Statistica Sinica **7**(1997), 815-840

SPLIT SELECTION METHODS FOR CLASSIFICATION TREES

Wei-Yin Loh and Yu-Shan Shih

University of Wisconsin–Madison and National Chung Cheng University

Abstract: Classification trees based on exhaustive search algorithms tend to be biased towards selecting variables that afford more splits. As a result, such trees should be interpreted with caution. This article presents an algorithm called QUEST that has negligible bias. Its split selection strategy shares similarities with the FACT method, but it yields binary splits and the final tree can be selected by a direct stopping rule or by pruning. Real and simulated data are used to compare QUEST with the exhaustive search approach. QUEST is shown to be substantially faster and the size and classification accuracy of its trees are typically comparable to those of exhaustive search.

Key words and phrases: Decision trees, discriminant analysis, machine learning.

1. Introduction

A classification tree is a rule for predicting the class of an object from the values of its predictor variables. The tree is constructed by recursively partitioning a learning sample of data in which the class label and the values of the predictor variables for each case are known. Each partition is represented by a node in the tree.

Two approaches to split selection have been proposed in the statistical literature. The first and more popular approach examines all possible binary splits of the data along each predictor variable to select the split that most reduces some measure of node impurity. It is used, for example, by the THAID (Morgan and Sonquist (1963), Morgan and Messenger (1973)) and CART (Breiman, Friedman, Olshen and Stone (1984)) algorithms. If X is an ordered variable, this approach searches over all possible values c for splits of the form

$$X \leq c. \tag{1}$$

A case is sent to the left subnode if the inequality is satisfied and to the right subnode otherwise. The values of c are usually restricted to mid-points between consecutively ordered data values. If X is a categorical predictor (i.e., a predictor variable that takes values in an unordered set), the search is over all splits of the form

$$X \in A, \tag{2}$$

where A is a non-empty subset of the set of values taken by X.

To illustrate, Figure 1 shows a classification tree constructed using the Gini measure of impurity for the Iris data (Fisher (1936)). The data consist of 50 samples from each of three varieties of the flower. Observations on four variables (sepal length and width and petal length and width) are given for each sample. The tree splits first on petal length and then on petal width. Six of the 150 samples are misclassified by the tree, giving an apparent error rate of 4%. A jackknife estimate of the true error of the procedure is 5% ± 2%.

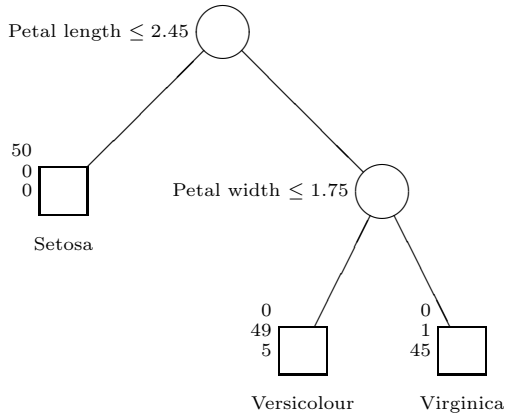


Figure 1. Iris data using exhaustive search with Gini index, 10-fold CV pruning, and 1-SE rule. The triple beside each terminal node gives the number of cases of Setosa, Versicolour and Virginica, respectively, in the node. A jackknife estimate of classification error is 0.053 ± 0.018.

There are two problems with the exhaustive search approach:

1. *Computational complexity.* An ordered variable with n distinct values at a node induces $(n-1)$ splits of the form (1). Therefore the order of computations at each node is linear in the number of distinct data values. In the case of a categorical variable, the order of computations increases exponentially with the number of categories, being $(2^{M-1}-1)$ for a variable with M values.
   More flexible splits may be obtained by combining variables. Ordered variables may be combined in a linear combination split of the form

$$\sum_{k=1}^{K} a_k x_k \le c. \tag{3}$$

   Exhaustively searching over these splits requires the data to be partitioned and the node impurity functions evaluated for every set of values of $\{a_1, \ldots, a_K, c\}$. This is a more difficult task than searching for splits of the form (1), especially since the objective function usually has multiple local maxima. These

computational problems have stimulated much research into approximate or heuristic solutions (Chou (1991), Murthy, Kasif and Salzberg (1994)).

2. *Bias in variable selection.* A more serious problem from the standpoint of tree interpretation is that unrestrained search tends to select variables that have more splits. This makes it hard to draw reliable conclusions from the tree structures. Doyle (1973) seems to be the first to warn of this in the context of the AID and THAID algorithms. More recently, Quinlan and Cameron-Jones (1995) observed that

> "... for any collection of training data, there are 'fluke' theories that fit the data well but have low predictive accuracy. When a very large number of hypotheses is explored, the probability of encountering such a fluke increases."

Numerical evidence of the bias will be given later in Table 2.

The FACT algorithm (Loh and Vanichsetakul (1988), Vanichsetakul (1986)) employs a computationally simpler approach. Instead of combining the problem of variable selection (X) with that of split point selection (c), FACT deals with them separately. At each node, an analysis of variance (ANOVA) F-statistic is calculated for each ordered variable. The variable with the largest F-statistic is selected and linear discriminant analysis (LDA) is applied to it to find c. Categorical variables are handled by transforming them into ordered variables. If there are J classes among the data in a node, this method splits the node into J subnodes.

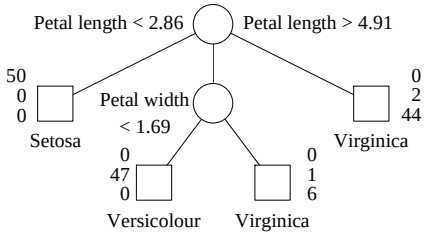
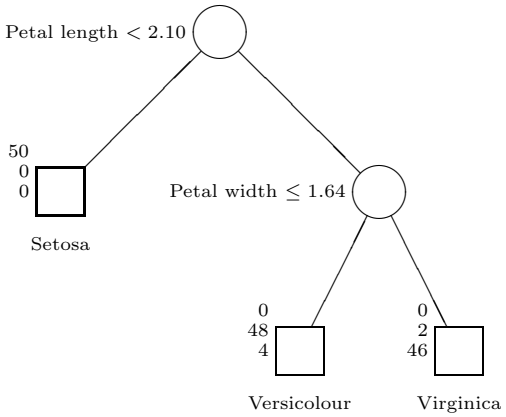


Figure 2. Iris data using the FACT method. The triple beside each terminal node gives the number of cases of Setosa, Versicolour and Virginica, respectively, in the node. A jackknife estimate of classification error is 0.033 ± 0.015.

Figure 2 shows the FACT tree for the Iris data. It uses the same two variables as in the exhaustive search method. The first split yields three nodes because there are three classes. The second split, however, produces only two nodes because only two classes are left in the data. The tree misclassifies three of the learning samples, and has a jackknife estimate of error of $3\% \pm 2\%$.

