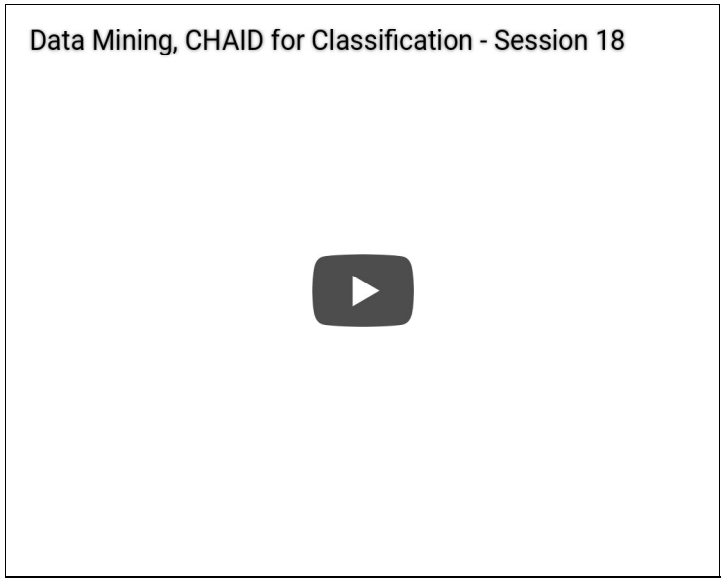
Popular Decision Tree: CHAID Analysis, Automatic Interaction Detection



General CHAID Introductory Overview

The acronym CHAID stands for *Chi*-squared Automatic Interaction Detector. It is one of the oldest tree classification methods originally proposed by Kass (1980). According to Ripley, 1996, the CHAID algorithm is a descendent of THAID developed by Morgan and Messenger, (1973). CHAID will "build" non-binary trees (i.e., trees where more than two branches can attach to a single root or node), based on a relatively simple algorithm that is particularly well suited for the analysis of larger datasets. Also, because the CHAID algorithm will often effectively yield many multi-way frequency tables (e.g., when classifying a categorical response variable with many categories, based on categorical predictors with many classes), it has been particularly popular in marketing

Variance Components
Statistical Advisor
Distribution Tables
References Cited
Send Comments
Business Solutions
Free Resources
About Textbook

research, in the context of market segmentation studies.

Both CHAID and C&RT techniques will construct trees, where each (non-terminal) node identifies a split condition, to yield optimum prediction (of continuous dependent or response variables) or classification (for categorical dependent or response variables). Hence, both types of algorithms can be applied to analyze regression-type problems or classification-type.

CHAID is a recursive partitioning method. To index

Basic Tree-Building Algorithm: CHAID and Exhaustive CHAID

The acronym CHAID stands for *Chi*-squared Automatic Interaction Detector. This name derives from the basic algorithm that is used to construct (non-binary) trees, which for classification problems (when the dependent variable is categorical in nature) relies on the *Chi*-square test to determine the best next split at each step; for regression-type problems (continuous dependent variable) the program will actually compute F-tests. Specifically, the algorithm proceeds as follows:

Preparing predictors. The first step is to create categorical predictors out of any continuous predictors by dividing the respective continuous distributions into a number of categories with an approximately equal number of observations. For categorical predictors, the categories (classes) are "naturally" defined.

Merging categories. The next step is to cycle through the predictors to determine for each predictor the pair of (predictor) categories that is least significantly different with respect to the dependent variable; for classification problems (where the dependent variable is categorical as well), it will compute a *Chi*-square test (Pearson *Chi*-square); for regression problems (where the dependent variable is continuous), F tests. If the respective test for a given pair of predictor categories is not statistically significant as defined by an alpha-to-merge value, then it will merge the respective predictor categories and repeat this step (i.e., find the next pair of categories, which now may include previously merged categories). If the statistical significance for the respective pair of predictor categories is significant (less than the respective alpha-to-merge value), then (optionally) it will compute a Bonferroni adjusted *p*-value for the set of categories for the respective predictor.

Selecting the split variable. The next step is to choose the split the predictor variable with the smallest adjusted *p*-value, i.e., the predictor variable that will yield the most significant split; if the smallest (Bonferroni) adjusted *p*-value for any predictor is greater than some alpha-to-split value, then no further splits will be performed, and the respective node is a terminal node.

Continue this process until no further splits can be performed (given the alpha-to-merge and alpha-to-split values).

CHAID and Exhaustive CHAID Algorithms. A modification to the basic CHAID algorithm, called Exhaustive CHAID, performs a more thorough merging and testing of predictor variables, and hence requires more computing time. Specifically, the merging of categories continues (without reference to any alpha-to-merge value) until only two categories remain for each predictor. The algorithm then proceeds as described above in the *Selecting the split variable* step, and selects among the predictors the one that yields the most significant split. For large datasets, and with many continuous predictor variables, this modification of the simpler CHAID algorithm may require significant computing time.

To index

General Computation Issues of CHAID

Reviewing large trees: Unique analysis management tools. A general issue that arises when applying tree classification or regression methods is that the final trees can become very large. In practice, when the input data are complex and, for example, contain many different categories for classification problems, and many possible predictors for performing the classification, then the resulting trees can become very large. This is not so much a computational problem as it is a problem of presenting the trees in a manner that is easily accessible to the data analyst, or for presentation to the "consumers" of the research.

Analyzing ANCOVA-like designs. The classic CHAID algorithms can accommodate both continuous and categorical predictor. However, in practice, it is not uncommon to combine such variables into analysis of variance/covariance (ANCOVA) like predictor designs with main effects or interaction effects for categorical and continuous predictors. This method of analyzing coded ANCOVA-like designs is relatively new. However, it is easy to see how the use of coded predictor designs expands these powerful classification and regression techniques to the analysis of data from experimental.

To index

CHAID, C&RT, and QUEST

For classification-type problems (categorical dependent variable), all three algorithms can be used to build a tree for prediction. QUEST is generally faster than the other two algorithms, however, for very large datasets, the memory requirements are usually larger, so using the QUEST algorithms for classification with very large input data sets may be impractical.

For regression-type problems (continuous dependent variable), the QUEST algorithm is not applicable, so only CHAID and C&RT can be used. CHAID will build non-binary trees that tend to be "wider". This has made the CHAID method particularly popular in market research applications: CHAID often yields many terminal nodes connected to a single branch, which can be conveniently summarized in a simple two-way table with multiple categories for each variable or dimension of the table. This type of display matches well the requirements for research on market segmentation, for example, it may yield a split on a variable *Income*, dividing that variable into 4 categories and groups of individuals belonging to those categories that are different with respect to some important consumer-behavior related variable (e.g., types of cars most likely to be purchased). C&RT will always yield binary trees, which can sometimes not be summarized as efficiently for interpretation and/or presentation.

As far as predictive accuracy is concerned, it is difficult to derive general recommendations, and this issue is still the subject of active research. As a practical matter, it is best to apply different algorithms, perhaps compare them with user-defined interactively derived trees, and decide on the most reasonably and best performing model based on the prediction errors. For a discussion of various schemes for combining predictions from different models, see, for example, Witten and Frank, 2000.

To index