International Statistical Review

International Statistical Review (2014), 82, 3, 329–348 doi:10.1111/insr.12016

Fifty Years of Classification and Regression Trees[1]

Wei-Yin Loh

Department of Statistics, University of Wisconsin, Madison, WI 53706, USA
E-mail: loh@stat.wisc.edu

Summary

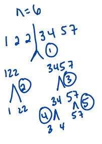
Fifty years have passed since the publication of the first regression tree algorithm. New techniques have added capabilities that far surpass those of the early methods. Modern classification trees can partition the data with linear splits on subsets of variables and fit nearest neighbor, kernel density, and other models in the partitions. Regression trees can fit almost every kind of traditional statistical model, including least-squares, quantile, logistic, Poisson, and proportional hazards models, as well as models for longitudinal and multiresponse data. Greater availability and affordability of software (much of which is free) have played a significant role in helping the techniques gain acceptance and popularity in the broader scientific community. This article surveys the developments and briefly reviews the key ideas behind some of the major algorithms.

Key words: Classification trees; regression trees; machine learning; prediction.

1 Introduction

As we reach the 50th anniversary of the publication of the first regression tree algorithm (Morgan & Sonquist, 1963), it seems appropriate to survey the numerous developments in the field. There have been previous reviews, but some are dated (e.g., Murthy, 1998) and others were written as brief overviews (e.g., Loh, 2008a; 2011; Merkle & Shaffer, 2009; Strobl *et al.*, 2011) or simple introductions intended for non-statistics audiences (e.g., De'ath & Fabricius, 2000; Harper, 2003; Lemon *et al.*, 2005). Owing to the large and increasing amount of literature (in statistics, computer science, and other fields), it is impossible, of course, for any survey to be exhaustive. We have therefore chosen to focus more attention on the major algorithms that have stood the test of time and for which software is widely available. Although we aim to provide a balanced discussion, some of the comments inevitably reflect the opinions of the author.

AID (ordered)



We say that X is an *ordered* variable if it takes numerical values that have an intrinsic ordering. Otherwise, we call it a *categorical* variable. Automatic Interaction Detection (AID) (Morgan & Sonquist, 1963) is the first regression tree algorithm published in the literature. Starting at the root node, AID recursively splits the data in each node into two children nodes. A split on an ordered variable X takes the form "$X \leq c$". If X has n distinct observed values, there are $(n-1)$ such splits on X. On the other hand, if X is a categorical variable having m distinct observed values, there are $(2^{m-1}-1)$ splits of the form "$X \in A$", where A is a subset of the X values. At any node t, let $S(t)$ denote the set of training data in t and let

?

[1] This paper is followed by discussions and a rejoinder.

variability within the node

$\bar{y}_t$ be the sample mean of Y in t. Let $\phi(t)$ denote the node "impurity" of t. Using the sum of squared deviations $\phi(t) = \sum_{i \in S(t)} (y_i - \bar{y}_t)^2$, AID chooses the split that minimizes the sum of the impurities in the two children nodes. Splitting stops when the reduction in impurity is less than a preset fraction of the impurity at the root node. The predicted Y value in each terminal node is the node sample mean. The result is a piecewise constant estimate of the regression function.

how is this chosen?

THeta Automatic Interaction Detection (THAID) (Messenger & Mandell, 1972) extends these ideas to classification, in which Y is a categorical variable. THAID chooses splits to maximize the sum of the number of observations in each modal category (i.e., the category with the most observations). Alternative impurity functions are the entropy, $\phi(t) = -\sum_j p(j|t) \log p(j|t)$, and the Gini index, $\phi(t) = 1 - \sum_j p^2(j|t)$, where $p(j|t)$ is the proportion of class j observations in node t. Messenger & Mandell (1972) attributed the Gini index to Light & Margolin (1971).

Figure 1 shows a classification tree model for the iris data that Fisher (1936) used to introduce linear discriminant analysis (LDA). Four measurements (petal length and width, and sepal length and width) were recorded on 150 iris flowers, with 50 from each of the Setosa, Versicolour, and Virginica types. The tree splits only on petal length and width.

Despite their novelty, or perhaps owing to it, AID and THAID did not initially attract much interest in the statistics community. Einhorn (1972) showed by simulation that AID can severely overfit the data. Doyle (1973) pointed out that if two or more variables are highly correlated, at most one may appear in the tree structure. This problem of *masking* can lead to spurious conclusions about the relative importance of the variables. Bishop *et al.* (1975) criticized AID for ignoring the inherent sampling variability of the data. Around the same time though, the idea of recursive partitioning was gaining steam in the computer science and engineering communities as more efficient algorithms for carrying out the search for splits began to appear (Chou 1969; Henrichon & Fu, 1973; Meisel & Michalopoulos, 1977; Payne & Meisel, 1977; Sethi & Chatterjee, 1991).

2 Classification Trees

require a categorical outcome, but any type of "feature"

We begin with classification trees because many of the key ideas originate here.

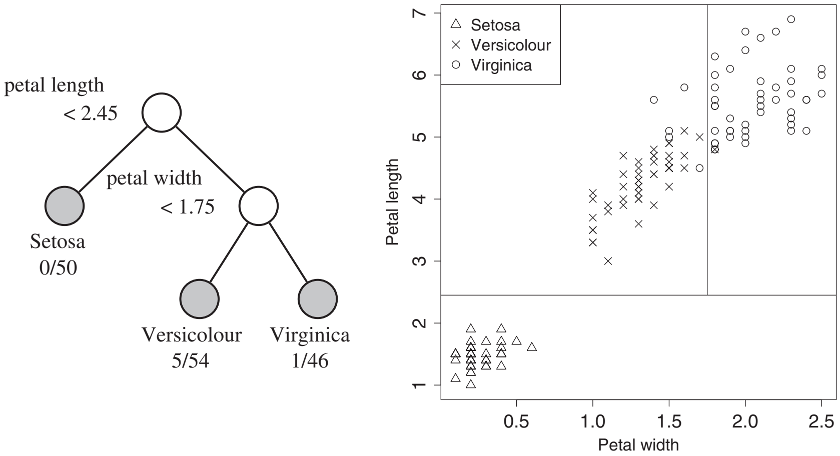


Figure 1. *Classification tree model for iris data. At each intermediate node, an observation goes to the left child node if and only if the stated condition is true. The pair of numbers beneath each terminal node gives the number misclassified and the node sample size.*

2.1 CART

Classification And Regression Trees (CART) (Breiman *et al.*, 1984) was instrumental in regenerating interest in the subject. It follows the same greedy search approach as AID and THAID, but adds several novel improvements. Instead of using stopping rules, it grows a large tree and then prunes the tree to a size that has the lowest cross-validation estimate of error. The pruning procedure itself is ingenious, being based on the idea of weakest-link cutting, with the links indexed by the values of a cost-complexity parameter. This solves the under-fitting and over-fitting problems of AID and THAID, although with increased computation cost. To deal with missing data values at a node, CART uses a series of "surrogate" splits, which are splits on alternate variables that substitute for the preferred split when the latter is inapplicable because of missing values. Surrogate splits are also used to provide an importance score for each X variable. These scores, which measure how well the surrogate splits predict the preferred splits, can help to detect masking. CART can also employ linear splits, that is, splits on linear combinations of variables, by stochastic search. Brown *et al.* (1996) proposed a linear programming solution as an alternative. Breiman *et al.* (1984) obtained conditions for all recursive partitioning techniques to be Bayes risk consistent. CART is available in commercial software. It is implemented as RPART (Therneau & Atkinson, 2011) in the R system (R Core Team 2014).

? what ?

2.2 CHAID

CHi-squared Automatic Interaction Detector (CHAID) (Kass, 1980) employs an approach similar to stepwise regression for split selection. It was originally designed for classification and later extended to regression. To search for an X variable to split a node, the latter is initially split into two or more children nodes, with their number depending on the type of variable. CHAID recognizes three variable types: categorical, ordered without missing values (called *monotonic*), and ordered with missing values (called *floating*). A separate category is defined for missing values in a categorical variable. If X is categorical, a node t is split into one child node for each category of X. If X is monotonic, t is split into 10 children nodes, with each child node defined by an interval of X values. If X is floating, t is split into 10 children nodes plus one for missing values. Pairs of children nodes are then considered for merging by using Bonferroni-adjusted significance tests. The merged children nodes are then considered for division, again by means of Bonferroni-adjusted tests. Each X variable is assessed with a Bonferroni-adjusted p-value, and the one with the smallest p-value is selected to split the node. CHAID is currently available in commercial software only.

2.3 C4.5

C4.5 (Quinlan, 1993) is an extension of the ID3 (Quinlan, 1986) classification algorithm. If X has m distinct values in a node, C4.5 splits the latter into m children nodes, with one child node for each value. If X is ordered, the node is split into two children nodes in the usual form "$X < c$". C4.5 employs an entropy-based measure of node impurity called *gain ratio*. Suppose node t is split into children nodes $t_1, t_2, \ldots, t_r$. Let $n(t)$ denote the number of training samples in t, and define $\phi(t) = -\sum_j p(j|t) \log p(j|t)$, $f_k(t) = n(t_k)/n(t)$, $\phi_X(t) = \sum_{k=1}^{r} \phi(t_k) f_k(t)$, $g(X) = \phi(t) - \phi_X(t)$, and $h(X) = -\sum_k f_k(t) \log f_k(t)$. The gain ratio of X is $g(X)/h(X)$. Although C4.5 takes almost no time on categorical variable splits, the strategy has the drawback that if X has many categorical values, a split on X may produce children nodes with so few observations in each that no further splitting is possible—see Loh (2008a) for an example. C4.5 trees are pruned with a heuristic formula instead of cross-validation.

