

Boosting Multi-Objective Regression Trees

Stephan R. Sain

Patrick S. Carmack

Department of Mathematics
University of Colorado at Denver
Denver, CO 80217

Department of Statistical Science
Southern Methodist University
Dallas, TX 75275

Abstract

Boosting was introduced to improve weak classifiers but it has also been shown to be a powerful learning tool applicable to a wide variety of situations and methods, including regression trees. Using a very general mixture approach to motivate splitting rules for regression trees, a boosting algorithm is developed for regression problems with multivariate response vectors. Details of the algorithm will be discussed as well as an example based on predicting cholesterol and triglyceride levels.

1 Introduction

Boosting is a powerful learning tool that can improve the performance of classifiers and regression functions. The procedure is similar to bagging and other related methods in that many versions of a classifier or regression function are fit to modified versions of the data and then combined to produce much stronger results. Friedman (2001) and Hastie, Tibshirani, and Friedman (2001) discuss boosting algorithms for regression trees. Regression trees are powerful nonparametric alternatives to traditional regression methodologies that capitalize on the complex relationships and higher-order interactions in data.

Sain and Carmack (2001) discuss regression trees for multivariate or multi-objective regression problems. Consider the multivariate multiple regression model given by

$$\mathbf{Y} = f(\mathbf{X}) + \mathbf{E}$$

where \mathbf{Y} is a $n \times p$ matrix, \mathbf{X} is an $n \times k$ matrix of predictor variables, \mathbf{E} is a $n \times p$ matrix of errors, and f is the unknown regression function. The traditional approach is to assume f is a linear function of the predictors, i.e. $f(\mathbf{X}) = \mathbf{X}\mathbf{B}$ where \mathbf{B} is a $k \times p$ matrix of regression parameters. However, regression trees offer an alternative when the linear assumption is not well-founded. In this work, boosting is used to improve regression tree estimates of f based on the mixture motivated splitting rules for multi-objective regression trees developed in Sain and Carmack (2001).

2 Regression Trees

Trees have been used extensively for classification and regression problems. Initially developed in the social sciences (Morgan and Sonquist, 1963), trees have been studied from different points-of-view, including machine learning (Quinlan, 1986 and 1993) and more statistical approaches. Breiman, et al. (1984), formalized many of the ideas and made a number of important contributions for tree construction in common use today.

Tree models can be written as

$$f(\mathbf{x}) = \sum_{i=1}^m \hat{\mu}_i I(\mathbf{x} \in R_i) \quad (1)$$

where the collection of R_i are disjoint partitions of the predictor variables and the $\hat{\mu}_i$ are the means of the response variables associated with the predictor variables in R_i . These models are built by first growing, which is a forward stepwise process in which additional terms or partitions are added to (1), and then pruning, which is analogous to backward model selection in which unimportant terms are removed from (1).

Trees are grown in a greedy fashion following the notion of recursive partitioning. This involves a series of yes/no questions that split the data or some subset of the data (referred to as the parent) into two groups (referred to as the children). These questions are based on the predictor variables and are of the form:

- “Is $x \leq c$?” where $a \leq c \leq b$ and $a \leq x \leq b$ if x is nominal, or
- “Is $x \in c$?” where $c \subset S$ and S is the set of possible values of x if x is categorical.

The process begins with the whole data set and each of the variables is examined in turn. The best split or partition of the data is chosen, i.e. an explanatory variable, x , and threshold, c , are chosen that make the children of the split more homogeneous (in terms of the response variables) than the parent. In other words, a constant model is initially proposed, i.e. $f(\mathbf{x}) = \hat{\mu}_1 I(\mathbf{x} \in R_1)$ where R_1 is the entire space of the predictor variables. The data are then split on a particular choice of predictor variable as above yielding

$$f(\mathbf{x}) = \hat{\mu}_2 I(\mathbf{x} \in R_2) + \hat{\mu}_3 I(\mathbf{x} \in R_3). \quad (2)$$

Subsequently, each of the children is split which involves replacing one of the terms in (2) with $\hat{\mu}_{2i} I(\mathbf{x} \in R_{2i}) + \hat{\mu}_{2i+1} I(\mathbf{x} \in R_{2i+1})$ for $i = 2$ or 3 . This process continues until some simple stopping criterion is met (usually some minimum number of observations are required in both the parent and children order to consider splitting).