It turns out that FACT is free of variable selection bias only when all the predictors are ordered variables. If some are categorical variables, it is not unbiased. Furthermore, because it uses a direct stopping rule, it is less effective than a method that employs bottom-up pruning (such as CART) in some applications.



Figreu 3. Iris data using the QUEST method with 10-fold CV pruning and 1-SE rule. The triple beside each terminal node gives the number of cases of Setosa, Versicolour and Virginica, respectively, in the node. A jackknife estimate of classification error is 0.040 ± 0.016.

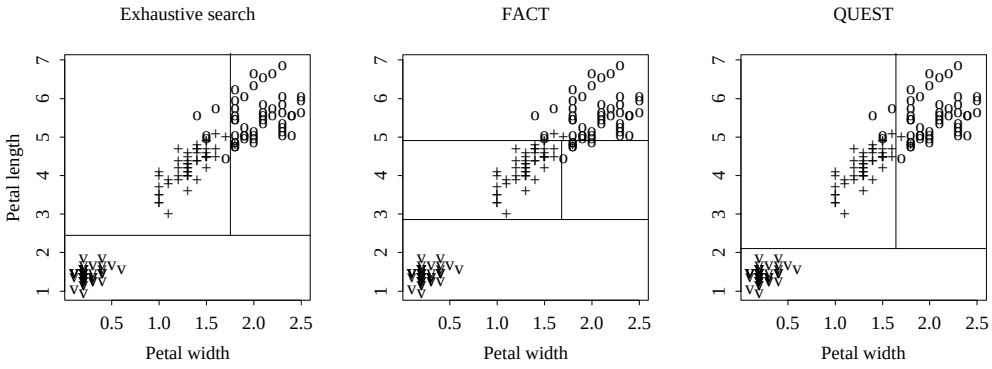


Figure 4. Plots of Iris data. Plot symbols refer to Setosa (v), Versicolour (+), and Virginica (o). The partitions in the plots correspond to the trees in Figures 1, 2, and 3.

The purpose of this paper is to present a new algorithm, called QUEST (for Quick, Unbiased, Efficient, Statistical Tree), that (i) has negligible variable selection bias, (ii) retains the computational simplicity of FACT, (iii) includes pruning as an option, and (iv) yields binary splits. The reason for binary splits

is so that the QUEST trees may be easily compared with exhaustive search trees in terms of stability of the splits and number of nodes.

Figure 3 shows the QUEST tree for the Iris data. The apparent error rate is higher than that for the FACT tree, but it is the same as that for the exhaustive search method. Its jackknife error estimate of $4\% \pm 2\%$ is in between those of the other two trees, although the differences are not statistically significant. Figure 4 shows how the data are partitioned by the three methods in the space of the two partitioning variables.

The remainder of the paper is organized as follows. Section 2 presents our method for split point selection, assuming that the variable has been chosen. Section 3 explains why the variable selection procedure in FACT is biased and how the bias is removed in QUEST. Section 4 extends the technique to linear combination splits. Section 5 considers the relative stability of the split points and Section 6 does the same for relative computational speed. The entire algorithms (including pruning) are compared in Section 7 by means of two datasets. The QUEST method is demonstrated to be much better than exhaustive search in terms of variable selection bias and computational cost. Neither approach dominates the other in terms of classification accuracy, stability of split points, or size of tree.

The following notations are used in the sequel. The dimension of the predictor space is denoted by K and the number of classes in the learning sample is denoted by J. The number of class j cases in the learning sample is denoted by N_j and the total sample size by $N = \sum_{j=1}^{J} N_j$. Given the data in a node t, the number of classes present is denoted by J_t ($J_t \leq J$), the number of class j cases by $N_j(t)$, and the total sample size by $N(t)$. The kth predictor variable and its observed value are denoted by X_k and x_k, respectively. We use $\pi(j)$ to denote the prior probability for class j. This may be specified by the user or estimated by the sample proportion N_j/N. We let $p(j,t) = \pi(j)N_j(t)/N_j$ denote the estimated probability that a class j object will fall into the partition represented by node t, and let $p(j|t) = p(j,t)/\sum_i p(i,t)$ denote the estimated posterior probability that an object belongs to class j given that it lands in t. The Gini index $i(t)$ of impurity at t is defined as $i(t) = 1 - \sum_j p^2(j|t)$.

2. Split Point Selection

2.1. Ordered variable

Suppose that an ordered variable X is selected to split a node. FACT employs LDA on this variable to construct the split. This has two disadvantages. The first is that the node is split into as many subnodes as there are classes. If J is large, this may deplete the learning sample so rapidly that the tree is too short to

reveal interesting features in the data. A second disadvantage is that the effect of unequal class variances is ignored.

These two problems are solved in the QUEST method as follows. To ensure binary splits when $J > 2$, we group the classes into two superclasses before application of discriminant analysis. However, to accommodate unequal variances, we use a modified form of quadratic discriminant analysis (QDA) on the two superclasses.

First suppose that there are only two classes. Traditional QDA estimates the class density functions with normal densities where the means and variances are estimated from the sample. Specifically, let $\bar{x}^{(j)}$ and s_j^2 denote the sample class mean and variance for the jth class ($j = 1, 2$). Let $\phi(x) = (2\pi)^{-1/2} \exp(-x^2/2)$ denote the standard normal density function. QDA splits the X-axis into three intervals $(-\infty, d_1)$, (d_1, d_2), and (d_2, ∞), where d_1 and d_2 are the roots of the equation

$$p(1|t)s_1^{-1}\phi\{(x - \bar{x}^{(1)})/s_1\} = p(2|t)s_2^{-1}\phi\{(x - \bar{x}^{(2)})/s_2\}. \tag{4}$$

In order to obtain a binary split, QUEST uses only one of the two roots as split point: the one that is closer to the sample mean of each class. Figure 5 compares the split points from the QDA method with those from the exhaustive search method. In each plot, the QUEST split point is marked with a dotted line and the exhaustive search split point with a dashed line. The data are the expected order statistics from four pairs of distributions that are indicated by the solid density curves. $N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2, T2(μ) denotes a t-distribution with 2 degrees of freedom centered at μ, Chisq(ν) denotes a chi-square distribution with ν degrees of freedom, Beta(p, q) denotes a beta distribution with parameters p and q, and $U(0, 1)$ denotes a uniform distribution on the unit interval. In the case of the beta-uniform pair, the splits are not unique for either method. The 'ideal' split is where the density curves intersect. By this criterion, QUEST is better than exhaustive search for two pairs of distributions and worse for the other two pairs.

For $J > 2$, a preliminary grouping of the classes into two superclasses is needed. This is carried out by applying a 2-means clustering algorithm (that minimizes the within-cluster sum of squares) to the J sample class X-means. If the class means are identical, the class with the most number of cases becomes superclass A and the other classes form superclass B. If there are two or more classes with the same maximum number of cases, the one with the smallest index among them is chosen to form A. The procedure may be stated as follows.