If there are missing values, the gain function is changed to $g(X) = F\{\phi(t) - \phi_X(t)\}$, where F is the fraction of observations in a node non-missing in X. The $h(X)$ function is extended by the addition of a "missing value" node t_{r+1} in its formula. If an observation is missing the value of a split variable, it is sent to every child node with weights proportional to the numbers of non-missing observations in those nodes. Empirical evidence shows that C4.5 possesses excellent speed and good prediction accuracy, but its trees are often substantially larger than those of other methods (Lim *et al.*, 2000; Loh, 2009).

Source code for C4.5 can be obtained from www.rulequest.com/Personal/c4.5r8.tar.gz. It is also implemented as J48 in the WEKA (Hall *et al.*, 2009) suite of programs.

2.4 FACT and QUEST

Fast and Accurate Classification Tree (FACT) (Loh & Vanichsetakul, 1988) is motivated by recursive LDA, which generates linear splits. As a result, it splits each node into as many children nodes as the number of classes. To obtain univariate splits, FACT uses analysis of variance (ANOVA) F-tests to rank the X variables and then applies LDA to the most significant variable to split the node. Categorical X variables are transformed first to dummy 0–1 vectors and then converted to ordered variables by projecting the dummies onto the largest discriminant coordinate. Splits on the latter are expressed back in the form $X \in A$. Missing X values are estimated at each node by the sample means and modes of the non-missing ordered and categorical variables, respectively, in the node. Stopping rules based on the ANOVA tests are used to determine the tree size.

One weakness of the greedy search approach of AID, CART, and C4.5 is that it induces biases in variable selection. Recall that an ordered X variable taking n distinct values generates $(n-1)$ splits. Suppose X_1 and X_2 are two such variables with n_1 and n_2 distinct values, respectively. If $n_1 < n_2$ and both variables are *independent* of Y, then X_2 has a larger chance to be selected than X_1. The situation is worse if X_2 is a categorical variable, because the number of splits grows exponentially with n_2. Breiman *et al.* (1984, p. 42) noted this weakness in the CART algorithm, and White & Liu (1994) and Kononenko (1995) demonstrated its severity in C4.5. We will say that an algorithm is *unbiased* if it does not have such biases. Specifically, if all X variables are independent of Y, an unbiased algorithm gives each X the same chance of being selected to split a node.

FACT is unbiased if all the X variables are ordered, because it uses F-tests for variable selection. But it is biased toward categorical variables, because it employs LDA to convert them to ordered variables before application of the F-tests. Quick, Unbiased and Efficient Statistical Tree (QUEST) (Loh & Shih, 1997) removes the bias by using F-tests on ordered variables and contingency table chi-squared tests on categorical variables. To produce binary splits when the number of classes is greater than 2, QUEST merges the classes into two superclasses in each node before carrying out the significance tests. If the selected X variable is ordered, the split point is obtained by either exhaustive search or quadratic discriminant analysis. Otherwise, if the variable is categorical, its values are transformed first to the largest linear discriminant coordinate. Thus, QUEST has a substantial computational advantage over CART when there are categorical variables with many values. Linear combination splits are obtained by applying LDA to the two superclasses. The trees are pruned as in CART.

2.5 CRUISE

Whereas CART always yields binary trees, CHAID and C4.5 can split a node into more than two children nodes, their number depending on the characteristics of the X variable. Classification Rule with Unbiased Interaction Selection and Estimation (CRUISE) (Kim & Loh, 2001) is a descendent of QUEST. It splits each node into multiple children nodes, with their number depending on the number of distinct Y values. Unlike QUEST, CRUISE uses contingency table chi-squared tests for variable selection throughout, with the values of Y forming the rows and the (grouped, if X is ordered) values of X forming the columns of each table. We call these "main effect" tests, to distinguish them from "pairwise interaction" tests that CRUISE also performs, which are chi-squared tests cross-tabulating Y against Cartesian products of the (grouped) values of pairs of X variables. If an interaction test between X_i and X_j, say, is most significant, CRUISE selects X_i if its main effect is more significant than that of X_j, and vice versa. Split points are found by LDA, after a Box–Cox transformation on the selected X variable. Categorical X variables are first converted to dummy vectors and then to their largest discriminant coordinate, following FACT and QUEST. CRUISE also allows linear splits using all the variables, and it can fit a linear discriminant model in each terminal node (Kim & Loh, 2003).

Kim & Loh (2001) showed that CART is biased toward selecting split variables with *more* missing values and biased toward selecting surrogate variables with *fewer* missing values. The cause is due to the Gini index being a function of the class proportions and not the class sample sizes. CRUISE and QUEST are unbiased in this respect. CRUISE has several missing value imputation methods, the default being imputation by predicted class mean or mode, with class prediction based on a non-missing X variable.

2.6 GUIDE

Generalized, Unbiased, Interaction Detection and Estimation (GUIDE) (Loh, 2009) improves upon QUEST and CRUISE by adopting their strengths and correcting their weaknesses. One weakness of CRUISE is that there are many more interaction tests than main effect tests. As a result, CRUISE has a greater tendency to split on variables identified through interaction tests. GUIDE restricts their frequency by using the tests only if no main effect test is significant at a Bonferroni-corrected level. This reduces the amount of computation as well. Further, GUIDE uses a two-level search for splits when it detects an interaction between X_i and X_j, say, at a node t. First, it finds the split of t on X_i and the splits of its two children nodes on X_j that yield the most reduction in impurity. Then it finds the corresponding splits with the roles of X_i and X_j reversed. The one yielding the greater reduction in impurity is used to split t.

Besides univariate splits, GUIDE can employ bivariate linear splits of two X variables at a time. The bivariate linear splits can be given higher or lower priority over univariate splits. In the latter case, linear splits are considered only if no interaction tests are significant after another Bonferroni correction. Although bivariate linear splits may be less powerful than linear splits on all X variables together, the former are still applicable if the number of X variables exceeds the number of observations in the node.

Other improvements in GUIDE include (i) assigning missing categorical values to a "missing" category, (ii) fitting bivariate kernel or nearest-neighbor node models, and (iii) using the node chi-squared test statistics to form an importance score for each variable (Loh, 2012). Smyth *et al*. (1995) and Buttrey & Karo (2002) proposed fitting kernel density estimation and nearest-neighbor models, respectively, in the terminal nodes of a CART tree or a C4.5 tree. Executable codes for CRUISE, GUIDE, and QUEST are distributed free from http://www.stat.wisc.edu/~loh/.

2.7 CTREE and Other Unbiased Approaches

Conditional Inference Trees (CTREE) (Hothorn *et al*., 2006b) is another algorithm with unbiased variable selection. It uses p-values from permutation distributions of influence function-based statistics to select split variables. Monte Carlo or asymptotic approximations to the p-values are employed if they cannot be computed exactly. CTREE does not use pruning; it uses stopping rules based on Bonferroni-adjusted p-values to determine tree size. The algorithm is implemented in the R package PARTY.

Shih (2004), Shih & Tsai (2004), and Strobl *et al.* (2007a) proposed to correct the selection bias of CART by choosing splits based on p-values of the maximal Gini statistics. The solutions are limited, however, to ordered X variables and to classification and piecewise constant regression trees, and they increase computation cost.

2.8 Ensemble, Bayesian, and Other Methods

There is much interest recently on the use of ensembles of classifiers for predictions. In this approach, the predicted value of an observation is based on the majority "vote" from the predicted values of the classifiers in the ensemble. *Bagging* (Breiman, 1996) uses an ensemble of unpruned CART trees constructed from bootstrap samples of the data. *Random forest* (Breiman, 2001) weakens the dependence among the CART trees by using a random subset of X variables for split selection at each node of a tree. Hothorn & Lausen (2005) applied bagging to the original variables as well as the predicted values of other classifiers, such as LDA, nearest neighbor, and logistic regression.

Boosting (Freund & Schapire, 1997) sequentially constructs the classifiers in the ensemble by putting more weight on the observations misclassified in the previous step. Hamza & Larocque (2005) found random forest to be better than boosting CART, but Gashler *et al.* (2008) showed that random forest can perform poorly if there are irrelevant variables in the data. Dietterich (2000) reviewed ensemble methods in the computer science literature.

Another class of ensemble methods is Bayesian model averaging, where prior distributions are placed on the set of tree models and stochastic search is used to find the good ones. Chipman *et al.* (1998) used a prior distribution that can depend on tree size and shape, and Denison *et al.* (1998) used a truncated Poisson prior that puts equal weight on equal-sized trees. For split point selection on an ordered X variable, Chipman *et al.* (1998) used a discrete uniform prior on the observed values of X, and Denison *et al.* (1998) used a continuous uniform distribution on the range of X.

