses are expected to a lesser extent when dealing with frauds. Hence, researchers recommend the application of this technique in the fraud identification systems.

In the financial domain, another novel algorithm was suggested [1] which was typical of its example-dependent cost-sensitive nature. The fundament was constituted over the decision tree structure. Assorted example-dependent costs were included to an upstart cost-based impurity measure as well as an upstart cost-based pruning standards. The designed approach´s appreciation was done handling the real-life databases. Specifically, they were related to credit scoring, credit card fraud detection together with direct marketing. In compliance with findings presented in the paper the new algorithm reported the finest records for every single database.

— **Cost-sensitive decision tree algorithms in medicine.**

The usefulness of decision trees is striking in another field - medicine. Classification accompanied with cost reflection is enormously important in this domain. Its part, mainly in diagnostics, underline [3]. The authors attempted to identify, execute and screen an action plan of the cost-sensitive learning. They constructed an algorithm for decision tree induction which involved a several types of costs such as delayed costs, test costs or costs arising from risks. Next step was the action plan implementation with intention to practice and appraise cost-sensitive decision trees using medical data. Modeled trees serve for verification in accordance with particular action plans that incorporate special costs, ordinary costs and group costs.

The practicability of cost-sensitive decision trees is confirmed also by [8] in medical diagnosis. The effort was costs reduction concerning misdiagnosis as well as medical tests. To support this effort, the cost-sensitive machine learning algorithms were proposed for subsequent diagnosis modeling. Medical tests figured as attributes - the attribute costs were bounded there. Misdiagnosis represented misclassifications with corresponding misclassification costs. In this connection, various test strategies were processed. What is important, mentioned strategies accorded with diverse cases from diagnosis practice. Even more, they have undergone an empirical assessment and declared their effectiveness. This results to the instantaneous possibility for usage in real medical diagnosis.

CONCLUSION

In the sphere of classification, decision trees represent one of the most favorite and used algorithms. The presented paper dealt with decision trees and their comprehensive usage. Mentioned theme was examined from the cost-sensitive point of view. The numerous fields of decision trees utilization were stated and some of them were afterward analyzed. The modeling under cost-sensitive regard was disclosed and selected associated problems were outlined. The central attention was dedicated to the cost-sensitive decision tree algorithms in computer technologies, financial information technologies and medicine.

Acknowledgment

This work was supported by the Slovak Research and Development Agency under the Contract no. APVV-18-0413.

References

[1] Bahnsen, A.C., Aouada, D., Ottersten, B. (2015). Example-dependent cost-sensitive decision trees. Expert Systems with Applications, 42(19), 6609-6619.

[2] Drabiková, E., Škrabul'áková, E.F. (2017, May). Decision trees-a powerful tool in mathematical and economic modeling. In 2017 18th International Carpathian Control Conference (ICCC) (pp. 34-39). IEEE.

[3] Freitas, A., Costa-Pereira, A., Brazdil, P. (2007, September). Cost-sensitive decision trees applied to medical data. In International Conference on Data Warehousing and Knowledge Discovery (pp. 303-312). Springer, Berlin, Heidelberg.

[4] Kaur, H., Pannu, H.S., Malhi, A. K. (2019). A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions. ACM Computing Surveys (CSUR), 52(4), 36 pp.

[5] Krawczyk, B., Woźniak, M., Schaefer, G. (2014). Cost-sensitive decision tree ensembles for effective imbalanced classification. Applied Soft Computing, 14, Part C, 554-562.

[6] Li, X., Zhao, H., Zhu, W. (2015). A cost sensitive decision tree algorithm with two adaptive

mechanisms. Knowledge-Based Systems, 88, 24-33.
[7] Ling, C.X., Sheng, V.S., Bruckhaus, T., Madhavji, N. H. (2006, August). Maximum profit mining and its application in software development. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 929-934). ACM.
[8] Ling, C.X., Sheng, V.S., Yang, Q. (2006). Test strategies for cost-sensitive decision trees. IEEE Transactions on Knowledge and Data Engineering, 18(8), 1055-1067.
[9] Peterin, I., Schreyer, J., Škrabul'áková, E.F., Taranenko, A. (2018). A note on the Thue chromatic number of lexicographic products of graphs. Discussiones Mathematicae Graph Theory, 38(3), 635-643.
[10] Sahin, Y., Bulkan, S., Duman, E. (2013). A cost-sensitive decision tree approach for fraud detection. Expert Systems with Applications, 40(15), 5916-5923.
[11] Schreyer, J., Škrabul'áková, E. (2015). Total Thue colourings of graphs. European Journal of Mathematics, 1(1), 186-197.
[12] Siers, M.J., Islam, M.Z. (2015). Software defect prediction using a cost sensitive decision forest and voting, and a potential solution to the class imbalance problem. Information Systems, 51, 62-71.
[13] Siers, M.J., Islam, M. Z. (2018). Novel algorithms for cost-sensitive classification and knowledge discovery in class imbalanced datasets with an application to NASA software defects. Information Sciences, 459, 53-70.
[14] Zhao, H., Li, X. (2017). A cost sensitive decision tree algorithm based on weighted class distribution with batch deleting attribute mechanism. Information Sciences, 378, 303-316.

ACTA TECHNICA CORVINIENSIS – Bulletin of Engineering
ISSN: 2067-3809