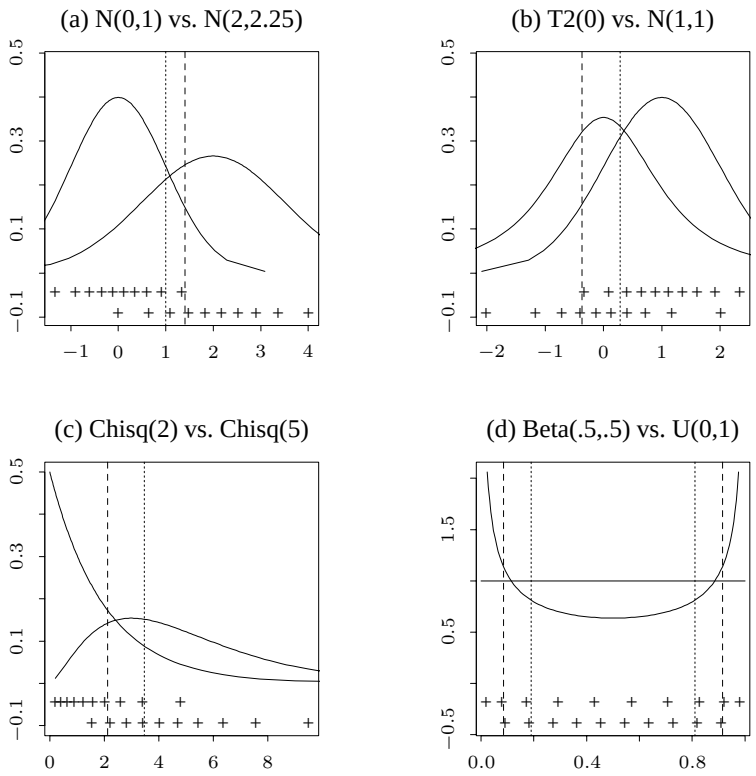


Figure 5. Comparison of split methods with different class populations. Data for each class indicated by '+' signs are 10 expected order statistics. True class densities are drawn with solid lines. QUEST splits are shown by dotted vertical lines and those for exhaustive search by dashed vertical lines. The 'ideal' split in each case is where the density curves intersect.

Algorithm 1. Split selection for an ordered variable

Let X be the selected variable to split node t.

1. *Apply the* 2*-means clustering algorithm of Hartigan and Wong* (1979) *to divide the J_t classes into two superclasses A and B, using the two most extreme sample means as initial cluster centers. If the sample means are identical, let A contain the most populous class and B contain the other classes.*
2. *Let $\bar{x}_A$ and s_A^2 denote the sample mean and variance of superclass A. Similarly, let $\bar{x}_B$ and s_B^2 denote the corresponding quantities for B. Let $p(A|t) = \sum_{j\in A} p(j|t)$ and $p(B|t) = 1 - p(A|t)$ denote the superclass priors.*
3. *Take logs on both sides of the equation*

$$p(A|t)s_A^{-1}\phi\{(x-\bar{x}_A)/s_A\} = p(B|t)s_B^{-1}\phi\{(x-\bar{x}_B)/s_B\}$$

to obtain the quadratic equation $ax^2 + bx + c = 0$, *where*

$$\begin{aligned} a &= s_A^2 - s_B^2 \\ b &= 2(\bar{x}_A s_B^2 - \bar{x}_B s_A^2) \\ c &= (\bar{x}_B s_A)^2 - (\bar{x}_A s_B)^2 + 2s_A^2 s_B^2 \log[\{p(A|t)s_B\}/\{p(B|t)s_A\}]. \end{aligned}$$

If $a = 0$ *and* $\bar{x}_A \neq \bar{x}_B$, *there is only one root given by*

$$x = (\bar{x}_A + \bar{x}_B)/2 - (\bar{x}_A - \bar{x}_B)^{-1} s_A^2 \log\{p(A|t)/p(B|t)\}. \tag{5}$$

The equation has no roots if $a = 0$ *and* $\bar{x}_A = \bar{x}_B$.

4. *The node is split at* $X = d$ *where* d *is defined as follows:*
   (a) *If* $a = 0$ *then*

$$d = \begin{cases} (\bar{x}_A + \bar{x}_B)/2 - (\bar{x}_A - \bar{x}_B)^{-1} s_A^2 \log \frac{p(A|t)}{p(B|t)}, & \bar{x}_A \neq \bar{x}_B, \\ \bar{x}_A, & \bar{x}_A = \bar{x}_B. \end{cases}$$

   (b) *Else if* $a \neq 0$, *then:*
      (i) *If* $b^2 - 4ac < 0$, *define* $d = (\bar{x}_A + \bar{x}_B)/2$. *It can be verified that* $b^2 - 4ac \geq 0$ *if* $p(A|t) = p(B|t)$.
      (ii) *Else if* $b^2 - 4ac \geq 0$, *then:*
         A. *Define* d *to be the root* $(2a)^{-1}\{-b \pm \sqrt{b^2 - 4ac}\}$ *that is closer to* $\bar{x}_A$, *provided this yields two nonempty nodes.*
         B. *Otherwise, define* $d = (\bar{x}_A + \bar{x}_B)/2$.

2.2. Categorical variable

FACT uses two steps to transform a categorical variable into an ordered one: (i) the sample values taken by the categorical variable are mapped into 0-1 dummy vectors, and (ii) the dummy vectors are projected onto their largest discriminant coordinate (called CRIMCOORD for short (see Gnanadesikan (1977))). The aim is to use the discriminatory information in the categorical variable to define the spacing and ordering of the transformed values.

A difficulty occurs when the within-class covariance matrix in the space of dummy vectors is singular. This situation occurs frequently because of the discrete nature of the dummy vectors. For example, if some categorical values have been been diverted to other nodes, they will not be present in the data in the current node. Therefore some of the components of the sample dummy vectors at the node will be identically zero and a singular sample covariance matrix results. When this occurs, FACT employs some other methods to map the dummy vectors to real numbers. We cannot use this solution here because it can split the node into more than two subnodes.

