Component-Based Decision Trees:
Empirical Testing on Data Sets of Account Holders in the Montenegrin Capital Market

Ljiljana Kašćelan, Faculty of Economics, University of Montenegro, Podgorica, Montenegro

Vladimir Kašćelan, Faculty of Economics, University of Montenegro, Podgorica, Montenegro

ABSTRACT

Popular decision tree (DT) algorithms such as ID3, C4.5, CART, CHAID and QUEST may have different results using same data set. They consist of components which have similar functionalities. These components implemented on different ways and they have different performance. The best way to get an optimal DT for a data set is one that use component-based design, which enables user to intelligently select in advance implemented components well suited to specific data set. In this article the authors proposed component-based design of the optimal DT for classification of securities account holders. Research results showed that the optimal algorithm is not one of the original DT algorithms. This fact confirms that the component design provided algorithms with better performance than the original ones. Also, the authors found how the specificities of the data influence the DT components performance. Obtained results of classification can be useful to the future investors in the Montenegrin capital market.

Keywords: Classification, Component Design, Data Mining, Decision Tree, Securities Account Holders

1. INTRODUCTION

Popular DT algorithms are implemented with "black-box"-approach, which implies that the logic is hidden from users and there is no possibility of ad hoc changes. These algorithms are changed and improved incrementally, and that can take a long time.

Component-based design implies "white-box"-construction of algorithms with help of standardized reusable components (RCs) obtained from original DT algorithms (Delibasic, Jovanovic, Vukicevic, Suknovic & Obradovic, 2011; Suknovic, Delibasic, Jovanovic, Vukicevic, Becejski-Vujaklija & Obradovic, 2012). It enables combination of advantages of various algorithms and their comparison, as well as the testing of influence of certain components on performance of the algorithms. Combining these components we can improve performance and get optimal DT algorithm for a specific data set.

DOI: 10.4018/IJORIS.2015100101

Copyright © 2015, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

In their research, Delibasic, Jovanovic, Vukicevic, Suknovic and Obradovic (2011) tested component design on 15 different data sets. They left an opened research question how the specificities of these data sets influence the DT components performance.

In this article, with the component-based design, we designed different DT algorithms and analyzed their results for data sets from Montenegrin capital market. By testing we determined the influence of individual components of DT algorithms on classification performance for observed data sets. The main contribution of our study is that we found how the performance of DT algorithms depends on the specificities of the data on which they are applied. We also showed that algorithms obtained with the component-based design can provide higher accuracy of classification, as well as the lower complexity of generated tree, than the original DT algorithms. Based on obtained classification results we determined features of investors and their corresponding portfolio for defined classes, and that can be useful for making future investment strategies for all stakeholders in the capital market.

This study is organized as follows. In the next section we give a review of related works, which served as a motivation for our research. In the third section we introduce the component-based design and a method for empirical testing. In the fourth section we define used tools and data sets, present the results of empirical testing and discuss the obtained results. In the conclusion we give our final considerations with real benefits of the study.

2. RELATED WORK

Existing DT algorithms are usually implemented with "black-box"- approach. Thereby the user specifies input data and parameters used for definition of appropriate model. Induction procedure is hidden from the user. The user has no possibility to change algorithms in order to improve such results. These algorithms are very hard to analyze and evaluate, because it is hard to determine which part of the algorithm had influence on its performance. Certain part of the algorithm is the best with one data set, while with the other data set, corresponding part of some other algorithm can be better. In this approach performance testing of different parts of algorithms over data sets, as well as combination of the most efficient parts from different algorithms, are not possible. Better performance with these algorithms can be achieved with incremental improvement of existing algorithms.

One of the first "black-box" DT algorithms is ID3 (Quinlan, 1986). This algorithm works only with categorical variables, it is based on "multi-way"-split and it uses "Information Gain"-measure for split quality. This evaluation measure is biased towards choosing attributes with more categories. Breiman, Friedman, Stone and Olshen (1984), proposed CART algorithm which works with both categorical and numerical variables, and for split evaluation it uses "Gini" measure. The algorithm supports only "binary"- splits. Algorithm C4.5 (Quinlan, 1993) is improvement of ID3 algorithm which can work both, with categorical and numerical data. It uses "multi-way"- split for categorical, and "binary" for numerical data. For split evaluation it uses "Gain Ratio"-measure, which is not biased towards attributes with several categories. It also includes three pruning algorithms: reduced error pruning, pessimistic error pruning and error based pruning. CHAID algorithm was proposed by Kass (1980). In this algorithm Chi-square test is used for evaluation of the split quality. QUEST algorithm (Loah & Shih, 1997), uses removal of insignificant attributes with chi-square test, for categorical, and ANOVA f-test, for numerical data.

The method we used in this article is "white box", i.e. component-based design of algorithms for DT induction. "White box"- approach enables the user to define "building blocks", i.e. RCs

Copyright © 2015, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

16 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/article/component-based-decision-trees/133602

Related Content

Employer Branding: Pioneering HR Innovations
Pallavi Pahuja, Ruby Sharmaand Aaruni Batta (2024). *Innovative Technologies for Increasing Service Productivity (pp. 212-219).*
www.irma-international.org/chapter/employer-branding/341252

Corporate Governance Challenges for Small and Medium Enterprises in the Constrained Zimbabwean Economy
Mufaro Dzingirai, Shingirai Sikomweand Noah Tshuma (2022). *International Journal of Applied Management Sciences and Engineering (pp. 1-14).*
www.irma-international.org/article/corporate-governance-challenges-for-small-and-medium-enterprises-in-the-constrained-zimbabwean-economy/299024

A Unified Classification Ecosystem for Auctions
Dimitrios M. Emirisand Charis A. Marentakis (2010). *International Journal of Operations Research and Information Systems (pp. 53-74).*
www.irma-international.org/article/unified-classification-ecosystem-auctions/45763

Highway Alignment Optimization Using Cost-Benefit Analysis Under User Equilibrium
Avijit Majiand Manoj K. Jha (2011). *International Journal of Operations Research and Information Systems (pp. 19-33).*
www.irma-international.org/article/highway-alignment-optimization-using-cost/58893

The Relationship Between Bitcoin and Stock Market
Xin Wang, Xi Chenand Peng Zhao (2020). *International Journal of Operations Research and Information Systems (pp. 22-35).*
www.irma-international.org/article/the-relationship-between-bitcoin-and-stock-market/250246