To ensure that all of the important structure is included in the tree, the growing process intentionally overfits to the training data. Breiman, et al. (1984) then suggest pruning the tree by removing the least important splits based on some cost-complexity measure. A popular alternative is cross-validation. The original tree can be considered a nested model and it is fairly straightforward to extract the best tree of each size (number of splits or terminal nodes). The training data are randomly split into k disjoint subsets (usually $k = 10$). A tree is fit using $k - 1$ subsets. Trees of each size are extracted from this tree and predictions are computed using the k th subset. The prediction error is stored for each size tree and the process is repeated until each subset has been held out for prediction. An estimate of the

optimal size of the tree can be found by choosing the size tree that minimizes the cross-validation classification or average prediction error, depending on the type of tree being used.

3 Mixture-Based Splitting Rules

Sain and Carmack (2001) suggest a method for generalizing trees for multivariate response variables by modeling the distribution of the response variables as a mixture. The form of (1) suggests that the response data at each terminal node can be assumed to come from a different component of a mixture distribution (McLachlan and Peel, 2000) of the form

$$g(\mathbf{y}) = \sum_{i=1}^m p_i g_i(\mathbf{y}; \theta_i)$$

where $0 < p_i < 1$ for all i and $\sum_i p_i = 1$ and g_i is some probability density function (taken here to be multivariate normal) with parameter vector θ_i .

Splitting rules can then easily be derived by examining the likelihood. A close examination of the likelihood shows that splitting rules are entirely local and are equivalent to finding the partition of the data that maximizes

$$L(\theta_l, \theta_r, \gamma) = \prod_{i:\gamma_i=l} f_l(\mathbf{y}_i, \theta_l) \prod_{i:\gamma_i=r} f_r(\mathbf{y}_i; \theta_r),$$

where f_r and f_l represent the probability density functions associated with the left and right children and γ represents the partition (a vector whose elements indicate which data points go to the left and which go to the right). Assuming that f_r and f_l are multivariate normal with different covariance matrices and applying some standard results, finding the split or partition that maximizes the likelihood or log-likelihood is equivalent to finding the partition that minimizes

$$n_l \log |\hat{\Sigma}_l| + n_r \log |\hat{\Sigma}_r|,$$

where $\hat{\Sigma}_l$ and $\hat{\Sigma}_r$ are the maximum likelihood estimates of the covariance matrices for the left and right children of the split. It should be noted that likelihood based splitting rules have been considered previously in the case of a univariate response, although from a different perspective. See, for example, Clark and Pregibon (1992). This approach is also similar to the mixture based clustering ideas of Banfield and Raftery (1993).

Different splitting rules can be derived by placing restrictions on the covariance structure. Keeping the covariance matrices spherical, but allowing different levels of variability, i.e. $\Sigma_i = \sigma_i \mathbf{I}$, yields a splitting rule of the form

$$n_l \log \text{tr} \hat{\Sigma}_l + n_r \log \text{tr} \hat{\Sigma}_r,$$

where tr indicates the trace of a matrix. Perhaps the most restrictive form forces each component to have equivalent spherical covariance structure, i.e. $\Sigma_i = \sigma \mathbf{I}$. This yields

$$\text{tr} W_l + \text{tr} W_r$$

where W_l and W_r are the matrices of sum-of-squares and cross-products. Interestingly, this criterion is exactly the least-squares criterion that is commonly used to fit regression trees, suggesting that it might be inappropriate in the case of heteroscedasticity.