A more consistent and elegant way to deal with singular matrices that does not split a node into more than two subnodes is to reduce the dimension of the dummy vectors before computation of the CRIMCOORDs. Suppose X is a categorical variable taking values in the set $\{c_1, \ldots, c_M\}$. Each value of X is first transformed into an M-dimensional 0-1 column vector $\boldsymbol{v} = (v_1, \ldots, v_M)'$ all of whose components are 0 except for the lth component, which is equal to 1, where l is defined implicitly through $X = c_l$. Let $\boldsymbol{v}_i^{(j)}$ denote the ith observed value of $\boldsymbol{v}$ in the jth class and define the M-dimensional column vectors

$$\bar{\boldsymbol{v}}^{(j)} = N_j^{-1} \sum_{i=1}^{N_j} \boldsymbol{v}_i^{(j)}, \quad \bar{\boldsymbol{v}} = N^{-1} \sum_{j=1}^{J} \sum_{i=1}^{N_j} \boldsymbol{v}_i^{(j)}.$$

Define the $M \times M$ matrices

$$\boldsymbol{B} = \sum_{j=1}^{J} N_j (\bar{\boldsymbol{v}}^{(j)} - \bar{\boldsymbol{v}})(\bar{\boldsymbol{v}}^{(j)} - \bar{\boldsymbol{v}})' \tag{6}$$

$$\boldsymbol{W} = \sum_{j=1}^{J} \sum_{i=1}^{N_j} (\boldsymbol{v}_i^{(j)} - \bar{\boldsymbol{v}}^{(j)})(\boldsymbol{v}_i^{(j)} - \bar{\boldsymbol{v}}^{(j)})'$$

$$\boldsymbol{T} = \sum_{j=1}^{J} \sum_{i=1}^{N_j} (\boldsymbol{v}_i^{(j)} - \bar{\boldsymbol{v}})(\boldsymbol{v}_i^{(j)} - \bar{\boldsymbol{v}})' \tag{7}$$

so that $\boldsymbol{T} = \boldsymbol{B} + \boldsymbol{W}$.

The largest CRIMCOORD is the projection $\boldsymbol{a}'\boldsymbol{v}$ that maximizes the ratio of between-classes to within-classes sum-of-squares $\boldsymbol{a}'\boldsymbol{B}\boldsymbol{a}/\boldsymbol{a}'\boldsymbol{W}\boldsymbol{a}$. The value of $\boldsymbol{a}$ is given, up to a scalar multiple, by the eigenvector corresponding to the largest eigenvalue of $\boldsymbol{W}^{-1}\boldsymbol{B}$, when the inverse $\boldsymbol{W}^{-1}$ exists (Mardia, Kent and Bibby (1979), p. 319), i.e., it is the solution to the matrix equation $(\boldsymbol{B} - \lambda \boldsymbol{W})\boldsymbol{a} = \boldsymbol{0}$. Since the solution of the latter equation is the same as that for the equation $(\boldsymbol{B} - \lambda \boldsymbol{T})\boldsymbol{a} = \boldsymbol{0}$ when $\boldsymbol{W}^{-1}$ exists, the solution $\boldsymbol{a}$ is also the eigenvector associated with the largest eigenvalue of $\boldsymbol{T}^{-1}\boldsymbol{B}$. The precise algorithm is as follows.

Algorithm 2. Split selection for a categorical variable

Suppose X is a categorical variable taking values in the set $\{c_1, \ldots, c_M\}$.

1. *Transform each value of X into an M-dimensional dummy column vector $\boldsymbol{v} = (v_1, \ldots, v_M)'$, where*

$$v_l = \begin{cases} 1, \textit{ if } X = c_l, \\ 0, \textit{ otherwise.} \end{cases}$$

Let $\boldsymbol{V}$ be the $N \times M$ data matrix consisting of the v-values.

2. *Let* $\boldsymbol{I}$ *denote the* $N \times N$ *identity matrix and let* $\mathbf{1}$ *be an* N*-column of* 1*'s. Let* $\boldsymbol{H} = \boldsymbol{I} - N^{-1}\mathbf{1}\mathbf{1}'$ *denote the centering matrix and obtain the singular value decomposition* $\boldsymbol{HV} = \boldsymbol{PDQ}'$ *with* $\boldsymbol{D} = diag(d_1, \ldots, d_M)$ *such that* $d_1 \geq \ldots \geq d_M \geq 0$.
3. *Let* ε *be the machine precision, i.e., the smallest floating point number such that if* $u = 1 + \varepsilon$, *then* $u > 1$. *Define an eigenvalue* d_m *as 'positive' if it satisfies* $d_m > \max(M, N)d_1\varepsilon$, *and as 'zero' otherwise* (*Math Works* (1991)). *The rank* r *of* $\boldsymbol{T}$ *is defined to be the number of 'positive' eigenvalues. Let* $\boldsymbol{F}$ *denote the* $M \times r$ *submatrix of* $\boldsymbol{Q}$ *consisting of its first* r *columns and let* $\boldsymbol{U} = diag(d_1^{-1}, \ldots, d_r^{-1})$.
4. *Reduce the dimension of* $\boldsymbol{v}$ *by transforming it to* $\boldsymbol{y} = \boldsymbol{F}'\boldsymbol{v}$.
5. *Define for each* j *the* $M \times N_j$ *matrix* $\boldsymbol{L}_j = (\bar{\boldsymbol{v}}^{(j)} - \bar{\boldsymbol{v}}, \ldots, \bar{\boldsymbol{v}}^{(j)} - \bar{\boldsymbol{v}})$ *and let* $\boldsymbol{G}$ *be the* $N \times M$ *matrix* $\boldsymbol{G} = (\boldsymbol{L}_1, \ldots, \boldsymbol{L}_J)'$, *so that* $\boldsymbol{B} = \boldsymbol{G}'\boldsymbol{G}$. *Perform a singular value decomposition of the matrix* $\boldsymbol{GFU}$ *and let* $\boldsymbol{a}$ *be the eigenvector associated with the largest eigenvalue.*
6. *Transform each* $\boldsymbol{v}$ *to* $\xi = \boldsymbol{a}'\mathbf{U}\mathbf{y} = \boldsymbol{a}'\boldsymbol{U}F'\boldsymbol{v}$. *This maps each* c_l *to a* ξ*-value.*
7. *Apply Algorithm* 1 *to the* ξ *data values to split the node.*
8. *Re-express a split of the form '*$\xi \leq \xi_0$*' to the form '*$X \in A$*'.*

Note that each categorical value is mapped into a CRIMCOORD value even if it is not represented in the learning sample at the node. As a result, there is no difficulty for the selected split to handle categorical values that appear in future test samples but that are absent from the learning sample.

Table 1. Four examples of CRIMCOORD transformations. The CRIMCOORD values are scaled so that the maximum value is 1 and the minimum value is -1 in each example.