2.9 Importance Scores

Many tree algorithms produce importance scores of the X variables. CART bases the scores on the surrogate splits, but because the latter are subject to selection bias, the scores are similarly biased. Sandri & Zuccolotto (2008) proposed a method to correct the bias. GUIDE uses as importance score a sum of weighted chi-squared statistics over the intermediate nodes, with node sample sizes as weights. A chi-squared approximation to the null distribution of the importance scores is used to provide a threshold for identifying the noise variables. Random forest derives its importance scores from changes in prediction error after random permutation of the X variable values. Strobl *et al.* (2007b) showed that the scores are biased toward correlated variables, and Strobl *et al.* (2008) proposed an alternative permutation scheme as a solution.

2.10 Comparisons

Figures 2 and 3 show the tree models and their partitions given by C4.5, CHAID, CRUISE, CTREE, GUIDE, QUEST, and Recursive PARTitioning (RPART) for the iris data. The X

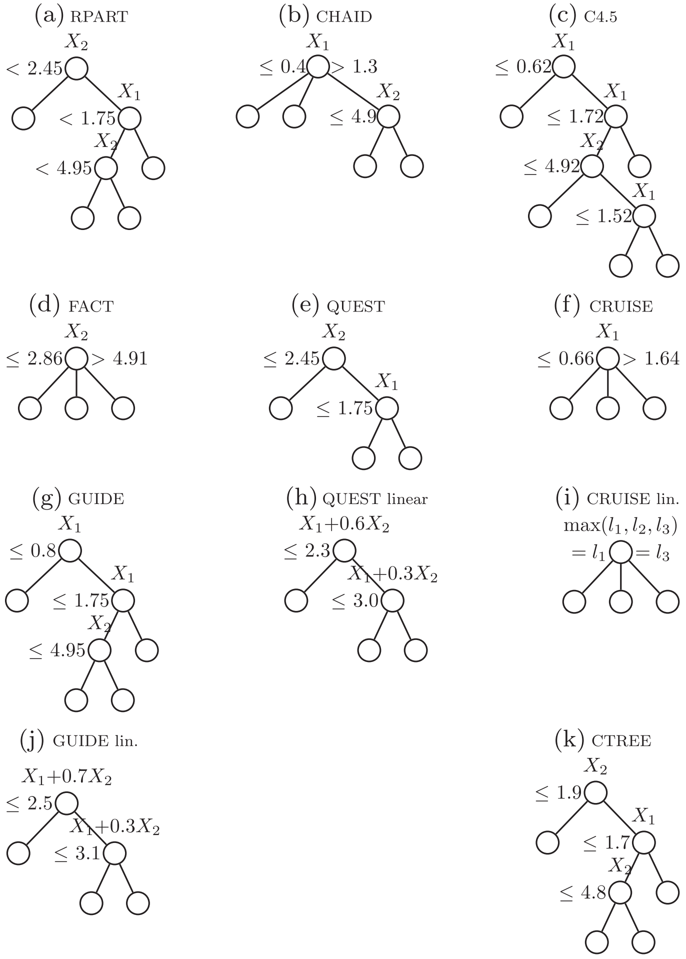


Figure 2. *Classification trees for iris data. X_1, X_2, X_3, and X_4 denote petal width and length, and sepal width and length, respectively. Functions $l_1 = -7 - 3X_1 + 9X_2$, $l_2 = -52 + 11X_1 + 21X_2$, $l_3 = -93 + 23X_1 + 25X_2$.*

variables are restricted to petal length and width for CHAID and for the CRUISE and QUEST linear split models to allow their partitions to be plotted in the space of these two variables. No such restriction is necessary for the other methods because they only split on these two variables. Although the tree structures may appear different, the methods give the same predictions for a large majority of the observations. The plots show that the CHAID split points are rather poor and that those of C4.5 and CTREE are at observed data values.

Lim *et al.* (2000) compared the prediction accuracy and computation speed of 33 classification algorithms on a large number of data sets without missing values. Twenty-two algorithms were classification trees; two were neural networks; and the others included LDA, nearest neighbor, logistic regression, and POLYchotomous regression and multiple CLASSification (POLYCLASS) (Kooperberg *et al.*, 1997), a logistic regression model based on linear splines and their tensor products. POLYCLASS and logistic regression were found to have the lowest and second lowest, respectively, mean error rates. QUEST with linear splits ranked fourth best overall. POLYCLASS was, however, among the slowest. C4.5 trees had on average about twice as many terminal nodes as QUEST.

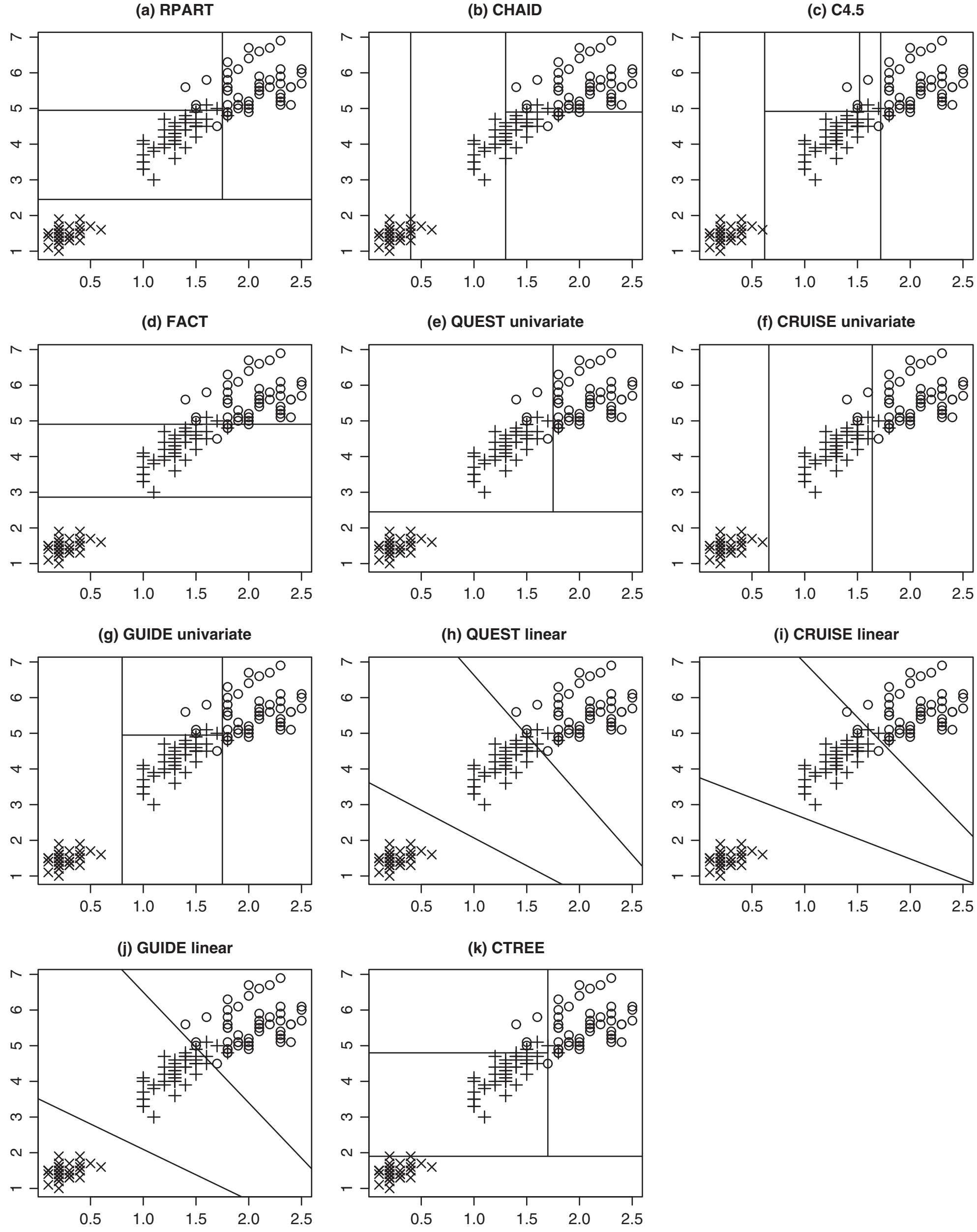


Figure 3. *Plots of petal length versus petal width with classification tree partitions; Setosa, Versicolour, and Virginica are marked by triangles, circles, and crosses, respectively.*

Perlich *et al.* (2004) compared logistic regression with C4.5 and found the former to be more accurate for small sample sizes and the latter better for large sample sizes. Loh (2009) found that among the newer classification tree algorithms, GUIDE had the best combination of accuracy and speed, followed by CRUISE and QUEST.

Ding & Simonoff (2010) studied the effectiveness of various missing value methods for classification trees with binary Y variables. After comparing several types of missing value mechanisms, they concluded that the use of a missing category to handle missing values (as used in CHAID and GUIDE) is best if the test sample has missing values and if missingness is not independent of Y.