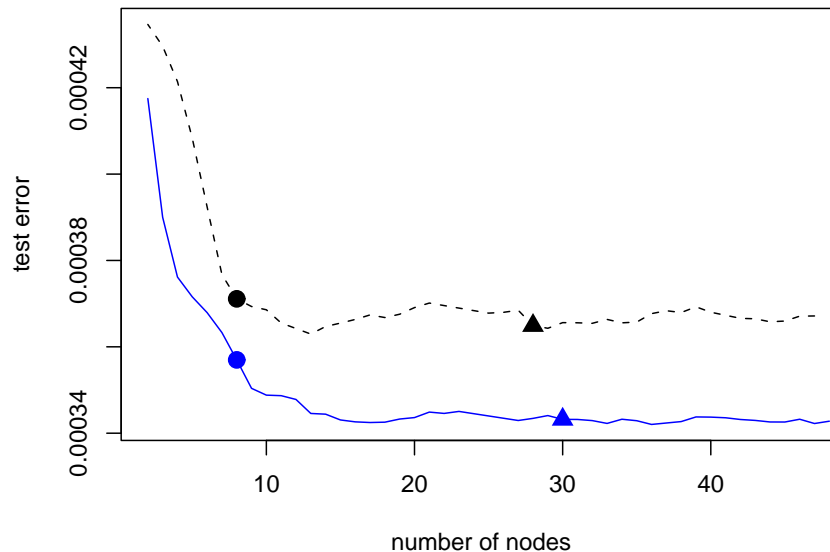


Figure 1: Testing error for trees fit with least-squares (dashed line) and the general covariance structure mixture criterion (solid line). Circles indicate cross-validation choices for the number of terminal nodes. Triangles indicate AIC choices for the number of terminal nodes.

3.1 The Turkish Heart Study

To demonstrate the methodology, we consider a study of environmental factors such as diet, work activity, etc. on cholesterol and triglyceride levels in the male Turkish population which is reported to have a high rate of coronary artery disease (Mahley, et al., 1995). The data consist of $n = 5310$ male volunteers (routine corporate physical exams, Lions/Rotary Clubs, etc.) and the data is randomly split into $n = 3200$ training and $n = 2110$ testing sets.

Diet is studied by proxy as suggested by the region of Turkey. In the study, there are six regions including Istanbul (western diet), Adana (lamb), Trabzon (diary), Kayseri (meat, diary), Aydin (sunflower and corn oil), and Ayvalik (olive oil). Other factors include age, blood pressure, weight and body mass, education, salary, type of work, and smoking. Salary is split into four categories ranging from \$200 - \$1000/month. Education is split into six categories including none; primary, middle, and high school; university; graduate school. Work has four categories including physical labor, office/shop, administrative/management, none.

Figure 1 shows the test error of the sequence of trees suggested by using two criterion, least-squares and the general covariance structure mixture criterion. Error was measured by taking the determinant of the covariance matrix of the difference between the testing response vectors and the predicted response vectors. Clearly, trees fit using the general covariance structure seem to outperform the trees fit using least-squares. The circles on the plot represent the optimal cross-validation trees while the triangles indicate trees with optimal tree size estimated by the AIC criterion. AIC is a penalized likelihood criterion of the form $aic = -2\log\text{-likelihood} + \#$ of parameters. The number of parameters is a bit difficult to compute since the splits are determined from the data. Our approach is similar to that used in the

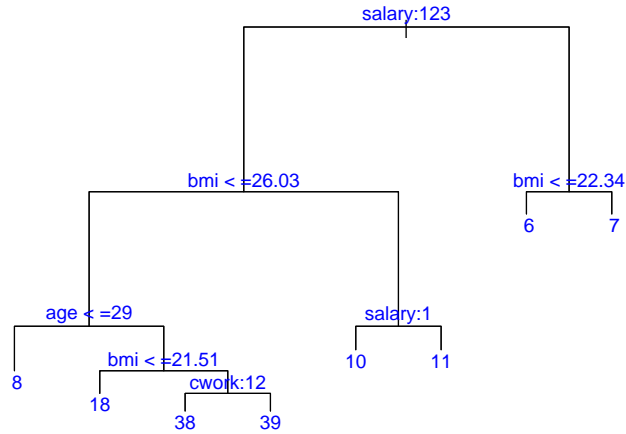


Figure 2: Final cross-validation tree using the general mixture criterion.

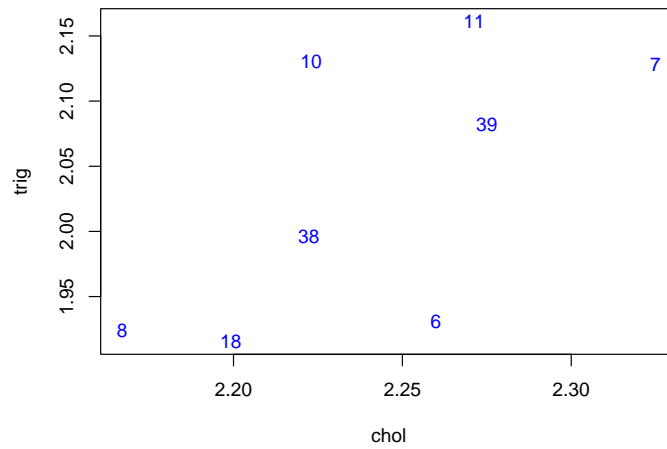


Figure 3: Scatterplot of the means of cholesterol and triglyceride. Plot characters are node numbers.