Data Set	$J = 2, M = 3, N_1 = N_2 = 10$	$\{4c_1, c_2, 5c_3\}, \{2c_1, 2c_2, 6c_3\}$
I	CRIMCOORD values	$\xi_1 = 1, \xi_2 = -1, \xi_3 = -0.273$
Data Set	$J = 2, M = 3, N_1 = N_2 = 10$	$\{5c_1, 5c_3\}, \{5c_1, 5c_3\}$
II	CRIMCOORD values	$\xi_1 = -1, \xi_2 = 0, \xi_3 = 1$
Data Set	$J = 2, M = 3, N_1 = 10, N_2 = 11$	$\{5c_1, 5c_3\}, \{5c_1, c_2, 5c_3\}$
III	CRIMCOORD values	$\xi_1 = \xi_3 = 1, \xi_2 = -1$
Data Set	$J = 3, M = 5,$	$\{5c_1, 5c_2\}, \{c_1, 5c_2, 3c_4, c_5\},$
	$N_1 = N_2 = N_3 = 10$	$\{c_1, 4c_2, 5c_3\}$
IV	CRIMCOORD values	$\xi_1 = -0.245, \xi_2 = -0.194,$
		$\xi_3 = 1, \xi_4 = \xi_5 = -1$

To illustrate, Table 1 shows four datasets and their CRIMCOORD transformations. Because eigenvectors are only defined by their directions and not their lengths, the CRIMCOORD values in the table are scaled so that the minimum

and maximum values in each example are -1 and 1, respectively. For example, there are two classes ($J = 2$) in Data Set I and the observations in its first class consist of 4 c_1's, 1 c_2 and 5 c_3's. In each dataset, c_i is transformed to ξ_i. Data Sets II–IV are chosen to demonstrate how our method handles different contingencies. In Data Set II, the two classes contain identical samples and one category (c_2) is absent from both samples. Since the categorical values contain no information about the classes, there is no right or wrong mapping in this case. Data Set III is obtained from Data Set II by adding the absent categorical value of c_2 to the second class sample. Now c_2 carries much information about the classes whereas c_1 and c_3 are still uninformative. The CRIMCOORD mapping $\xi_1 = \xi_3 = 1$ and $\xi_2 = -1$ is thus reasonable. Finally, in Data Set IV, two of the categorical values (c_4 and c_5) appear only in one class. These two values are indistinguishable between themselves but are highly informative for the second class. The CRIMCOORD transformation reflects this by mapping them to the same extreme value.

Some information is clearly lost in the projection step of the CRIMCOORD transformation. However, the other alternative of replacing each categorical variable with its dummy vector and letting these vectors compete with the ordered variables for splits has two undesirable effects. First, variable selection may be biased towards categorical variables because their dimensions are increased. Second, the class of splits on a categorical variable are restricted to splits on the dummy vector components. Since the vector components take values 0 or 1, the splits have the form (2) with the sets A being singletons or their complements.

3. Variable Selection

It is assumed in the preceding sections that a variable has been selected to split a node. We now explain how this is carried out in the FACT and QUEST methods. FACT uses statistical tests to choose the variable. First, if there are categorical variables, each is converted into a CRIMCOORD variable. Next, a threshold value F_0 is chosen and an ANOVA F-statistic is computed for every variable. If the largest F-statistic exceeds F_0, the variable with the largest F-value is selected to split the node. Otherwise, Levene's (1960) F-statistic for unequal variances is computed for each variable. If the largest Levene F-statistic is greater than F_0, the variable with the largest Levene F-value is used to split the node. Otherwise, if neither the largest ANOVA F-value nor the largest Levene F-value exceeds F_0, the node is split using the variable with the largest ANOVA F-value.

The reason for the inclusion of tests for variances is to avoid inefficient splits. Suppose, for example, that there are two ordered predictor variables and two classes such that along X_1, the class distributions are normal with equal means

but unequal variances. Suppose further that the distribution of X_2 is the same for both classes. In this case, splits on X_2 should be avoided since the partitions they induce merely convert the original problem into two smaller but similar problems.

When all the variables are ordered, the FACT procedure is unbiased in the sense that each variable has an equal chance of being selected if they are independent and have no relationship with the class variable. On the other hand, when there are categorical variables, the ANOVA F-statistics computed from the CRIMCOORD variables tend to be stochastically larger than those for the ordered variables. Thus, even when all the variables are independent of each other and of the class variable, categorical variables are more likely to be chosen than ordered variables.

In order to remove this bias, it is necessary to use another method to rank categorical variables for split selection. We use the Pearson contingency table χ^2-test of independence between the class variable and the categorical variable. The test is easy to compute and its statistical significance can be approximated via the chi-square distribution with $(J_t - 1)(M_t - 1)$ degrees of freedom, where M_t is the number of distinct categories present in the learning sample in node t.

We therefore obtain a P-value from each variable based on the appropriate χ^2 or F-test. Call this Stage I. If the smallest P-value is less than a predefined threshold (determined via the Bonferroni method for multiple comparisons), the corresponding variable is selected. Otherwise, Levene's F-test for unequal variances is computed for each ordered variable. Call this Stage II. If the smallest P-value from the Stage II tests is less than another Bonferroni threshold, the corresponding variable is selected. Otherwise, the variable with the smallest P-value from Stage I is selected. Like FACT, this procedure is not exactly unbiased when categorical variables are present. However, the Bonferroni correction ensures that the bias is practically negligible. The detailed algorithm follows.

Algorithm 3. Variable selection

Let $\alpha \in (0, 1)$ be a pre-specified level of significance. Assume that $X_1, \ldots, X_{K_1}$ are ordered variables and $X_{K_1+1}, \ldots, X_K$ are categorical variables. Given node t, let $x_{ik}^{(j)}$ denote the value of the kth variable for the ith case in the jth class ($i=1, \ldots, N_j(t)$; $j = 1, \ldots, J_t$; $k = 1, \ldots, K_1$).

1. *If $K_1 \geq 1$, compute the ANOVA F-statistic F_k for each X_k, $k = 1, \ldots, K_1$. Let k_1 be the smallest integer such that $F_{k_1} = \max\{F_k : k = 1, \ldots, K_1\}$ and define $\hat{\alpha}_1 = \Pr\{F_{J_t-1,N(t)-J_t} > F_{k_1}\}$, where F_{ν_1,ν_2} denotes the F-distribution with ν_1 and ν_2 degrees of freedom.*
2. *If $K > K_1$, compute the P-value $\hat{\beta}(k)$ of the contingency table chi-square test of independence between class labels and category values for $k = K_1+1, \ldots, K$. The degrees of freedom in each case are given by $(n_r-1)\times(n_c-1)$, where n_r and*

n_c are the numbers of rows and columns of the table with nonzero totals. Let k_2 be the smallest integer such that $\hat{\beta}(k_2) = \min\{\hat{\beta}(k) : k = K_1 + 1, \ldots, K\}$ and define $\hat{\alpha}_2 = \hat{\beta}(k_2)$.

