Cornell University
Cornell Statistical Consulting Unit

StatNews #62

Building Classification and Regression Trees Using SAS Enterprise Miner
April 2004
Updated 2012

Tree methods have been available at Cornell since the introduction of software for Classification and Regression Trees (Breiman et. al., 1984). Despite their attractiveness and usefulness, tree methods have been underutilized. One possible reason is that familiar software has not been readily available. Now it is possible to implement tree methods using SAS Enterprise Miner (EM).

Tree methods are used to study the relationship between a dependent variable and a set of predictor variables. If the dependent variable is categorical, then classification trees are created. If the dependent variable is continuous, then regression trees are created. The predictor variables can be either continuous or categorical variables (see StatNews #19 and # 55 for more information).

EM is not installed automatically with SAS versions 9.2 or later, but it can be added by the user. EM offers functions such as trees, neural networks, clustering, time series, regression, and other models. To use EM, open SAS and select Solutions/Analysis/Enterprise Miner from the menu.

EM uses a point-and-click graphical user interface to data mining. The user can select a representative sample of the data, apply exploratory techniques, modify and select most important variables, model the variables to predict outcomes, and asses the model's accuracy.

EM's functions are encapsulated as nodes. To build a tree, the user selects nodes from the Tool Bar, drags them onto the Diagram Workspace, and connects them in desired sequence. The SAS implementation of trees allows binary and multi-way splits based on nominal, ordinal, and interval inputs. The user chooses the splitting criteria and other options that determine the method of tree construction. The options include the popular features of CHAID (Chi-squared automatic interaction detection), and CART (Classification and Regression Trees) algorithms.

To test out EM, a public repository data set used in similar applications, the Pima Indians Diabetes data set (768 subjects, 268 have diabetes) was used. The data set can be obtained at http://ftp.ics.uci.edu/pub/machine-learning-databases/pima-indians-diabetes. It contains eight predictor variables, and the dependent variable is binary, diabetes present or not.

As EM uses a "training and validation" approach to data mining, the data set was split into training and validation data sets (70% and 30% respectively, a common approach) using stratified random sampling. This allowed for equal proportions of diabetes cases present in both data sets. The Gini index was selected as splitting criterion, and total leaf impurity was selected as assessment criterion. The analysis was run ten different times, with each having different training and validation data sets of size 538 and 230, respectively. EM produced similar but not identical trees, with splits occurring in slightly different order. For the 10 runs, the validation misclassification error rate ranged from 0.24 to 0.30 (in range with error rates reported in the literature for the Pima Indians data set using various other algorithms to build trees).

Note that SAS Enterprise Miner is not the only commonly available software that can execute CART algorithms. Other popular statistical programs such as JMP, SPSS, Stata, and R, all have modules or packages that can do Classification and regression Trees.

If you need assistance with classification and regression trees, contact the CSCU.

Author: Simona Despa

(This newsletter was distributed by the Cornell Statistical Consulting Unit. Please forward it to any interested colleagues, students, and research staff. Anyone not receiving this newsletter who would like to be added to the mailing list for future newsletters should contact us at cscu@cornell.edu. Information about the Cornell Statistical Consulting Unit and copies of previous newsletters can be obtained at http://www.cscu.cornell.edu).

EM offers a powerful and easy way to implement trees. By using a representative sample of data instead of the entire database, it allows users to explore large quantities of data in a short amount of time.

If you need assistance with using EM for classification and regression trees, contact the CSCU.

Author: Simona Despa

(This newsletter was distributed by the Cornell Statistical Consulting Unit. Please forward it to any interested colleagues, students, and research staff. Anyone not receiving this newsletter who would like to be added to the mailing list for future newsletters should contact us at cscu@cornell.edu. Information about the Cornell Statistical Consulting Unit and copies of previous newsletters can be obtained at http://www.cscu.cornell.edu).