multivariate adaptive regression spline methodology of Friedman (1991) in that an additional penalty is added for the splits. In other words, the number of parameters in the general covariance structure case is $(\# \text{ of terminal nodes}) \times (p + \frac{1}{2}p(p+1) + c)$ where p is the dimensionality of the response and $c = 3$ taken as the penalty for each data-based split.

Figures 2 and 3 show the final cross-validation tree. Traditionally such summary measures as the node mean are shown for each terminal node in the tree. However, this can begin to appear rather cluttered with multi-objective trees. Our approach is to link the tree with a scatterplot matrix. Here the node number is plotted at each terminal node as well as on the scatterplot in Figure 3. Interpreting the plots is straightforward. For example, the initial split puts the highest level of salary to the right and then splits on body mass (nodes 6 and 7). These people tend to be executives in Istanbul and, from the scatterplot, have higher levels of cholesterol. For those with higher body mass (node 7), both cholesterol and triglycerides are at higher levels.

4 Boosting

Boosting was introduced as a way to improve weak classifiers. The idea is to successively fit the classifier to the training data, each time giving more weight to misclassified training points. Friedman (2001) generalized this approach to fit models of the form

$$f(\mathbf{x}; \alpha, \beta) = \sum_{i=1}^M \beta_i b(\mathbf{x}, \alpha_i)$$

where β are expansion coefficients and b are simple functions (wavelet/spline basis functions, trees, etc.) with parameter vector α . See also Hastie, et al. (2001). These models are fit using the following a greedy approximate algorithm

1. Initialize $f_0(\mathbf{x})$.
2. For $i = 1, \dots, M$
 - (a) Compute

$$(\alpha_i, \beta_i) = \arg \min_{\alpha, \beta} L(\mathbf{Y}, \mathbf{f}_{i-1} + \beta \mathbf{b}).$$

where

$$\mathbf{b} = (b(\mathbf{x}_1; \alpha), \dots, b(\mathbf{x}_N; \alpha))'$$

- (b) Set $f_i(\mathbf{x}) = f_{i-1}(\mathbf{x}) + \beta_i b(\mathbf{x}, \alpha_i)$.

In this setting, this algorithm fits a (small) tree and computes the residuals. Another (small) tree is fit to the residuals and the whole process is repeated a large number of times. Predictions are then built up sequentially. In a sense, this procedure is analogous to numerical optimization where one starts with initial values and takes steps towards the optimal value, usually in the direction indicated by the estimated gradient.

4.1 The Turkish Heart Study

Continuing the example presented in the previous section, Figure 4 shows the test error for the boosted tree using the general covariance structure criterion as well as the best cross-validation and AIC tree. The boosted tree model uses individual trees with three terminal nodes and iteratively refits the tree using the algorithm presented previously.

It is also possible to use the boosted tree to obtain information about the relative importance of the underlying variables as well as a graphical representation of the functional relationship between the predictor variables and the response variables. Breiman, et al. (1984) suggested a measure of importance based on the improvement associated with each split. In this setting, that amounts to recording the reduction in the likelihood based criterion associated with each split and associating this reduction with the predictor variable used in the split. These reductions are then summed over all nodes in a particular tree. Hastie, et al. (2001) note that, for boosted trees, this measure can just be summed over all trees used in the final fit. These importances are often scaled to range from 0 to 100.

