Distinguishing the Forest from the TREES:
A Comparison of Tree Based Data Mining Methods
Attachment to Paper

By

Richard Derrig, PhD, **OPAL Consulting LLC** **41 Fosdyke Street** **Providence** **Rhode Island, 02906, U.S.A.** **Phone: 401-861-2855** **email: richard@derrig.com**	**Louise Francis, FCAS, MAAA** **Francis Analytics and Data Mining** **706 Lombard Street** **Philadelphia** **Pennsylvania, 19147, U.S.A.** **Phone: 215-923-1567** **email: louise_francis@msn.com**

Software for modeling nonlinear dependencies and testing the models

Four software products were included in our fraud comparison: They are CART, Treenet, S-PLUS (R) and Insightful Miner[1]. The following is a discussion of some of the features of the products that was not included in the paper "Distinguishing the Forest From the Trees".

CART and Treenet are Salford Systems stand-alone software products that each performs one technique. CART (Classification and Regression Trees) does tree analysis and Treenet applies stochastic gradient boosting using the method described by Freidman (2001). The Salford System software tested produce SAS code[2] that can be used to implement the model in a production stage. Also the products contain a procedure for handling missing values using surrogate variables. At any given split point, CART and

[1] Software used in the comparison were based on 1) software licensed to the authors 2) free software and 2) software that the authors were granted temporary use of by the company licensing the software.

[2] The SAS code is generally relatively easy to edit if some other language is used to implement the model.

Treenet find the variable that is next in importance in influencing the target variable and they use this variable to replace the missing data. The specific statistic used to rank the variables and find the surrogates is described in Brieman et al. (1993). Different versions of CART and Treenet handle different size databases. The number of levels on categorical variables affects how much memory is needed, as more levels necessitate more memory. The 128k version of each product was used for this analysis. With approximately 100,000 records in the training data, occasional memory problems were experienced and it became necessary to sample fewer records. One of the very useful features of the Salford Systems software is that all the products rank variables in importance[3].

S-PLUS and R are comprehensive statistical languages used to perform a range of statistical analyses including exploratory data analysis, regression, ANOVA, generalized linear models, trees and neural networks. Both S-PLUS and R are derived from S, a statistical programming language originally developed by Bell Labs. The S progeny, S-PLUS and R, are popular among academic statisticians. S-PLUS is a commercial product sold by Insightful which has a true GUI interface that facilitates easier handling of some functions. Insightful also supplies technical support. The S-Plus programming language is widely used by analysts who do serious number crunching. They find it more effective, especially for processes that are frequently repeated. R is free open source statistical software that is supported largely by academic statisticians and computer science faculty. It has only limited GUI functionality and the data mining functions must be accessed through the language. Most code written for S-PLUS will also work for R.

[3] See Derrig and Francis (2007).

One notable difference is that data must be converted to text mode to be read by R (a bit of an inconvenience, but usually not an insurmountable one). Fox (2002) points out some of the differences between the two languages, where they exist. The S-PLUS procedures used here in the analytic comparison are found in both S-PLUS and R. However one ensemble tree method used in this research, Random Forest, appears only to be available in R but not S-PLUS. The S-PLUS (R) procedures used were: the tree function for decision trees and the glm (generalized linear models) for logistic regression and Random Forest. S-PLUS (R) incorporates relatively crude methods for handling missing values. These include eliminating all records with a missing value on any variable, an approach which is generally not recommended (Francis 2005, Allison 2002). S-PLUS also creates a new category for missing values (on categorical variables) and allows aborting the analysis if a missing value is found. In general, it is necessary to preprocess the data (at least the numeric variables where the software has no missing value method[4]) to make a provision for the missing values. In the fraud comparison, a constant not in the range of the data was substituted into the variable and an indicator dummy variable for missing was created for each numeric variable with missing values. S-PLUS and R are generally not considered optimal choices for analyzing large databases. After experiencing some difficulty reading training data of about 100,000 records into S-PLUS, the database was reduced to contain only the variables used in the analysis. Once the data was read into S-PLUS, few problems were experienced. Another eccentricity is that the S-PLUS tree

[4] S-PLUS would convert the numeric variable into a categorical variable with a level for every numeric value that is in the training data, including missing data, but the result would have far too many categories to be feasible.

function can only handle 32 levels on any given categorical variable, so in the preprocessing the number of levels may need to be reduced[5].

The Insightful Miner is a data mining suite that contains the most common data mining tools: regression, logistic regression, trees, ensemble trees, neural networks and Naïve Bayes[6]. As mentioned earlier, Insightful also markets S-PLUS. However, the Insightful Miner has been optimized for large databases and contains methods (Naïve Bayes) which are not part of S-PLUS (R). The Naïve Bayes, Tree and Ensemble Tree procedures from Insightful Miner are used here in the fraud comparison. The insightful Miner has several procedures for automatically handling missing values. These are 1) drop records with missing values, 2) randomly generate a value, 3) replace with the mean, 4) replace with a constant and 5) carry forward the last observation. Each missing value was replaced with a constant. In theory, the data mining methods used, such as trees, should be able to partition records coded for missing from the other observations with legitimate categorical or numeric values and separately estimate their impact on the target variable (possible after allowing for interactions with other variables). Server versions of the Insightful Miner generate code that can be used in deploying the model, but the version used in this analysis did not have that capability. As mentioned above some preprocessing was necessary for the Naïve Bayes procedure.

[5] Generally by collapsing sparsely populated categories into an "all other" category.
[6] It also contains some dimension reduction methods such as clustering and Principal Components which are also contained in S-PLUS.