3 Regression

AID and CART construct piecewise constant regression trees by using the node mean of Y as predicted value and the sum of squared deviations as node impurity function. Subsequent developments fall under one of two directions: (i) piecewise linear or higher order least-squares models and (ii) piecewise constant or linear models with other loss functions.

3.1 Least Squares

Although it is straightforward conceptually to extend the CART algorithm to piecewise linear models, this can be too time consuming in practice because it fits a linear model in each child node for *every* potential split of a node. To reduce the amount of computation, Alexander & Grimshaw (1996) proposed fitting a simple linear regression model in each node, with the linear predictor being the X variable yielding the smallest sum of squared residuals. M5 (Quinlan, 1992) and its implementation M5′ (Wang & Witten, 1996) fit a piecewise multiple linear tree model by using a less exhaustive but much faster approach. M5 first grows a piecewise constant tree and then fits a stepwise multiple linear model in each node t, using as linear predictors only those variables that are used to split the nodes below t. Thus, M5 avoids having to fit two linear models for every potential split. But because they are originally piecewise constant models, the M5 trees tend to be quite large. Torgo (1997) took a similar approach, but allowed kernel regression and nearest-neighbor models in addition to linear models in the terminal nodes.

Smoothed and Unsmoothed Piecewise POlynomial Regression Trees (SUPPORT) (Chaudhuri *et al.*, 1994) uses a different approach that applies classification tree techniques to the residuals. At each node, it first fits a linear model to the data and classifies the observations into two classes according to the signs of their residuals. Then, as in FACT, it performs two-sample tests of differences between the class means and the class variances for each X variable. The most significant X is selected to split the node with the split point being the average of the two class means. As a result, only one linear model needs to be fitted at each node. Conditions for asymptotic consistency of the function estimate and its derivatives from recursive partitioning methods are given in Chaudhuri *et al.* (1994).

It is harder to achieve unbiased variable selection in piecewise multiple linear regression trees because an X variable can be used in one or both of two roles: (a) as a candidate for split selection (called a “split” variable) and (b) as a linear predictor in the linear model (called a “fit” variable). Because the residuals are uncorrelated with split-and-fit variables, but are not necessarily uncorrelated with split-only variables, the p-values of the former tend to be stochastically larger than those of the latter. As a result, SUPPORT is biased toward selecting split-only variables. One way to correct the bias is to scale down the p-values (or scale up the test statistic values) of the split-and-fit variables. GUIDE (Loh, 2002) uses bootstrap calibration to find the scale factor.

There are also extensions in other directions. CTREE (Hothorn *et al.*, 2006b) uses permutation tests to construct unbiased piecewise constant regression trees for univariate, multivariate, ordinal, or censored Y variables. Regression Trunk Approach (RTA) (Dusseldorp & Meulman, 2004) combines the regression tree approach with linear regression to detect interactions between a treatment variable and ordered X variables. RTA first fits a linear main effects model to all the data. Then it uses the residuals to construct a piecewise constant regression tree model for each treatment group. Simultaneous Threshold Interaction Modeling Algorithm (STIMA) (Dusseldorp *et al.*, 2010) improves upon RTA by estimating the linear regression and tree models simultaneously.

Ciampi *et al.* (2002) proposed the use of soft thresholds (sigmoidal functions) instead of hard thresholds (indicator functions) for splits on ordered variables. Chipman *et al.* (2002) extended the Bayesian approach of Chipman *et al.* (1998) to piecewise linear regression trees. Fan & Gray (2005) and Gray & Fan (2008) used a genetic algorithm for tree construction. Guerts *et al.* (2006) proposed selecting splits from randomly picked subsets of split variables and split points. Su *et al.* (2004) extended CART by using maximum likelihood to choose the splits in a piecewise constant regression model, but instead of the usual negative log-likelihood, they used Akaike information criterion (AIC) and an independent test sample to prune the tree.

Yildiz & Alpaydin (2001, 2005a, 2005b) and Gama (2004) compared linear splitting with linear node modeling. Their results suggest that linear splitting and linear fitting yield similar gains in prediction accuracy, and both are superior to univariate splits and constant node models. Loh *et al.* (2007) showed that the prediction accuracy of piecewise linear regression trees can be improved by truncating or Winsorizing the fitted values. Kim *et al.* (2007) and Loh (2008b) showed that using only one or two regressor variables in the node models can be useful for data visualization.

3.1.1 Baseball example

To compare the methods, we use them to predict the 1987 salaries (in thousands of dollars) of 263 professional baseball players. The data, from Statlib (http://lib.stat.cmu.edu), contain 22 predictor variables, of which six are categorical, as shown in Table 1. They were used for a poster session contest at an American Statistical Association meeting. After reviewing the submitted solutions and performing their own analysis, Hoaglin & Velleman (1995) chose the following model fitted to log-salary:

$$\log(\texttt{Salary}) = \beta_0 + \beta_1 \texttt{Runcr/Yrs} + \beta_2\sqrt{\texttt{Run86}} + \beta_3 \min[(\texttt{Yrs}-2)_+, 5] + \beta_4(\texttt{Yrs}-7)_+.$$

Reasons for the data transformations include dealing with the range restriction on salary, collinearity, variance heterogeneity, and other difficulties typically encountered in linear regression.

As tree models are not limited by these difficulties, we fit salary without transformations to any variables. Figure 4 shows the RPART, CTREE, and two GUIDE tree models (one with

Table 1. *Predictor variables for baseball data.*

`Bat86`	# times at bat in 1986	`Batcr`	# times at bat during career
`Hit86`	# hits in 1986	`Hitcr`	# hits during career
`Hr86`	# home runs in 1986	`Hrcr`	# home runs during career
`Run86`	# runs in 1986	`Runcr`	# runs during career
`Rb86`	# runs batted in 1986	`Rbcr`	# runs batted in during career
`Wlk86`	# walks in 1986	`Wlkcr`	# walks during career
`Leag86`	league at end of 1986 (2 cat.)	`Leag87`	league at start of 1987 (2 cat.)
`Team86`	team at end of 1986 (24 cat.)	`Team87`	team at start of 1987 (24 cat.)
`Div86`	division at end of 1986 (2 cat.)	`Yrs`	# years in the major leagues
`Pos86`	position in 1986 (23 cat.)	`Puto86`	# put outs in 1986
`Asst86`	# assists in 1986	`Err86`	# errors in 1986

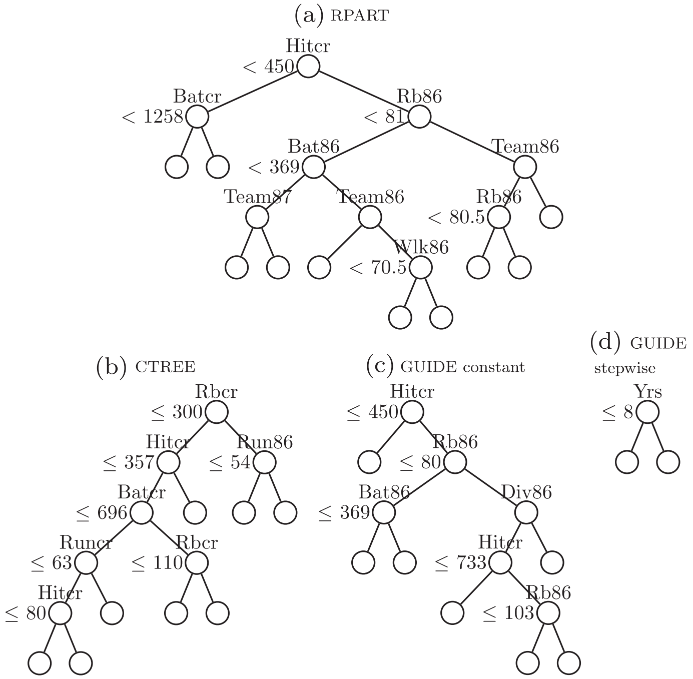


Figure 4. *Regression tree models for baseball data.*

a constant fitted in each node and the other with a stepwise linear model in each node). The initial splits of the RPART and GUIDE piecewise constant trees are identical except for one split point. RPART, however, shows a preference for splits on the two `Team` variables, which have more than eight million ($2^{23} - 1$) splits each. None of the piecewise constant models select `Yrs`, which features prominently in the Hoaglin–Velleman model. The piecewise linear GUIDE model, in contrast, splits just once, and on `Yrs`. The tree is short because each node is fitted with a multiple linear model.

Figure 5 plots the observed versus fitted values of the tree models and those from ordinary least squares, Hoaglin and Velleman, Random forest, and GUIDE forest. The two forest methods are similar, their only difference being that Random forest uses the CART algorithm for split selection and GUIDE forest uses its namesake algorithm. Although the ordinary least-squares model compares quite favorably with the others, it has a fair number of negative fitted values. The piecewise constant models are easily identified by the vertical stripes in the plots. As every model has trouble predicting the highest salaries, we conclude these salaries cannot be adequately explained by the variables in the data. We note that the GUIDE stepwise model (which uses a single tree) fits the data about as well as Random forest (which uses 500 trees) and that GUIDE forest fits the data visibly better than Random forest here.

