Visualization of Decision Trees that Analyze Medical Data

Sungyun Bae[*], Seongmin Mun[§], Gyeongcheol Choi[†], Suhyun Lim[‡], Sunjoo Bang[¶], Sangjoon Son[**], Changhyung Hong[§§], Hyunjung Shin[††], Kyungwon Lee[‡‡]
Life media interdisciplinary program[*,§,†,‡], UMR 7114 MoDyCo - CNRS[§], Department of industrial engineering[¶,††], Department of psychiatry[**,§§], Department of digital media[‡‡]
Ajou university, University Paris Nanterre[§], South korea, France[§]
{roah[*], stat34[§], ckc6842[†], hyun0979[‡], smalsunjoo[¶], sjsonpsy[**], antiagint[§§], shin[††], kwlee[‡‡]}@ajou.ac.kr

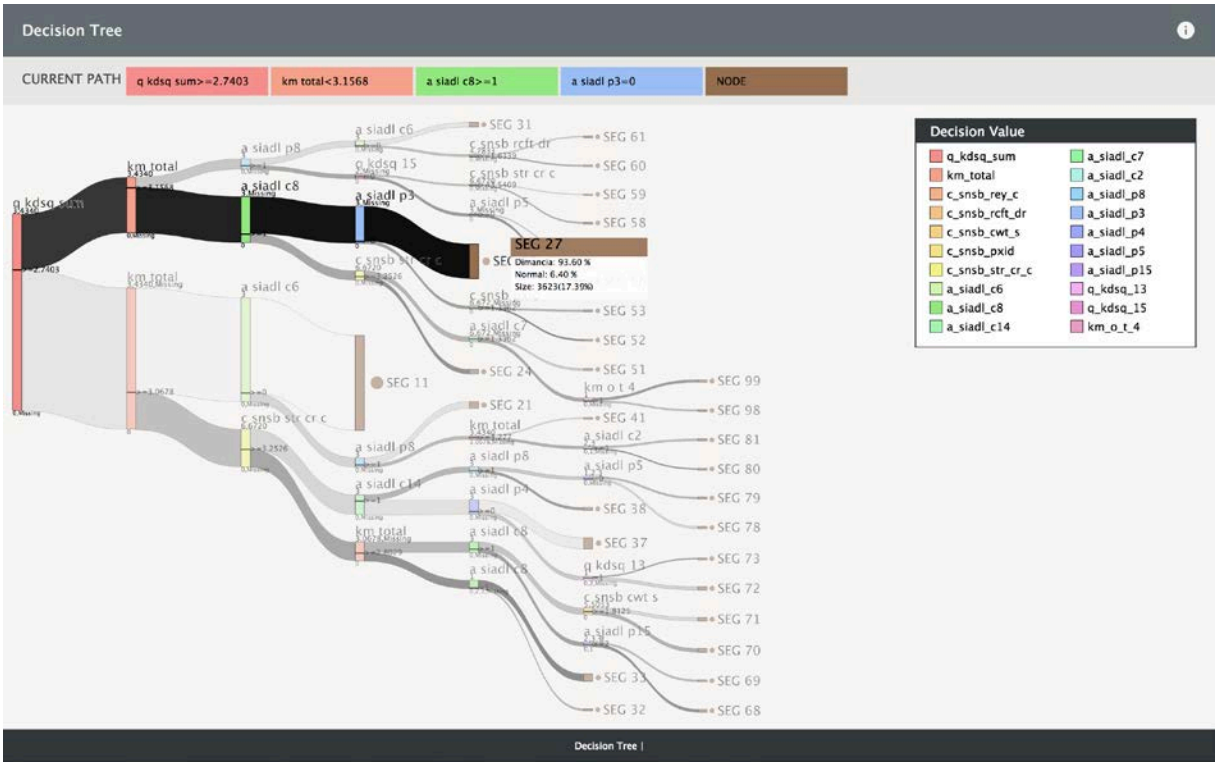

Figure 1: A visualization of the decision-making tree analysis using the Sankey diagram. The left shows a visualization of the decision trees, and the upper right shows the index of the decision variables.