A plot of the variable importance is shown in Figure 5. This type of plot can aid in the interpretation of the tree structure. Often variables are masked, meaning

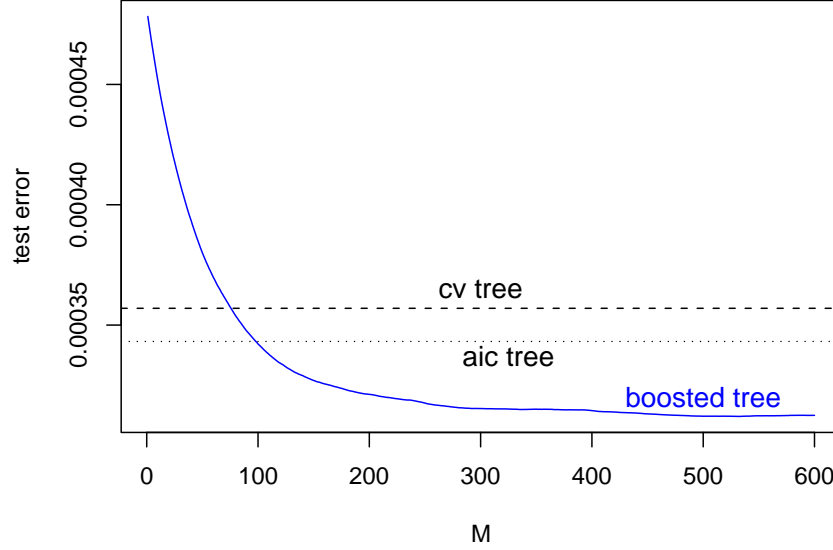


Figure 4: Test error for boosted tree (solid line) as well as the cross-validation and AIC tree.

that they are strongly associated, but are often not chosen for a split since some other variable is slightly better. In Figure 5, variables such as weight and body mass, region (diet), and age seem important (as expected).

Taking a closer look at such variables as age, one can examine plots of partial dependence functions (Hastie, et al., 2001). These are defined as

$$f_S(X_S) = \frac{1}{n} \sum_{i=1}^n f(X_S, x_{iC})$$

where x_{iC} are the values of X_C occurring in the training data. For univariate functions, this is fairly straightforward to implement and construct a plot. For multivariate functions, a little more care must be taken. For example, consider examining the effect of age on cholesterol and triglyceride shown in Figure 6. Bins with equal numbers of data points are constructed from the age variable and the average age, cholesterol, and triglyceride are computed. These are plotted in the figure with the plot character denoted the average age. As one might expect, as age increases so do the levels of cholesterol and triglycerides. For the bins representing the older ages, levels of cholesterol and triglyceride seem to level out or actually decrease representing a common effect in this type of non-longitudinal data.

A second partial variable plot is shown in Figure 7 using region (diet) as a predictor variable. In this case, the average cholesterol and triglyceride values are shown as a function of the regions. Istanbul clearly shows much larger cholesterol values. It is difficult to see a direct link with diet since there are a number of other confounding factors, such as salary and work.

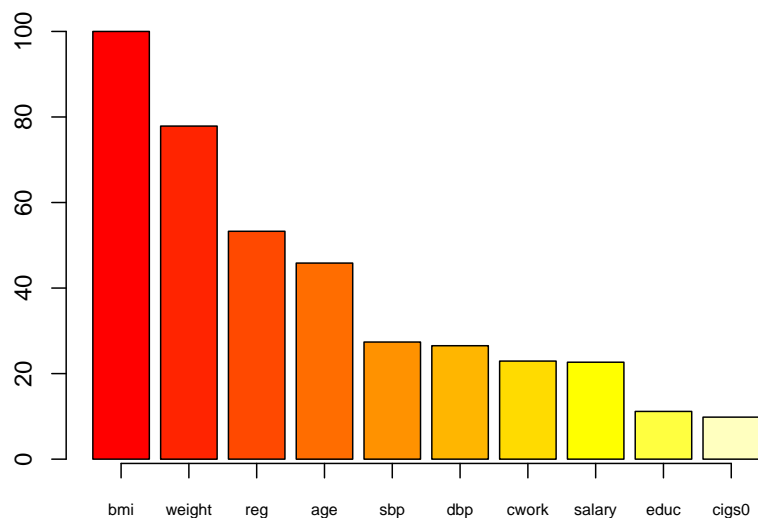


Figure 5: Variable ranking based on the boosted tree.

5 Concluding Remarks

The mixture based splitting rules have been shown to be effective for multi-objective regression trees. The boosting approach can offer additional improvement. The size of trees used in the boosting algorithm can have a dramatic effect, although for most cases trees with fewer than five terminal nodes seem to work well. Often, stumps (two terminal nodes) are adequate. The number of trees to add to the final model seems to be less important as long as enough (usually several hundred) are used. A better evaluation of these issues as well as a comparison with other “committee” methods (e.g. bagging) is currently under way.

