SPSS®

white paper

AnswerTree Algorithm Summary

AnswerTree offers four powerful algorithms that enable you to build the best model, for any type of data. This document summarizes and compares the different types of algorithms used in AnswerTree: CHAID, Exhaustive CHAID, Classification and Regression Trees (C&RT) and QUEST.

What is a Classification Tree?

A classification tree is an empirical rule for predicting the class of an object from values of predictor variables.

Common features of classification tree methods

- *Merging* – relative to the target variable, non-significant predictor categories are grouped with the significant categories.
- *Splitting* – selecting the split point. Variable to split population is chosen by comparison to all others.
- *Stopping* – rules which determine how far to extend the splitting of nodes.
- *Pruning* – removing branches that add little to the predictive value of the tree are removed.

Validation and error estimation

The methods used for evaluating and comparing a given classifier are the same regardless of which methods are used for generation. Measurement of true error vs. apparent error, and validation of classifiers using separate or resampled data are performed identically for CHAID, C&RT, and QUEST.

CHAID Method

CHAID (Chisquare-Automatic-Interaction-Detection) was designed to handle categorical variables only. SPSS has extended both algorithms to handle nominal categorical, ordinal categorical and continuous dependent variables. Continuous independent variables are discretized prior to evaluation.

CHAID algorithms

A CHAID tree is a decision tree that is constructed by splitting subsets of the space into two or more child (nodes) repeatedly, beginning with the entire data set.

CHAID

To determine the best split at any node, Kass (1980) merges any allowable pair of categories of the predictor variable (the set of allowable pairs is determined by the type of predictor variable being studied) if there is no statistically significant difference within the pair with respect to the target variable. The process is repeated until no non-significant pair is found. The resulting set of categories of the predictor variable is the best split with respect to that predictor variable.

This process is followed for all predictor variables. The split that is the best prediction is selected, and the node is split. The process repeats recursively until one of the stopping rules is triggered. Kass's approach does save computer time. However, it is not guaranteed to find the split that is the "best" at each node. Only an exhaustive search of all possible category subsets will accomplish this.

Exhaustive CHAID

Exhaustive search CHAID was proposed by Biggs et al. (1991). Biggs suggests finding the best split by merging similar pairs continuously until only a single pair remains. The set of categories with the largest significance is taken to be the best split for that predictor variable.

This process is followed for all predictor variables. T predictor that gives the best prediction is selected, and the node is split. The process repeats recursively until one of the stopping rules is triggered.

Components of CHAID Analysis

Basic components in a CHAID analysis are as follows:

- *One or more predictor variables.* Predictor variables can be continuous, ordinal categorical, or nominal categorical.
- *One target variable.* The target variable can be nominal categorical, ordinal categorical or continuous. The order characteristics of the target variable define the method to be used in the segmentation model. Variations of the algorithm are used to take advantage of the information contained it nominal, ordinal and continuous variables.
- *Settings for various CHAID parameters.* The settings include significance levels used in merging and splitting and a criterion to stop the splitting process. Case weight variables and frequency weight variables are implemented.

Comments for CHAID Analysis

Use CHAID if:

- You want to find non-binary splits.
- The classification model that is produced by CHAID is measurably better than that which is produced by the other methods.

Classification and Regression Tree (C&RT) Method

The Classification and Regression Trees (C&RT) method of Breiman et al. (1984) generates binary decision trees. The C&RT tree is constructed by splitting subsets of the data set using all predictor variables to create two child nodes repeatedly, beginning with the entire data set. The best predictor is chosen using a variety of impurity or diversity measures. The goal is to produce subsets of the data which are as homogeneous as possible with respect to the target variable.

The C&RT algorithm

Fore each split, each predictor is evaluated to find the best cut point (continuous predictors) or groupings of categories (nominal and ordinal predictors) based on improvement

score, or reduction in impurity. Then the predictors are compared, and the predictor with the best improvement is selected for the split. The process repeats recursively until one of the stopping rules is triggered.