3. *Define $k' = k_1$ if $\hat{\alpha}_1 \leq \hat{\alpha}_2$; otherwise define $k' = k_2$.*
4. *If $\min(\hat{\alpha}_1, \hat{\alpha}_2) < \alpha/K$, select variable $X_{k'}$ to split the node.*
5. *Otherwise, if $\min(\hat{\alpha}_1, \hat{\alpha}_2) \geq \alpha/K$, then*
   (a) *Compute the ANOVA F-statistics $F_k^{(z)}$ $(k = 1, \ldots, K_1)$ for the ordered variables based on the absolute deviations $z_{ik}^{(j)} = |x_{ik}^{(j)} - \bar{x}_k^{(j)}|$, where $\bar{x}_k^{(j)} = N_j(t)^{-1} \sum_{i=1}^{N_j(t)} x_{ik}^{(j)}$. Let k'' be the smallest integer such that $F_{k''}^{(z)} = \max\{F_k^{(z)} : k = 1, \ldots, K_1\}$.*
   (b) *Compute $\tilde{\alpha} = \Pr\{F_{J_t-1,N(t)-J_t} > F_{k''}^{(z)}\}$. If $\tilde{\alpha} < \alpha/(K+K_1)$, select variable $X_{k''}$ to split the node. Otherwise, select variable $X_{k'}$.*

A simulation experiment was carried out to compare the effect of several factors on the variable selection bias of the QUEST, FACT and exhaustive search methods. The factors are: variable type and quantity, number of distinct data values, and number of classes. The simulations are carried out with $\alpha = 0.05$ (in Algorithm 3) for the QUEST method and $F_0 = 4$ for the FACT method. The estimated probabilities are based on 10,000 simulation trials. Therefore the simulation standard errors are less than 0.005. The predictor variables are all mutually independent.

We first study the simple null case, where there are only two predictor variables both independent of the class variable. An unbiased variable selection procedure should select each variable with probability 0.5. The results are given in Table 2, where U_m denotes a uniform distribution on the integers $\{1, 2, \ldots, m\}$.

The QUEST and FACT methods are almost unbiased when the variables are ordered. As expected, the FACT method shows a large bias towards categorical variables. When both variables are categorical, FACT tends to select the one with more categories.

The exhaustive search method is also biased towards categorical variables, although its bias is not as large as that of FACT. However, unlike FACT, the bias is not limited to the presence of categorical variables. When one variable is continuous (e.g., normal, t_2, or exponential) and the other is discrete with few distinct values (U_4), exhaustive search shows a large bias towards the continuous variables because they afford more splits.

If all the predictor variables are uninformative for predicting the class variable, an effective pruning procedure would likely remove all the nodes except the root node. In this case, the bias of the exhaustive search method towards variables with more splits does not adversely affect the final result. The danger occurs when a data set consists of a mix of informative and noise variables, and

the noise variables have many more splits than the informative variables. Then there is a high probability that the noise variables will be chosen to split the top nodes of the tree. Pruning will produce either a tree with misleading structure or no tree at all.

Table 2. Probabilities of correct variable selection when the predictor variables have no discriminatory power. A method is unbiased if it selects X_1 with a constant probability of 0.5. Estimates are based on 10,000 replications and learning class sample sizes of 200 and 100 for $J = 2$ and $J = 5$, respectively. Simulation standard errors are less than 0.005. "Cat. U_n" denotes a categorical distribution uniformly distributed on the integers 1, ..., n, and "Ord. U_n" denotes an ordered uniform distribution on the same integers.

Number of classes J	Learning sample size N	Distribution of X_1	Distribution of X_2	$P(X_1$ is selected) QUEST method	FACT method	Exhaustive search
2	400	Normal	t_2	0.492	0.507	0.507
		Normal	Exponential	0.501	0.514	0.508
		t_2	Exponential	0.515	0.526	0.503
		Normal	Ord. U_4	0.508	0.522	0.895
		t_2	Ord. U_4	0.520	0.533	0.897
		Exponential	Ord. U_4	0.500	0.512	0.894
		Cat. U_4	Cat. U_{15}	0.497	0.017	0.036
		Cat. U_4	Normal	0.501	0.831	0.176
		Cat. U_{15}	Normal	0.497	0.997	0.869
		Cat. U_4	Exponential	0.493	0.821	0.180
		Cat. U_{15}	Exponential	0.502	0.998	0.874
		Cat. U_4	Ord. U_4	0.492	0.832	0.610
		Cat. U_{15}	Ord. U_4	0.501	0.998	0.980
5	500	Normal	t_2	0.488	0.483	0.501
		Normal	Exponential	0.510	0.506	0.498
		t_2	Exponential	0.518	0.514	0.504
		Normal	Ord. U_4	0.526	0.522	0.898
		t_2	Ord. U_4	0.520	0.515	0.898
		Exponential	Ord. U_4	0.519	0.509	0.897
		Cat. U_4	Cat. U_{15}	0.504	0.011	0.020
		Cat. U_4	Normal	0.497	0.859	0.179
		Cat. U_{15}	Normal	0.488	0.999	0.909
		Cat. U_4	Exponential	0.476	0.855	0.173
		Cat. U_{15}	Exponential	0.475	1.000	0.905
		Cat. U_4	Ord. U_4	0.495	0.861	0.638
		Cat. U_{15}	Ord. U_4	0.486	1.000	0.990

To see how noise variables can prevent an informative variable from being

selected, 20 variables were employed in another simulation experiment. One variable has discriminatory power while the other 19 are noise. Different combinations of variable types were included (e.g., all ordered variables, all categorical variables, and 1 ordered and 19 categorical variables). Two classes were used with 100 learning samples in each class. Table 3 shows the probabilities that the informative variable is selected. The class distributions for the informative variable X_1 are given in the first 2 columns of the Table. The notation $E(0, b)$ denotes an exponential distribution with density $b^{-1}\exp(-x/b)$, $x > 0$. A_4 denotes the discrete distribution on the integers $\{1, 2, 3, 4\}$ with probabilities $P(X = 1) = P(X = 2) = P(X = 3) = 2/9$, $P(X = 4) = 1/3$ and B_4 the discrete distribution with $P(X = 1) = P(X = 2) = P(X = 3) = 1/5$, $P(X = 4) = 2/5$. The results show that in every case, the probability that the exhaustive search method selects the informative variable is less than 1/20, the probability of random selection. Again, FACT and QUEST are equally good when all the variables are ordered. However, when categorical variables are present, the FACT method performs even worse than exhaustive search.

Table 3. Probabilities of correct variable selection when X_1 is informative and $X_2, \ldots, X_{20}$ are noise. 10,000 replications, 2 classes, and learning sample sizes of 100 for each class.

Distribution of X_1		Noise distribution	$P(X_1$ is selected)		
Class 1	Class 2	of $X_2, \ldots, X_{20}$	QUEST method	FACT method	Exhaustive search
Ord. U_4	Ord. A_4	$N(0,1)$	0.157	0.153	0.046
Ord. U_4	Ord. A_4	t_2	0.177	0.178	0.044
Ord. U_4	Ord. A_4	$E(0,1)$	0.150	0.148	0.048
Cat. U_4	Cat. B_4	Cat. U_{15}	0.404	0.008	0.031
$N(0,1)$	$N(0.25,1)$	Cat. U_{15}	0.380	0.001	0.029
$E(0,1)$	$E(0,1.3)$	Cat. U_{15}	0.430	0.001	0.033
Ord. U_4	Ord. B_4	Cat. U_{15}	0.412	0.001	0.029

4. Linear Combination Splits

The FACT approach towards linear combination splits is essentially recursive LDA on all the variables, where categorical variables are first transformed to their CRIMCOORDS. The QUEST version is similar, except that a prior grouping of the classes into two superclasses is carried out if $J > 2$. The steps may be stated as follows.