ABSTRACT

A decision tree is one of the predictive models in data mining and has been widely used in recent medical-data analyses [1]. In the finding of an important node, the existing decision-tree visualizations are used to show the ratio of the sample number to the target variable. Additional decision-variable information, however, is needed since the decision tree for the analysis of the medical data is important for the identification of the variables. Therefore, this study proposes a visualization that can be used to easily find the important terminal nodes and grasp the decision variables.

CCS CONCEPTS

• **Information systems** → **Information-system applications**; *Decision-support systems*; Expert systems

KEYWORDS

Information visualization, Big-data visualization, Visual analytics, Decision-support systems

1 INTRODUCTION

A decision tree is easier to interpret and can more easily interpret data than other statistical methods, and this is because the tree structure is used as the starting point to interpret the data from the root node to the terminal node. The form of the analysis results can be easily understood and utilized in comparison with other quantitative methods. The decision-tree analysis has been widely used in recent medical data for prediction and classification values [2]. It is characterized by the economic advantage of a reduction of the number of test items, especially with the diagnosis of diseases. To demonstrate the merits of such decision trees, it is more important to grasp the decision variables in the existing decision tree. In the existing node-linked diagram form, it is difficult to grasp the variables. The visualization presented in this study emphasizes the explanations of the decision variables rather than the existing visualizations, enabling the identification of the important variables through visualization.

2 Visualization

2.1 Design Guidelines

The visualization sample data was obtained using data where the dementia rate is the target variable and the psychological scores of the demented patients are the input variables. The CHAID-using regression tree that is among the decision-tree algorithms is visualized because the target variable, the dementia/patient ratio, is the continuous type. Figure 1 shows the visualization.

2.2 Visualization

The sample data for the visualization was formulated using data where the dementia rate is the target variable and the psychological scores of the demented patients are the input variables [1]. The CHAID-using regression tree is visualized from among the decision-tree algorithms because the target variable, the dementia/patient ratio, is of the continuous type. Figure 2 shows the visualization.

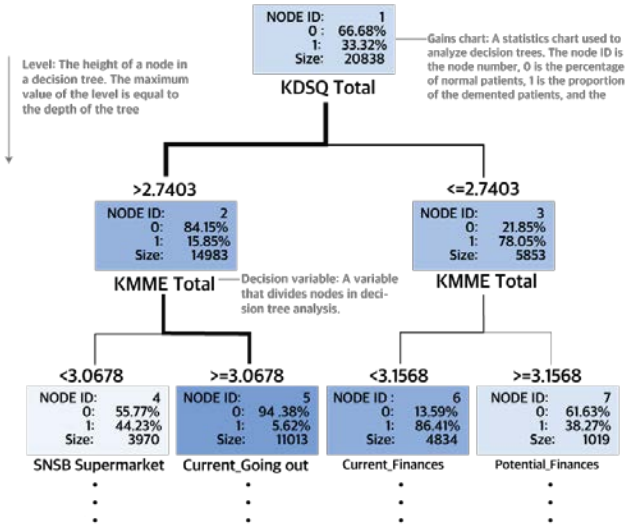

Figure 2: A part of the SAS-provided visualization of the decision-tree analysis in the form of a node-link diagram. The larger the line, the greater the number of involved patients. The inequality above the node represents the partitioning criterion in the decision variable.

3 CONCLUSIONS

The visualization that was formed for this study visually improved the expression of the information that is not intuitive in the existing graph. By improving a number of the expressions of the statistics, decision variables, and hierarchical structures, it became possible to quickly find the main terminal nodes. It is inconvenient, however, to compare the statistical information to the mouse, because it is difficult to compare the nodes to the small number of observations.

ACKNOWLEDGMENTS

This work was supported by the 2017 BK21 Program, Ajou University.

REFERENCES

[1] Bang, S., Son, S., Roh, H., Lee, J., Bae, S., Lee, K., &Shin, H. (2017). Quad-phased data mining modeling for dementia diagnosis. BMC Medical Informatics and Decision Making, 17(1), 60.

[2] Bhojani, S. H., & Bhatt, N. (2016). Data Mining Techniques and Trends–A Review. Global Journal For Research Analysis, 5(5).