Components of C&RT analysis

Basic components in a C&RT analysis are as follows:

- *One or more predictor variables.* Predictor variables can be continuous, ordinal categorical, nominal categorical variables.
- *One target variable.* The target variable can be nominal categorical, ordinal categorical or continuous variable. The nature of the target variable usually defines the method to be used in the segmentation model.
- *Settings for various C&RT parameters.* The settings include priors for categorical target variable, impurity measures and misclassification costs, the variable used as the case weight variable (if any), and the variable used as frequency weight variable.

Comments for C&RT Analysis

Use C&RT if:

- You want to restrict your tree to binary splits.
- The classification model produced by C&RT is measurably better than that which is produced by the other methods.
- You want cost matrices to be considered for variable selection.
- Cost complexity pruning or direct stopping rules are required.

QUEST

QUEST stands for Quick, Unbiased, Efficient, Statistical Tree. The original method is described in Loh and Shih (1997). It is a tree-structured classification algorithm that yields a binary decision tree like C&RT. The reason for yielding a binary tree is that a binary tree may allows techniques such as pruning, direct stopping rules and surrogate splits to be used. Unlike CHAID and C&RT, which handle variable selection and split point selection simultaneously during the tree growing process, QUEST deals with them separately.

It is well known that exhaustive search methods such as C&RT tend to select variables with more discrete values, which can afford more splits in the tree growing process. This introduces bias into the model, which reduces the generalizability of results. Another limitation of C&RT is the computational investment in searching for splits. QUEST method is designed to address these problems. QUEST was demonstrated to be much better than exhaustive search methods in terms of variable selection bias and computational cost. In terms of classification accuracy, variability of split points and tree size, however, there is still no clear winner when univariate splits are used.

QUEST algorithm

For each split, the association between each predictor variable and the target is computed using the ANOVA F-test or Levene's test (for ordinal and continuous predictors)

or Pearson's chi-square (for nominal predictors). If the target variable is multinomial, two-means clustering is used to create two superclasses. The predictor having the highest association with the target variable is selected for splitting. Quadratic Discriminant Analysis (QDA) is applied to find the optimal splitting point for the predictor variable. The process repeats recursively until one of the stopping rules is triggered.

Components of QUEST analysis

Basic components in a QUEST analysis are as follows:

- *One or more predictor variables.* Predictor variables can be continuous, ordinal or nominal variables.
- *One target variable.* The target variable must be nominal.
- *Settings for various QUEST parameters.* The settings include alpha level for variable selection, priors for the categorical target variable, profit values and misclassification costs, and the variable used as the frequency weight variable.

Comments on QUEST analysis

Use Quest:

- With categorical dependent variables only.
- If it is important to have an unbiased tree.
- If you have a large or complex data set and need an efficient algorithm for estimating the tree.
- Or C&RT, if you want to restrict your tree to binary splits.
- If the classification model produced by QUEST is measurably better than that produced by the other methods.
- Or C&RT to handle missing values by surrogate splits.
- Case weight is ignored in the QUEST option.
- As in C&RT, direct stopping rules and cost-complexity pruning can be applied to a QUEST tree.
- Like CHAID, the cost matrix is not directly involved in the QUEST tree-growing process. However, for a symmetric cost matrix, cost information can be incorporated into the model by adjusting the priors based on the cost matrix.

Stopping Rules

Each of the methods recursively splits nodes until one of the stopping rules is triggered. The following conditions will cause the algorithm to terminate:

- The maximum tree depth has been reached.
- No more splits can be made, because all terminal nodes meet one or more of the following conditions:
  - There is no significant predictor variable left to split the node.
  - The number of cases in the terminal node is less than the minimum number of cases for parent nodes.
  - If the node were split, the number of cases in one or more child nodes would be less than the minimum number of cases for child nodes.

Validation and risk estimation

Once a tree has been built, its predictive value can be assessed. The same methods of assessing risk (error) are used for all tree-growing methods.