3.2 Poisson, Logistic, and Quantile Regression

Efforts have been made to extend regression tree methods beyond squared error loss. Ciampi (1991) extended CART to fit a generalized linear regression model in each node, choosing the split that most reduces the sum of deviances in the children nodes. The trees are pruned by significance tests or the AIC.

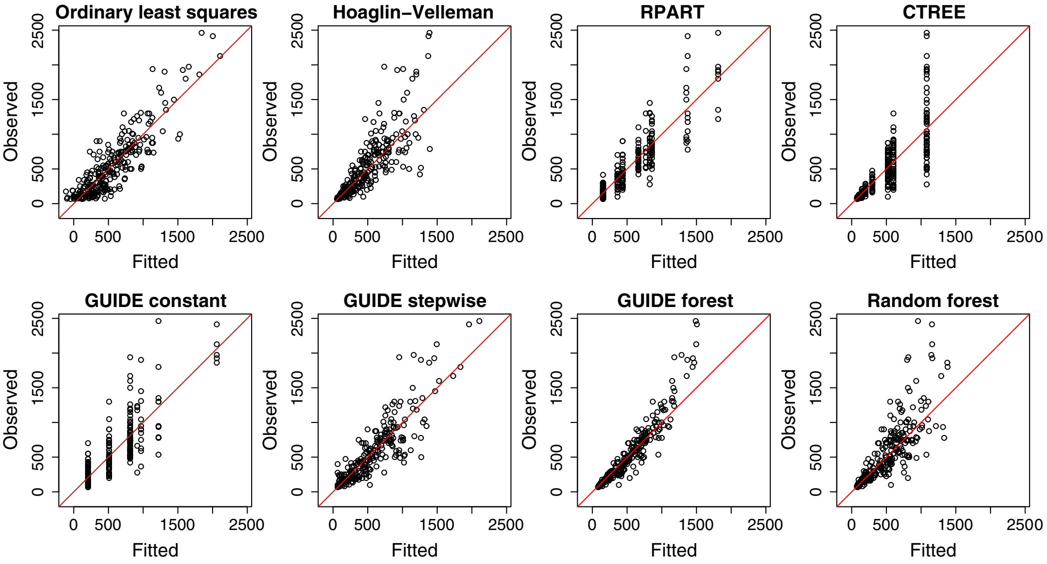


Figure 5. *Plots of observed versus fitted values for baseball data.*

Chaudhuri *et al.* (1995) extended SUPPORT to piecewise linear Poisson and logistic regression. For Poisson regression, they used adjusted Anscombe residuals. For logistic regression, they estimated the probability function at each node by using both logistic regression and a nearest-neighbor method and defined the "residual" as the difference between the two estimated values. Ahn & Chen (1997) used similar ideas to construct logistic regression trees for clustered binomial data.

Chaudhuri & Loh (2002) and Loh (2006b) extended GUIDE to quantile and Poisson regression, respectively. LOTUS (Chan & Loh, 2004; Loh, 2006a) uses the same ideas to fit a linear logistic regression model in each node. To attain unbiasedness, LOTUS uses a trend-adjusted chi-squared test for X variables that are used for splitting and fitting.

Landwehr *et al.* (2005) constructed logistic regression trees by using the LogitBoost (Friedman *et al.*, 2000) technique to fit a logistic regression model to each node and the C4.5 method to split the nodes. Choi *et al.* (2005) used the GUIDE split selection method to construct regression trees for overdispersed Poisson data. Lee & Yu (2010) employed the CART approach to model ranking response data.

MOdel-Based recursive partitioning (MOB) (Zeileis *et al.*, 2008) fits least-squares, logistic, and other models, using the score functions of M-estimators. Special cases include standard maximum and pseudo-likelihood models. It achieves unbiased variable selection by choosing split variables on the basis of structural break tests for the score function. The split point (for ordered X) or split set (for categorical X) is obtained by maximizing the change in an objective function. The tree is not pruned; instead, stopping rules based on Bonferroni-adjusted p-values are used to control tree growth. MOB has been extended to psychometric models such as the Bradley–Terry model (Strobl *et al.*, 2011) and the Rasch model (Strobl *et al.*, 2010) and to generalized linear models and maximum likelihood models with linear predictors (Rusch & Zeileis, 2013). The algorithm is not unbiased if some X variables are used for both fitting and splitting.

3.3 Censored Response Variables

Gordon & Olshen (1985) extended CART to censored response variables by fitting a Kaplan–Meier survival curve to the data in each node and using as node impurity the minimum Wasserstein distance between the fitted Kaplan–Meier curve and a point-mass function. Segal (1988) chose splits to maximize a measure of between-node difference instead of within-node homogeneity. The measures include two-sample rank statistics such as the logrank test (Peto & Peto, 1972). Davis & Anderson (1989) adapted CART to fit a constant hazard to each node, using exponential log-likelihood as impurity function.

Ciampi *et al.* (1986) compared stepwise Cox regression, correspondence analysis, and recursive partitioning models for censored response data. Stepwise Cox regression finds a prognostic index (a linear combination of the X variables) and then partitions the data at the quartiles of the index. Correspondence analysis converts each ordered X into a vector of indicator variables (one indicator for each observed value) and groups the Y values into a small number of categories. The first canonical variable is used as the prognostic index to partition the data. Recursive partitioning converts each categorical variable into a vector of indicators and partitions the data on the indicators. RECursive Partitioning and AMalgamation (RECPAM) (Ciampi *et al.*, 1988) extends these ideas to allow merging of terminal nodes. For regression with censored response data, the split criterion is a dissimilarity measure such as likelihood ratio or the logrank, Wilcoxon–Gehan, and Kolmogorov–Smirnov statistics. For classification, the split criterion is the multinomial likelihood. Splits may be univariate or Boolean intersections of univariate splits. Missing values may be given a separate category or be dealt with through surrogates splits as in CART. Importance scores are given by the sum of the dissimilarities of each variable over all the nodes. Tree size is determined by cross-validation or AIC.

LeBlanc & Crowley (1992) fitted a proportional hazards model with the hazard rate in node t being $\lambda_t(u) = \theta_t \lambda_0(u)$, where θ_t is a constant and $\lambda_0(u)$ is the baseline hazard function. For tree construction and pruning, the baseline cumulative hazard $\Lambda_0(u)$ is estimated by the Nelson–Aalen estimator (Aalen, 1978). The θ_t is a one-step estimate from the full maximum likelihood. Split selection and pruning are based on the one-step deviance. The rest of the algorithm follows CART. LeBlanc & Crowley (1993) used logrank test statistics to select splits and the sum of logrank test statistics over intermediate nodes as measure of goodness of split for pruning. Crowley *et al.* (1995) noted that, without a node impurity measure, cross-validation pruning cannot be employed with Segal's (1988) approach. They also showed that the split selection method of Gordon & Olshen (1985), based on L_p and Wasserstein metrics, can perform poorly even with mild censoring. Bacchetti & Segal (1995) considered left-truncated survival times and splits on time-dependent covariates, by letting each observation go into both child nodes at the same time. This approach precludes classifying each subject in exactly one terminal node. They noted that splits on time-dependent variables can yield unstable Kaplan–Meier estimates of the survival functions.

Jin *et al.* (2004) used between-node variance of restricted mean survival time as node impurity to construct survival trees. For clustered survival data, Gao *et al.* (2004) fitted a proportional hazards model with subject frailty to each node; see also Su & Fan (2004) and Fan *et al.* (2006). Hothorn *et al.* (2004) used bagging to obtain an ensemble of survival trees and obtained a Kaplan–Meier survival curve for each subject from the bootstrap observations belonging to the same terminal nodes as the subject. Molinaro *et al.* (2004) used inverse probabilities of censoring as weights to construct trees for censored data. Hothorn *et al.* (2006a) employed the idea to predict mean log survival time from random forests with case weights. Ishwaran *et al.* (2004) applied the random forest (Breiman, 2001) technique to construct relative risk forests using piecewise proportional hazards; see also Ishwaran *et al.* (2006) who obtained variable importance scores. Clarke & West (2008) fitted Bayesian Weibull tree models to uncensored survival data with split criteria on the basis of Bayes factors, and Garg *et al.* (2011) used a similar approach to fit exponential models. Cho & Hong (2008) constructed median regression trees by using the Buckley–James (1979) method to estimate the survival times of the censored observations and then fitting a piecewise constant quantile regression model to the completed data.

Loh (1991) and Ahn & Loh (1994) extended SUPPORT to piecewise proportional hazards models. Ahn (1994a, 1994b, 1996a, 1996b) did the same for piecewise parametric survival models.

3.4 Longitudinal and Multiresponse Variables

Segal (1992) was among the first to extend CART to longitudinal data by using as node impurity a function of the likelihood of an autoregressive or compound symmetry model. If there are missing response values, the expectation–maximization (EM) algorithm (Laird & Ware, 1982) is used to estimate the parameters. Abdolell *et al.* (2002) used the same approach, but with a likelihood-ratio test statistic as impurity function.

Zhang (1998) extended CART to multiple binary response variables, using as node impurity the log-likelihood of an exponential family distribution that depends only on the linear terms and the sum of second-order products of the responses. Zhang & Ye (2008) applied the technique to ordinal responses by first transforming them to binary-valued indicator functions; see also Zhang & Singer (2010). Their approach requires covariance matrices to be computed at every node.