1. Transform each categorical variable into a CRIMCOORD variable.
2. Apply Steps 2–6 of Algorithm 2, with the ordered and transformed categorical variables in place of $\boldsymbol{v}$, and K in place of M, to find the linear projection values ξ.

3. Apply Algorithm 1 to the ξ-values to find the split.

5. Variability of Split Points

It is recently reported that some classification methods are unstable in the sense that small changes in the learning sample may lead to large changes in the classifiers. As a result, these classifiers tend to vary substantially in their classification accuracy. Breiman (1996a) noted that CART and neural networks are in this category, while LDA and nearest neighbor methods are stable. Breiman (1996b) pointed out that LDA achieves its low variability by having a limited set of models to fit the data. When this set is inadequate, LDA can perform poorly.

Because QUEST employs a form of QDA to recursively generate a tree structure, it may be viewed as a hybrid between LDA and CART. Thus it is natural to enquire about its stability too. We will study this problem by looking at the variability of the split points in this section.

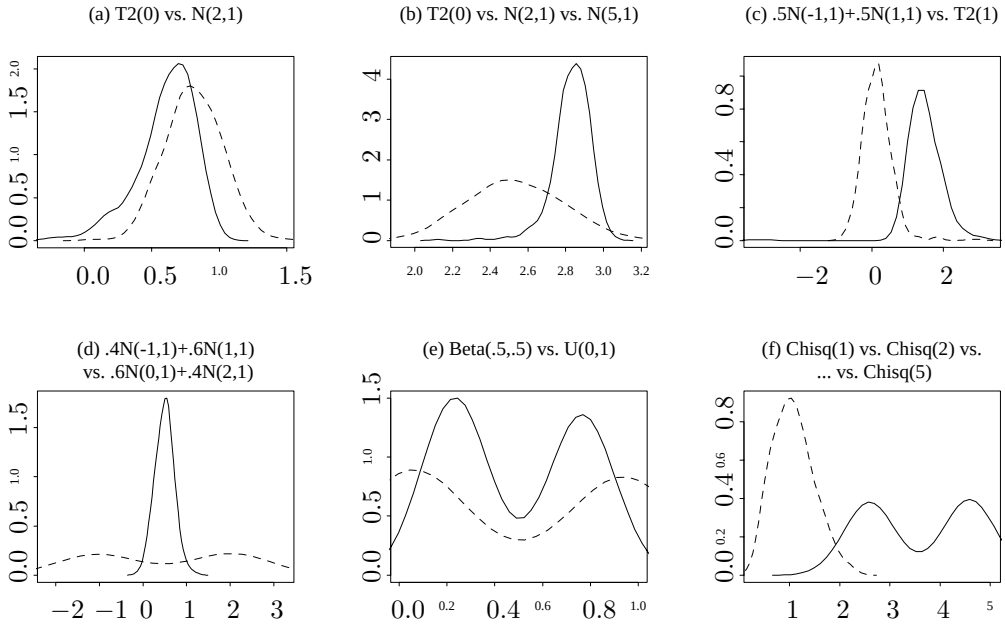


Figure 6. Distributions of split points for various class populations based on 2–5 classes, sample size 200 for each class and 500 replicates. Solid curve is for QUEST; dashed curve for exhaustive search.

Figure 6 shows the simulated distributions of the split points obtained via QUEST (solid lines) and exhaustive search (dashed lines) for one variable having different class distributions. The class sample size used in each plot is 200, and

500 simulation trials were employed to construct the density estimates. Following is a brief description of each plot.

Plot (a). One class is $N(2,1)$ and the other class has a t_2 distribution. The curves show that the two methods have similar variability. This is interesting because the QUEST method has to estimate the infinite variance of the t_2 distribution.

Plot (b). A third class with $N(5,1)$ distribution is added to the two classes in the previous case. Now the QUEST method has distinctly lower variability.

Plot (c). One class is bimodal, being a half-half mixture of $N(-1,1)$ and $N(1,1)$. The other class is a t_2 distribution centered at 1. The variability of the two methods seem to be roughly the same.

Plot (d). Again there are two classes, but both are bimodal normal mixtures. One class is $0.4N(-1,1) + 0.6N(1,1)$ and the other is $0.6N(0,1) + 0.4N(2,1)$. QUEST is clearly less variable than exhaustive search.

Plot (e). One class has a U-shaped beta distribution and the other class is uniform on the unit interval. As observed previously in Figure 5(d), there are two 'ideal' split points. This explains the bimodal nature of the two split point distributions. QUEST is again less variable.

Plot (f). There are five classes in this example, having chi-square distributions with degrees of freedom $1, \ldots, 5$. The exhaustive search method is clearly less variable. Recall that QUEST forms superclasses according to the spacing of the sample class means. When the population class means are equally spaced as here, it is as likely for the leftmost two class means to be grouped to form one superclass as it is for the leftmost three. This accounts for the bimodal shape of the solid curve in the plot.

These simulation results suggest that the variability of the QUEST split point is less than or equal to that of exhaustive search, except when the class means are symmetrically located. To throw more light on the situation, we now approach the question from a large-sample perspective. We only consider two classes because the problem is too unwieldy otherwise.

Let the class distribution and density functions be denoted by $F_j(x)$ and $f_j(x)$, $j = 1, 2$. Suppose that the densities are continuous functions of x and that there is a unique value x_0 such that $f_1(x_0) = f_2(x_0)$. Then assuming equal priors, the split $x = x_0$ is 'ideal' because it minimizes the misclassification rate.

5.1. Asymptotics for exhaustive search

The Gini measure of node impurity before splitting is 1/4. For any split at x into two subnodes, the impurity in one subnode is $i_1(x) = p_1(1 - p_1)$, where $p_1 = F_1(x)/[F_1(x) + F_2(x)]$, and that in the other subnode is $i_2(x) = p_2(1 - p_2)$, where $p_2 = [1 - F_1(x)]/[2 - F_1(x) - F_2(x)]$. Let $\bar{F}(x) = [F_1(x) + F_2(x)]/2$ and

$\bar{f}(x) = [f_1(x) + f_2(x)]/2$. Then the decrease in impurity obtained by splitting at x is given by

$$\Delta(x) = 1/4 - \bar{F}(x)i_1(x) - [1 - \bar{F}(x)]i_2(x)$$

which simplifies to

$$\Delta(x) = \frac{[F_1(x) - F_2(x)]^2}{16\bar{F}(x)[1 - \bar{F}(x)]}. \tag{8}$$

