CHAID

Chi-square Automatic Interaction Detector (CHAID) was a technique created by Gordon V. Kass in 1980. CHAID is a tool used to discover the relationship between variables. CHAID analysis builds a predictive medel, or tree, to help determine how variables best merge to explain the outcome in the given dependent variable. In CHAID analysis, nominal, ordinal, and continuous data can be used, where continuous predictors are split into categories with approximately equal number of observations. CHAID creates all possible cross tabulations for each categorical predictor until the best outcome is achieved and no further splitting can be performed. In the CHAID technique, we can visually see the relationships between the split variables and the associated related factor within the tree. The development of the decision, or classification tree, starts with identifying the target variable or dependent variable; which would be considered the root. CHAID analysis splits the target into two or more categories that are called the initial, or parent nodes, and then the nodes are split using statistical algorithms into child nodes. Unlike in regression analysis, the CHAID technique does not require the data to be normally distributed.

Merging: In CHAID analysis, if the dependent variable is continuous, the *F* test is used and if the dependent variable is categorical, the chi-square test is used. Each pair of predictor categories are assessed to determine what is least significantly different with respect to the dependent variable. Due to these steps of merging, a Bonferroni adjusted p-value is calculated for the merged cross tabulation.

Decision tree components in CHAID analysis:

In CHAID analysis, the following are the components of the decision tree:

1. **Root node:** Root node contains the dependent, or target, variable. For example, CHAID is appropriate if a bank wants to predict the credit card risk based upon information like age, income, number of credit cards, etc. In this example, credit card risk is the target variable and the remaining factors are the predictor variables.
2. **Parent's node:** The algorithm splits the target variable into two or more categories. These categories are called parent node or initial node. For the bank example, high, medium and low categories are the parent's nodes.
3. **Child node:** Independent variable categories which come below the parent's categories in the CHAID analysis tree are called the child node.
4. **Terminal node:** The last categories of the CHAID analysis tree are called the terminal node. In the CHAID analysis tree, the category that is a major influence on the dependent variable comes first and the less important category comes last. Thus, it is called the terminal node.

Related Pages:

- Cluster Analysis