For longitudinal data observed at very many times, Yu & Lambert (1999) treated each response vector as a random function and reduced the dimensionality of the data by fitting each trajectory with a spline curve. Then they used the estimated coefficients of the basis functions as multivariate responses to fit a regression tree model.

De'ath (2002) avoided the problem of covariance estimation by using as node impurity the total sum of squared deviations from the mean across the response variables. Larsen & Speckman (2004) used the Mahalanobis distance, but estimated the covariance matrix from the whole data set.

Hsiao & Shih (2007) showed that multivariate extensions of CART are biased toward selecting variables that allow more splits. They proposed using chi-squared tests of conditional independence (conditioning on the components of the response vector) of residual signs versus grouped X values to select the split variables. The method may lack power if the effects of the X variables are not in the same direction across all the Y variables.

Lee (2005) applied the GUIDE approach to multiple responses with ordered X variables by fitting a generalized estimating equation model to the data in each node and taking the average of the Pearson residuals over the responses variables, for each observation. The observations are classified into two groups according to the signs of the average residuals, and the X with the smallest p-value from two-sample t-tests is chosen to split the node. Although unbiased, the method is not sensitive to all response trajectory shapes. Loh & Zheng (2013) solved this problem by using the residual vector patterns, rather than their averages, to choose the split variables. The solution is applicable to data observed at random time points.

Sela & Simonoff (2012) proposed the RE-EM method (Sela & Simonoff, 2011), which fits a model consisting of the sum of a random effects term and a tree-structured term.

The procedure mimics the EM algorithm (Laird & Ware, 1982) by iterating between estimating the tree structure, assuming that the random effects are correct, and estimating the random effects, assuming that the tree structure is correct.

4 Conclusion

Research in classification and regression trees has seen rapid growth, and applications are increasing at an even greater rate. Interpretability of the tree structures is a strong reason for their popularity among practitioners, but so are reasonably good prediction accuracy, fast computation speed, and wide availability of software.

Despite 50 years of progress, however, many hard problems remain. One of them is how best to deal with missing values. Ding & Simonoff (2010) made a good start, but their results apply only to classification with binary responses. Much of the difficulty is due to missing value techniques interacting with other algorithm components and with the type of variables and the causes of the missingness. Another challenging problem is how to deal with time-varying covariates in regression trees for longitudinal and censored response data. This is not surprising given that traditional (non-tree) solutions require various model assumptions that are hard to justify in a tree-structured framework. Computationally efficient approaches to search for effective linear combination splits is yet another elusive problem, especially in the regression context. It is harder if the linear splits are coupled with linear model fits in the nodes, because the latter are already quite effective in reducing prediction error. Thus, the linear splits need to be so much more effective to justify the increase in computation and loss of interpretability. In this age of large data sets, there are also new problems, such as algorithms that scale well with sample size (Dobra & Gehrke, 2002; Gehrke, 2009) and incremental tree construction algorithms for streaming data (Alberg *et al.*, 2005; Potts & Sammut, 2011; Taddy *et al.*, 2012).

The rise of ensemble and other methods has made it difficult for single-tree methods to compete in terms of prediction accuracy alone. Comparisons based on real and simulated data sets suggest that the accuracy of the best single-tree algorithm is on average about 10% less than that of a tree ensemble, although it is certainly not true that an ensemble always beats a single tree (Loh, 2009). An ensemble of, say, 500 trees is, however, often practically impossible to understand. Importance scores can rank order the variables, but they do not explain how the variables influence the predictions. Thus, the biggest advantage of single-tree models remains their model interpretability, although interpretability rapidly diminishes with tree size. But because inferences from the tree structures can be compromised by selection bias, future algorithms will need to be unbiased to be useful in applications where interpretability is important.

Acknowledgements

The author is grateful to Probal Chaudhuri for suggesting and encouraging him to write this survey and to four anonymous reviewers for their helpful comments. Research for this article was partially supported by US Army Research Office grant W911NF-09-1-0205, National Institutes of Health grant P50CA143188, and Eli Lilly and Company. CART® is a registered trademark of California Statistical Software, Inc.

References

Aalen, O.O. (1978). Nonparametric inference for a family of counting processes. *Ann. Stat.*, **6**, 701–726.

Abdolell, M., LeBlanc, M., Stephens, D. & Harrison, R.V. (2002). Binary partitioning for continuous longitudinal data: categorizing a prognostic variable. *Stat. Med.*, **21**, 3395–3409.

Ahn, H. (1994a). Tree-structured exponential regression modeling. *Biometrical J.*, **36**, 43–61.

Ahn, H. (1994b). Tree-structured extreme value model regression. *Commun. Stat.-Theor. M.*, **23**, 153–174.

Ahn, H. (1996a). Log-gamma regression modeling through regression trees. *Commun. Stat.-Theor. M.*, **25**, 295–311.

Ahn, H. (1996b). Log-normal regression modeling through recursive partitioning. *Comput. Stat. Data Anal.*, **21**, 381–398.

Ahn, H. & Chen, J. (1997). Tree-structured logistic model for over-dispersed binomial data with application to modeling developmental effects. *Biometrics*, **53**, 435–455.

Ahn, H. & Loh, W.-Y. (1994). Tree-structured proportional hazards regression modeling. *Biometrics*, **50**, 471–485.

Alberg, D., Last, M. & Kandel, A. (2012). Knowledge discovery in data streams with regression tree methods. *Wil. Interdiscip. Rev.: Data Mining and Knowledge Disc.*, **2**, 69–78.

Alexander, W.P. & Grimshaw, S.D. (1996). Treed regression. *J. Comput. Graph. Stat.*, **5**, 156–175.

Bacchetti, P. & Segal, M.R. (1995). Survival trees with time-dependent covariates: application to estimating changes in the incubation period of AIDS. *Lifetime Data Anal.*, **1**, 35–47.

Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis.* Cambridge, MA: MIT Press.

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984). *Classification and Regression Trees.* Belmont: Wadsworth.

Breiman, L. (1996). Bagging predictors. *Mach. Learn.*, **24**, 123–140.

Breiman, L. (2001). Random forests. *Mach. Learn.*, **45**, 5–32.

Brown, D.E., Pittard, C.L. & Park, H. (1996). Classification trees with optimal multivariate decision nodes. *Pattern. Recogn. Lett.*, **17**, 699–703.

Buckley, J.J. & James, I.R. (1979). Linear regression with censored data. *Biometrika*, **66**, 429–436.

Buttrey, S.E. & Karo, C. (2002). Using k-nearest-neighbor classification in the leaves of a tree. *Comput. Stat. Data Anal.*, **40**, 27–37.

Chan, K.-Y. & Loh, W.-Y. (2004). LOTUS: an algorithm for building accurate and comprehensible logistic regression trees. *J. Comput. Graph. Stat.*, **13**, 826–852.

Chaudhuri, P., Huang, M.-C., Loh, W.-Y. & Yao, R. (1994). Piecewise-polynomial regression trees. *Stat. Sinica*, **4**, 143–167.

Chaudhuri, P., Lo, W.-D., Loh, W.-Y. & Yang, C.-C. (1995). Generalized regression trees. *Stat. Sinica*, **5**, 641–666.

Chaudhuri, P. & Loh, W.-Y. (2002). Nonparametric estimation of conditional quantiles using quantile regression trees. *Bernoulli*, **8**, 561–576.

Chipman, H.A., George, E.I. & McCulloch, R.E. (1998). Bayesian CART model search. *J. Amer. Statist. Assoc.*, **93**, 935–948.

Chipman, H.A., George, E.I. & McCulloch, R.E. (2002). Bayesian treed models. *Mach. Learn.*, **48**, 299–320.

Cho, H.J. & Hong, S.-M. (2008). Median regression tree for analysis of censored survival data. *IEEE T. Syst. Man Cy. A.*, **38**, 715–726.

Choi, Y., Ahn, H. & Chen, J. (2005). Regression trees for analysis of count data with extra Poisson variation. *Comput. Stat. Data Anal.*, **49**, 893–915.

Chou, P.A. (1991). Optimal partitioning for classification and regression trees. *IEEE T. Pattern. Anal.*, **13**, 340–354.

Ciampi, A. (1991). Generalized regression trees. *Comput. Stat. Data. Anal.*, **12**, 57–78.

Ciampi, A., Couturier, A. & Li, S.L. (2002). Prediction trees with soft nodes for binary outcomes. *Stat. Med.*, **21**, 1145–1165.

Ciampi, A., Hogg, S.A., McKinney, S. & Thiffault, J. (1988). RECPAM: A computer program for recursive partition and amalgamation for censored survival data and other situations frequently occurring in biostatistics. *Comput. Meth. Prog. Bio.*, **26**, 239–256.

Ciampi, A., Thiffault, J., Nakache, J-P. & Asselain, B. (1986). Stratification by stepwise regression, correspondence analysis and recursive partition: A comparison of three methods of analysis for survival data with covariates. *Comput. Stat. Data Anal.*, **4**, 185–204.