- *Nominal and ordinal target variables.* For categorical targets, each node assigns a predicted category to all cases belonging to it. The risk estimate is the proportion of all cases incorrectly classified.
- *Continuous target variables.* For continuous targets, each node predicts the value as the mean value of cases in the node. The risk estimate is the within-node variance about each node's mean, averaged over all nodes. (In other words, it is the mean squared error within nodes.)

Two verification methods are available in AnswerTree: partitioning and cross-validation. These methods allow you to estimate how well you can expect your tree to generalize new data.

- *Partitioning.* This method requires you to "set aside" part of your data when building the tree. When the tree-growing process is complete, a risk estimate is computed based on classifying the held-out data with the tree. This can help you identify models that overfit the data – that is, models that incorporate idiosyncratic information from your specific sample that does not apply to other data or the population as a whole.
- *Cross-validation*. This method uses all of the data to build the tree. The risk estimate is computed by partitioning the data into k separate groups or folds (where k is specified by the user). Next, k trees are built using the same growing criteria as the tree being evaluated. The first tree uses all folds except the first, the second tree uses all folds except the second, and so on, until each fold has been excluded once. For each of these trees, a risk estimate is computed, and the cross-validated risk estimate is the average of these k risk estimates for the k trees, weighted by number of cases in each fold.

About SPSS

SPSS Inc. is a leader in the large markets for analytic software: business intelligence (data mining and market research), quality improvement and scientific research. SPSS products and services transform organizations by helping users leverage information to grow revenues and improve processes. More than two million people use SPSS, the world's best-selling tool for desktop analysis, to create and distribute information for better decision making.

Celebrating its 30th anniversary in 1998, SPSS is based in Chicago with more than 40 offices, distributors and partners worldwide. Products run on leading computer platforms and many are translated into 10 local languages. In 1997, the company employed nearly 800 people worldwide and generated net revenues of approximately $110 million.

Contacting SPSS

To place an order or to get more information, call your nearest SPSS office or visit our World Wide Web site at **www.spss.com**

SPSS Inc.		+1.312.651.3000
	Toll-free:	+1.800.543.2185
SPSS Argentina		+541.814.5030
SPSS Asia Pacific		+65.245.9110
SPSS Australasia		+61.2.9954.5660
	Toll-free:	+1800.024.836
SPSS Belgium		+32.162.389.82
SPSS Benelux		+31.183.636711
SPSS Brasil Ltda		+011.5505.3644
SPSS Central and Eastern Europe		+44.(0)1483.719200
SPSS Czech Republic		+420.2.24813839
SPSS Denmark		+45.45412000
SPSS East Mediterranea & Africa		+972.9.9526701
SPSS Federal Systems		+1.703.527.6777
SPSS Finland		+358.9.524.801
SPSS France		+33.1.5535.2700
SPSS Germany		+49.89.4890740
SPSS Hellas		+30.1.7251925
SPSS Hispanoportuguesa		+34.91.447.37.00
SPSS Hong Kong		+852.2.811.9662
SPSS Ireland		+353.1.496.9007
SPSS Israel		+972.9.9526700
SPSS Italia		+39.51.252573
SPSS Japan		+81.3.5466.5511
SPSS Kenya		+254.2.577.262
SPSS Korea		+82.2.3446.7651
SPSS Latin America		+1.312.651.3226
SPSS Malaysia		+60.3.704.5877
SPSS Mexico		+52.5.682.87.68
SPSS Middle East & South Asia		+91.80.545.0582
SPSS Polska		+48.12.6369680
SPSS Russia		+7.095.125.0069
SPSS Scandinavia		+46.8.506.105.50
SPSS Schweiz		+41.1.266.90.30
SPSS Singapore		+65.533.3190
SPSS South Africa		+27.11.807.3189
SPSS Taiwan		+886.2.25771100
SPSS UK		+44.1483.719200