References

- Banfield, J.D. and Raftery, A.E. (1993), “Model-Based Gaussian and Non-Gaussian Clustering,” *Biometrics*, Vol. 49, pp. 803-821.
- Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984), *Classification and Regression Trees*, Chapman and Hall.
- Clark, L.A. and Pregibon, D. (1992), “Tree-Based Models,” In *Statistical Models in S*, J.M. Chambers and T.J. Hastie (eds.), Wadsworth and Brooks/Cole, Pacific Grove, Ca.
- Friedman, J. (1991), “Multivariate Adaptive Regression Splines (with discussion),” *Annals of Statistics*, Vol. 19, pp. 1-141.
- Friedman, J. (2001), “Greedy Function Approximation: the Gradient Boosting Machine,” *Annals of Statistics*, To appear.
- Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning*, New York: Springer.

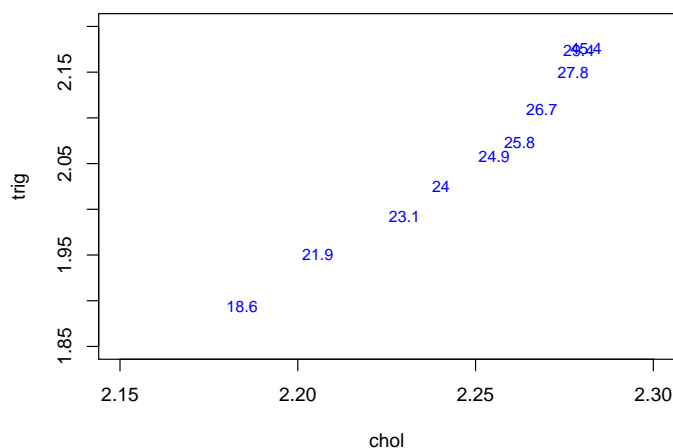


Figure 6: Partial variable plot of cholesterol and triglyceride by age.

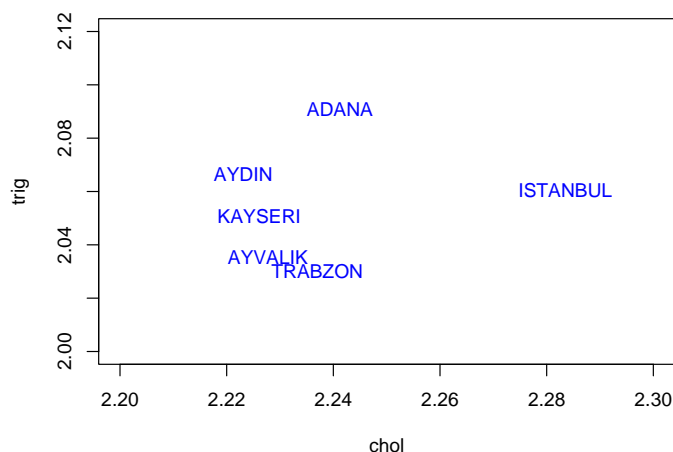


Figure 7: Partial variable plot of cholesterol and triglyceride by region.

Mahley, R.W., et al. (1995), "Turkish Heart Study: Lipids, Lipoproteins, and Apolipoproteins," *Journal of Lipid Research*, **36**, 839-859.

McLachlan, G. and Peel, D. (2000), *Finite Mixture Models*, New York: John Wiley.

Morgan, J.N. and Sonquist, J.A. (1963), "Problems in the analysis of survey data, and a proposal," *Journal of the American Statistical Association*, Vol. 58, pp. 415-434.

Quinlan, J.R. (1986), "Induction of decision trees," *Machine Learning*, Vol. 1, pp. 81-106.

Quinlan, J.R. (1993), *C4.5: Programs for Machine Learning*, San Mateo, CA: Morgan Kaufman.

Sain, S.R. and Carmack, P.S. (2001), "A Mixture Approach for Multivariate Regression Trees," *Proceedings of the American Statistical Association*, To appear.

Therneau, T.M. and Atkinson, E.J. (1997), "An Introduction to Recursive Partitioning Using the RPART Routines," Technical Report, The Mayo Foundation.