Visualization Based Sankey Diagram Decision Making Tree Analysis

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§

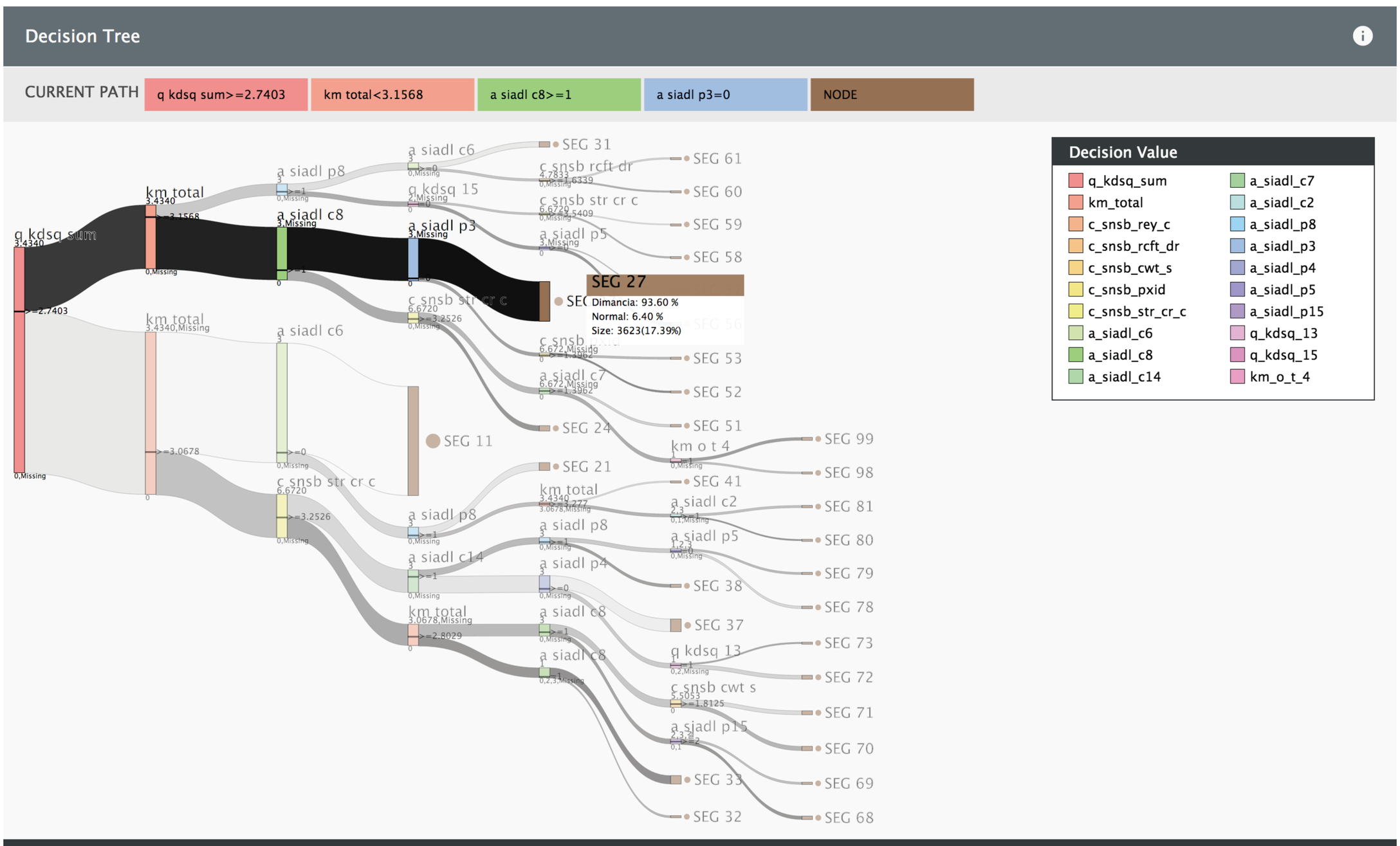

Figure 1: A visualization of the decision tree analysis using the Sankey diagram. The left shows a visualization of the decision trees, and the upper right shows the index of the decision variables.

Abstract

A decision tree 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.

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.

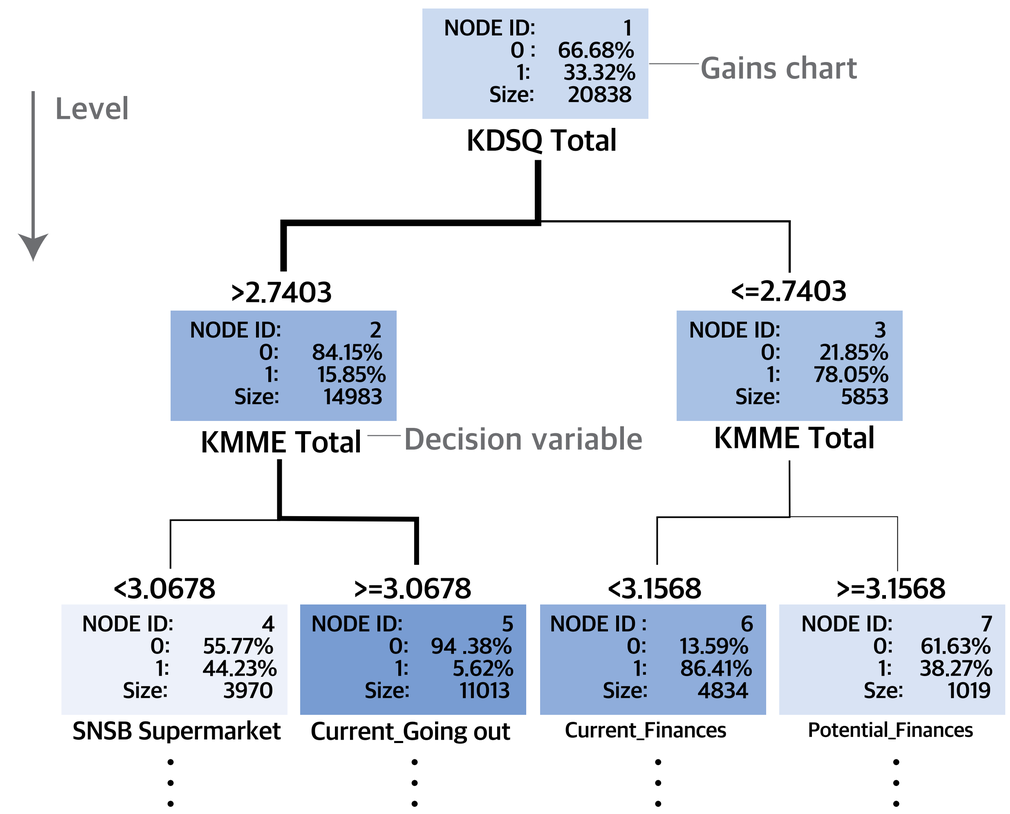

Figure 2: A part of the SAS-provided visualization of the decision-tree analysis in the form of a node-link diagram. The larger the line, the greater the number of involved patients. The inequality above the node represents the partitioning criterion in the decision variable.

- **Gains chart**

A statistics chart used to analyze decision trees.
Gains chart usually have node IDs,'%' for each class, and the number of nodes the node contains.

- **Decision variable**

A variable that divides nodes in decision tree analysis.
The first is the most important variable in determining the target variable. The same variable can appear repeatedly in a tree.

- **Level**

The height of a node in a decision tree.
The maximum value of the level is equal to the depth of the tree.

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 [2]. 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.

Conclusion

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.

References

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

[2] 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.