Clarke, J. & West, M. (2008). Bayesian Weibull tree models for survival analysis of clinico-genomic data. *Stat. Meth.*, **5**, 238–262.

Crowley, J., LeBlanc, M., Gentleman, R. & Salmon, S. (1995). Exploratory methods in survival analysis. In *Analysis of Censored Data*, vol. 27, Eds. H.L. Koul & J.V. Deshpande, pp. 55–77. Institute of Mathematical Statistics, IMS Lecture Notes-Monograph Series: Hayward, CA.

Davis, R.B. & Anderson, J.R. (1989). Exponential survival trees. *Stat. Med.*, **8**, 947–961.

De'ath, G. (2002). Multivariate regression trees: a new technique for modeling species-environment relationships. *Ecology*, **83**, 1105–1117.

De'ath, G. & Fabricius, K.E. (2000). Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology*, **81**, 3178–3192.

Denison, D.G.T., Mallick, B.K. & Smith, A.F.M. (1998). A Bayesian CART algorithm. *Biometrika*, **85**, 363–377.

Dietterich, T.G. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems, LBCS-1857,* pp. 1–15. New York: Springer.

Ding, Y. & Simonoff, J.S. (2010). An investigation of missing data methods for classification trees applied to binary response data. *J. Mach. Learn. Res.*, **11**, 131–170.

Dobra, A. & Gehrke, J.E. (2002). SECRET: A scalable linear regression tree algorithm. In *Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp. 481–487. Edmonton, Canada: ACM Press.

Doyle, P. (1973). The use of automatic interaction detector and similar search procedures. *Oper. Res. Quart.*, **24**, 465–467.

Dusseldorp, E., Conversano, C. & Van Os, B.J. (2010). Combining an additive and tree-based regression model simultaneously: STIMA. *J. Comput. Graph. Stat.*, **19**, 514–530.

Dusseldorp, E. & Meulman, J.J. (2004). The regression trunk approach to discover treatment covariate interaction. *Psychometrika*, **69**, 355–374.

Einhorn, H.J. (1972). Alchemy in the behavioural sciences. *Public Opin. Quart.*, **36**, 367–378.

Fan, G. & Gray, J.B. (2005). Regression tree analysis using TARGET. *J. Comput. Graph. Stat.*, **14**, 1–13.

Fan, J., Su, X., Levine, R.A., Nunn, M.E. & LeBlanc, M. (2006). Trees for correlated survival data by goodness of split, with applications to tooth prognosis. *J. Amer. Statist. Assoc.*, **101**, 959–967.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugenic.*, **7**(2), 179–188.

Freund, Y. & Schapire, R.E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, **55**(1), 119–139.

Friedman, J., Hastie, T. & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *Ann. Stat.*, **38**(2), 337–374.

Gama, J. (2004). Functional trees. *Mach. Learn.*, **55**, 219–250.

Gao, F., Manatunga, A.K. & Chen, S. (2004). Identification of prognostic factors with multivariate survival data. *Comput. Stat. Data Anal.*, **45**, 813–824.

Garg, L., McClean, S., Meenan, B.J. & Millard, P. (2011). Phase-type survival trees and mixed distribution survival trees for clustering patients' hospital length of stay. *Informatica*, **22**, 57–72.

Gashler, M., Giraud-Carrier, C.G. & Martinez, T.R. (2008). Decision tree ensemble: small heterogeneous is better than large homogeneous. In *Seventh International Conference on Machine Learning and Applications*, pp. 900–905. Washington, DC: IEEE Computer Society.

Gehrke, J. (2009). Scalable decision tree construction. In *Encyclopedia of Database Systems*, Eds. L. Liu & T. Ozsu, pp. 2469–2474. New York: Springer.

Gordon, L. & Olshen, R.A. (1985). Tree-structured survival analysis. *Cancer Treat. Rep.*, **69**, 1065–1069.

Gray, J.B. & Fan, G. (2008). Classification tree analysis using TARGET. *Comput. Stat. Data Anal.*, **52**, 1362–1372.

Guerts, P., Ernst, D. & Wehenkel, L. (2006). Extremely randomized trees. *Mach. Learn.*, **63**, 3–42.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I.H. (2009). The WEKA data mining software: an update. *SIGKDD Explorations*, **11**, 10–18.

Hamza, M. & Larocque, D. (2005). An empirical comparison of ensemble methods based on classification trees. *J. Stat. Comput. Sim.*, **75**, 629–643.

Harper, P.R. (2005). A review and comparison of classification algorithms for medical decision making. *Health Policy*, **71**, 315–331.

Henrichon, Jr., E.G. & Fu, K.-S. (1969). A nonparametric partitioning procedure for pattern classification. *IEEE T. Comput.*, **C-18**, 614–624.

Hoaglin, D.C. & Velleman, P.F. (1995). A critical look at some analyses of major league baseball salaries. *Am. Stat.*, **49**, 277–285.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A. & van der Laan, M.J. (2006a). Survival ensembles. *Biostatistics*, **7**, 355–373.

Hothorn, T., Hornik, K. & Zeileis, A. (2006b). Unbiased recursive partitioning: a conditional inference framework. *J. Comput. Graph. Stat.*, **15**, 651–674.

Hothorn, T. & Lausen, B. (2005). Bundling classifiers by bagging trees. *Comput. Stat. Data Anal.*, **49**, 1068–1078.

Hothorn, T., Lausen, B., Benner, A. & Radespiel-Tröger, M. (2004). Bagging survival trees. *Stat. Med.*, **23**, 77–91.

Hsiao, W.-C. & Shih, Y.-S. (2007). Splitting variable selection for multivariate regression trees. *Stat. Probab. Lett.*, **77**, 265–271.

Ishwaran, H., Blackstone, E.H., Pothier, C.E. & Lauer, M. (2004). Relative risk forests for exercise heart rate recovery as a predictor of mortality. *J. Amer. Statist. Assoc.*, **99**, 591–600.

Ishwaran, H., Kogalur, U.B., Blackstone, E.H. & Lauer, M. (2006). Random survival forests. *Ann. Appl. Stat.*, **2**, 841–860.

Jin, H., Lu, Y., Stone, K. & Black, D.M. (2004). Survival analysis based on variance of survival time. *Med. Decis. Making*, **24**, 670–680.

Kass, G.V. (1980). An exploratory technique for investigating large quantities of categorical data. *Ann. Appl. Stat.*, **29**, 119–127.

Kim, H. & Loh, W.-Y. (2001). Classification trees with unbiased multiway splits. *J. Amer. Statist. Assoc.*, **96**, 589–604.

Kim, H. & Loh, W.-Y. (2003). Classification trees with bivariate linear discriminant node models. *J. Comput. Graph. Stat.*, **12**, 512–530.

Kim, H., Loh, W.-Y., Shih, Y.-S. & Chaudhuri, P. (2007). Visualizable and interpretable regression models with good prediction power. *IIE Transactions*, **39**, 565–579.

Kononenko, I. (1995). On biases in estimating multi-valued attributes. In *14th International Joint Conference on Artificial Intelligence*, pp. 1034–1040. Burlington, MA: Morgan Kaufmann.

Kooperberg, C., Bose, S. & Stone, C.J. (1997). Polychotomous regression. *J. Amer. Statist. Assoc.*, **92**, 117–127.

Laird, N.M. & Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Landwehr, N., Hall, M. & Frank, E. (2005). Logistic model trees. *Mach. Learn.*, **59**, 161–205.

Larsen, D.R. & Speckman, P.L. (2004). Multivariate regression trees for analysis of abundance data. *Biometrics*, **60**, 543–549.

LeBlanc, M. & Crowley, J. (1992). Relative risk trees for censored survival data. *Biometrics*, **48**, 411–425.

LeBlanc, M. & Crowley, J. (1993). Survival trees by goodness of split. *J. Amer. Statist. Assoc.*, **88**, 457–467.

Lee, P.H. & Yu, P.L.H. (2010). Distance-based tree models for ranking data. *Comput. Stat. Data Anal.*, **54**, 1672–1682.

Lee, S.K. (2005). On generalized multivariate decision tree by using GEE. *Comput. Stat. Data Anal.*, **49**, 1105–1119.

Lemon, S.C., Roy, J., Clark, M.A., Friedman, P.D. & Rakowski, W. (2003). Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression. *Ann. Behav. Med.*, **26**, 172–181.

Light, R.J. & Margolin, B.H. (1971). An analysis of variance for categorical data. *J. Amer. Statist. Assoc.*, **66**, 534–544.

Lim, T.-S., Loh, W.-Y. & Shih, Y.-S. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Mach. Learn.*, **40**, 203–228.

Loh, W.-Y. (1991). Survival modeling through recursive stratification. *Comput. Stat. Data Anal.*, **12**, 295–313.

Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Stat. Sinica*, **12**, 361–386.

Loh, W.-Y. (2006a). Logistic regression tree analysis. In *Handbook of engineering statistics,* Ed. H. Pham, pp. 537–549. New York: Springer.

Loh, W.-Y. (2006b). Regression tree models for designed experiments. In *Second E. L. Lehmann Symposium*, Vol. 49, Ed. J. Rojo. IMS Lecture Notes-Monograph Series, pp. 210–228. Bethesda, MD: Institute of Mathematical Statistics.

Loh, W.-Y. (2008a). Classification and regression tree methods. In *Encyclopedia of Statistics in Quality and Reliability*, Eds. F. Ruggeri, R. Kenett & F.W. Faltin, pp. 315–323. Chichester, UK: Wiley.

Loh, W.-Y. (2008b). Regression by parts: fitting visually interpretable models with GUIDE. In *Handbook of Data Visualization,* Eds. C. Chen, W. Härdle & A. Unwin. pp. 447–469. New York: Springer.

Loh, W.-Y. (2009). Improving the precision of classification trees. *Ann. Appl. Stat.*, **3**, 1710–1737.

Loh, W.-Y. (2011). Classification and regression trees. *Wil. Interdiscip. Rev.: Data Mining and Knowledge Disc.*, **1**, 14–23.

Loh, W.-Y. (2012). Variable selection for classification and regression in large p, small n problems. In *Probability approximations and beyond*, Vol. 205, Eds. A. Barbour, H.P. Chan & D. Siegmund. Lecture Notes in Statistics—Proceedings. pp. 133–157. New York: Springer.

Loh, W.-Y., Chen, C.-W. & Zheng, W. (2007). Extrapolation errors in linear model trees. *ACM Trans. On Knowledge Disc. From Data*, **1**. DOI: 10.1145/1267066.1267067.

Loh, W.-Y. & Shih, Y.-S. (1997). Split selection methods for classification trees. *Stat. Sinica*, **7**, 815–840.

Loh, W.-Y. & Vanichsetakul, N. (1988). Tree-structured classification via generalized discriminant analysis (with discussion). *J. Amer. Statist. Assoc.*, **83**, 715–728.

Loh, W.-Y. & Zheng, W. (2013). Regression trees for longitudinal and multiresponse data. *Ann. Appl. Stat.*, **7**(1), 495–522.

Meisel, W.S. & Michalopoulos, D.A. (1973). A partitioning algorithm with application in pattern classification and the optimization of decision trees. *IEEE T. Comput.*, **C-22**, 93–103.

Merkle, E.C. & Shaffer, V.A. (2011). Binary recursive partitioning: Background, methods, and application to psychology. *Brit. J. Math. Stat. Psy.*, **64**, 161–181.

Messenger, R. & Mandell, L. (1972). A modal search technique for predictive nominal scale multivariate analysis. *J. Amer. Statist. Assoc.*, **67**, 768–772.

Molinaro, A.M., Dudoit, S. & van der Laan, M.J. (2004). Tree-based multivariate regression and density estimation with right-censored data. *J. Multivariate Anal.*, **90**, 154–177.

Morgan, J.N. & Sonquist, J.A. (1963). Problems in the analysis of survey data, and a proposal. *J. Amer. Statist. Assoc.*, **58**, 415–434.

Murthy, S.K. (1998). Automatic construction of decision trees from data: a multi-disciplinary survey. *Data Min. Knowl. Disc.*, **2**, 345–389.

Payne, H.J. & Meisel, W.S. (1977). An algorithm for constructing optimal binary decision trees. *IEEE T. Comput.*, **C-26**, 905–916.

Perlich, C., Provost, F. & Simonoff, J.S. (2004). Tree induction vs. logistic regression: A learning-curve analysis. *J. Mach. Learn. Res.*, **4**, 211–255.

Peto, R. & Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *J. R. Stat. Soc. Ser.-A*, **135**, 185–207.

Potts, D. & Sammut, C. (2005). Incremental learning of linear model trees. *Mach. Learn.*, **61**, 5–48.

Quinlan, J.R. (1986). Induction of decision trees. *Mach. Learn.*, **1**, 81–106.

Quinlan, J.R. (1992). Learning with continuous classes. In *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence,* pp. 343–348. Singapore: World Scientific.

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

R Core Team. (2014). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0.

Rusch, T. & Zeileis, A. (2013). Gaining insight with recursive partitioning of generalized linear models. *J. Stat. Comput. Sim.*, **83**(7), 1301–1315.

Sandri, M. & Zuccolotto, P. (2008). A bias correction algorithm for the Gini variable importance measure in classification trees. *J. Comput. Graph. Stat.*, **17**, 611–628.

Segal, M.R. (1988). Regression trees for censored data. *Biometrics*, **44**, 35–47.

Segal, M.R. (1992). Tree structured methods for longitudinal data. *J. Amer. Statist. Assoc.*, **87**, 407–418.

Sela, R.J. & Simonoff, J.S. (2011). *Reemtree: Regression trees with random effects*. R package version 0.90.3.

Sela, R.J. & Simonoff, J.S. (2012). RE-EM trees: A data mining approach for longitudinal and clustered data. *Mach. Learn.*, **86**, 169–207.

Sethi, I.K. & Chatterjee, B. (1977). Efficient decision tree design for discrete variable pattern recognition problems. *Pattern Recogn.*, **9**, 197–206.

Shih, Y.S. (2004). A note on split selection bias in classification trees. *Comput. Stat. Data Anal.*, **45**, 457–466.

Shih, Y.S. & Tsai, H.W. (2004). Variable selection bias in regression trees with constant fits. *Comput. Stat. Data Anal.*, **45**, 595–607.

Smyth, P., Gray, A. & Fayyad, U.M. (1995). Retrofitting decision tree classifiers using kernel density estimation. In *Proceedings of the 12th International Conference on Machine Learning,* pp. 506–514. Burlington, MA: Morgan Kaufmann.

Strobl, C., Boulesteix, A. & Augustin, T. (2007a). Unbiased split selection for classification trees based on the Gini index. *Comput. Stat. Data Anal.*, **52**, 483–501.

Strobl, C., Boulesteix, A., Kneib, T., Augustin, T. & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, **9**, 307.

Strobl, C., Boulesteix, A., Zeileis, A. & Hothorn, T. (2007b). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, **8**, 25.

Strobl, C., Kopf, J. & Zeileis, A. (2010). A new method for detecting differential item functioning in the Rasch model. Tech. Rep. 92, Department of Statistics, Ludwig-Maximilians-Universitaet Muenchen.

Strobl, C., Malley, J. & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forests. *Psychol. Methods*, **14**, 323–348.

Strobl, C., Wickelmaier, F. & Zeileis, A. (2011). Accounting for individual differences in Bradley–Terry models by means of recursive partitioning. *J. Educ. Behav. Stat.*, **36**(2), 135–153.

Su, X. & Fan, J. (2004). Multivariate survival trees: A maximum likelihood approach based on frailty models. *Biometrics*, **60**, 93–99.

Su, X. G., Wang, M. & Fan, J.J. (2004). Maximum likelihood regression trees. *J. Comput. Graph. Stat.*, **13**, 586–598.

Taddy, M.A., Gramacy, R.B. & Polson, N.G. (2011). Dynamic trees for learning and design. *J. Amer. Statist. Assoc.*, **106**, 109–123.

Therneau, T.M. & Atkinson, B. (2011). *rpart: Recursive partitioning*, R port by Brian Ripley. R package version 3.1-50.

Torgo, L. (1997). Functional models for regression tree leaves. In *Proceedings of the Fourteenth International Conference on Machine Learning,* Ed. D.H. Fisher. pp. 385–393. Burlington, MA: Morgan Kaufmann.

Wang, Y. & Witten, I.H. (1996). *Induction of model trees for predicting continuous classes*, Working paper series, Department of Computer Science, University of Waikato.

White, A.P. & Liu, W.Z. (1994). Technical note: bias in information-based measures in decision tree induction. *Mach. Learn.*, **15**, 321–329.

Yildiz, O.T. & Alpaydin, E. (2001). Omnivariate decision trees. *IEEE T. Neural. Networ.*, **12**, 1539–1546.

Yildiz, O.T. & Alpaydin, E. (2005a). Linear discriminant trees. *Int. J. Pattern Recogn.*, **19**, 323–353.

Yildiz, O.T. & Alpaydin, E. (2005b). Model selection in omnivariate decision trees. In *Machine learning: ECML 2005, Proceedings*, Vol. 3720, Eds. J. Gama, R. Camacho, P. Brazdil, A. Jorge & L. Torgo. Lecture Notes in Artificial Intelligence, pp. 473–484. Berlin: Springer-Verlag.

Yu, Y. & Lambert, D. (1999). Fitting trees to functional data, with an application to time-of-day patterns. *J. Comput. Graph. Stat.*, **8**, 749–762.

Zeileis, A., Hothorn, T. & Hornik, K. (2008). Model-based recursive partitioning. *J. Comput. Graph. Stat.*, **17**, 492–514.

Zhang, H. (1998). Classification trees for multiple binary responses. *J. Amer. Statist. Assoc.*, **93**, 180–193.

Zhang, H. & Singer, B.H. (2010). *Recursive Partitioning and Applications*, 2nd ed. New York: Springer.

Zhang, H. & Ye, Y. (2008). A tree-based method for modeling a multivariate ordinal response. *Stat. and Its Interface*, **1**, 169–178.

[*Received February 2012, accepted February 2013*]