Taking the derivative with respect to x and equating the result to 0 shows that $\Delta'(x) = 0$ if and only if

$$2\bar{F}(x)[1 - \bar{F}(x)][f_1(x) - f_2(x)] = \bar{f}(x)[F_1(x) - F_2(x)][1 - 2\bar{F}(x)]. \tag{9}$$

Suppose that $\Delta(x)$ has a unique maximum at x_e. Let $\hat{F}_j$ be the empirical distribution based on a sample of size n from F_j, $j = 1, 2$. Let $\Delta_n(x)$ be defined as in (8) with $\hat{F}_j$ in place of F_j. That is, $\Delta_n(x)$ is the decrease in impurity based on a sample of size n. Suppose that

$$\Delta(x_e) > \limsup_{|x|\to\infty} \Delta(x). \tag{10}$$

Then $\lim_{n\to\infty} \sup_x |\Delta_n(x) - \Delta(x)| = 0$ a.s. by the Glivenko-Cantelli lemma. Let $\epsilon > 0$ and $\hat{x}_e$ be the maximizing value of $\Delta_n(x)$. With probability 1,

$$-\epsilon < \Delta_n(x) - \Delta(x) < \epsilon \text{ for all large } n \text{ and all } x.$$

By definition of $\hat{x}_e$ and x_e, we have $\Delta_n(x_e) \le \Delta_n(\hat{x}_e)$ and $\Delta(\hat{x}_e) \le \Delta(x_e)$. Therefore with probability 1,

$$\begin{aligned} 0 &\le \Delta(x_e) - \Delta(\hat{x}_e) \\ &= [\Delta(x_e) - \Delta_n(x_e)] + [\Delta_n(\hat{x}_e) - \Delta(\hat{x}_e)] - [\Delta_n(\hat{x}_e) - \Delta_n(x_e)] \\ &< 2\epsilon \text{ for all large } n. \end{aligned}$$

It follows from assumption (10) that $\hat{x}_e \to x_e$ a.s. Hence the split point converges to its population counterpart. Note that if F_1 and F_2 are normal with the same variance, then $x_e = x_0$.

5.2. Asymptotics for QUEST

Let η_j and σ_j^2 denote the mean and variance of the F_j, and let s_j^2 be the sample variance based on sample size n, $j = 1, 2$. In the case that the priors are equal, the roots of (4) simplify to

$$\bar{Y} + (s_1^2 - s_2^2)^{-1}[(\bar{Y} - \bar{X})s_2^2 \pm s_1 s_2 |\bar{X} - \bar{Y}|].$$

Let $\hat{x}_q$ denote the root closer to $\bar{Y}$, i.e., $\hat{x}_q$ is the QUEST split. Then

$$\begin{aligned}\hat{x}_q &= \bar{Y} + (s_1^2 - s_2^2)^{-1}[(\bar{Y} - \bar{X})s_2^2 - \text{sgn}(\bar{Y} - \bar{X})s_1 s_2 |\bar{X} - \bar{Y}|] \\ &= (s_1 + s_2)^{-1}(\bar{Y}s_1 + \bar{X}s_2).\end{aligned}$$

As $n \to \infty$, we have $\hat{x}_q \to x_q$ a.s. and $\sqrt{n}(\hat{x}_q - x_q) \xrightarrow{D} N(0, \omega^2)$, where $x_q = (\sigma_1 + \sigma_2)^{-1}(\eta_2\sigma_1 + \eta_1\sigma_2)$, $\omega^2 = 2\sigma_1^2\sigma_2^2/(\sigma_1 + \sigma_2)^2$. Clearly, $x_q = x_0$ if the class distributions are normal. Therefore the split points via exhaustive search and via QUEST converge to the same value if F_1 and F_2 are normal with equal variance.

Consider now the situation where the densities are not necessarily normal but are reflections of each other about the point x_0, i.e.,

$$F_2(x - x_0) = 1 - F_1(x_0 - x) \ \ \forall x. \tag{11}$$

Then $\sigma_1 = \sigma_2$, $\eta_1 + \eta_2 = 2x_0$, and hence $x_q = x_0$. On the other hand, (11) implies that (9) is satisfied with $x_e = x_0$. Therefore $\hat{x}_q$ and $\hat{x}_e$ again converge to the same limit.

5.3. Disjoint supports: power distributions

While the preceding analysis shows that $\hat{x}_q$ typically converges at the $\sqrt{n}$-rate, the convergence rate of $\hat{x}_e$ is harder to determine in general. There is one special situation, however, where the asymptotic distribution of $\hat{x}_e$ is amenable to theoretical analysis. This is the situation where the distributions have disjoint support.

Suppose that the class distributions satisfy condition (11) with $x_0 = 0$. Then the population density functions satisfy the condition $f_2(x) = f_1(-x)$, i.e., the class populations are reflections of each other about the point $x = 0$. Let σ^2 denote the common population variance and suppose that $\eta_1 \neq 0$ and $\sigma^2 < \infty$. With probability converging to 1 as $n \to \infty$, the split point using the QUEST method will be equal to $(\bar{X} + \bar{Y})/2$, which is $\sqrt{n}$-consistent.

Now suppose further that $\inf\{x : F_1(x) > 0\} = 0$. Then the split point obtained with the exhaustive search method is $(X_{(1)} + Y_{(n)})/2$. It follows from the theory of extreme order statistics (Galambos (1978)) that if $F_1(x)$ satisfies the condition $\lim_{\delta \to 0} F_1(\delta x)/F_1(\delta) = x^{-\gamma}$, $x > 0$ for some constant $\gamma > 0$, then

$$\lim_{n \to \infty} P\{X_{(1)} + Y_{(n)} < xF_1^{-1}(1/n)\} = P(W_1 + W_2 < x),$$

where W_1 is a random variable with distribution function $P(W_1 \leq x) = 1 - \exp(-x^\gamma)$ and W_2 is an independent copy of $-W_1$. The rate of convergence to 0 of the split point $(X_{(1)} + Y_{(n)})/2$ therefore depends on the limiting behavior of $F_1^{-1}(1/n)$.

A simple example that illustrates the range of possible rates of convergence is provided by the family of power function densities with $f_1(x) = px^{p-1}$, $0 < x < 1$, and $p > 0$. Then $F_1(x) = x^p$, $F_1^{-1}(1/n) = n^{-1/p}$, and $\gamma = p$. Hence

$$\lim_{n\to\infty} P\{n^{1/p}(X_{(1)} + Y_{(n)})/2 \le x\} = P\{(W_1 + W_2)/2 \le x\}.$$

The convergence to 0 of the split point $(X_{(1)} + Y_{(n)})/2$ is slower or faster than that of $(\bar{X} + \bar{Y})/2$ depending on whether $p > 2$ or $p < 2$. If $p = 2$, the splits points are both $\sqrt{n}$-consistent, although with different limit distributions.

Table 4. Computational speed of QUEST relative to exhaustive search method for ordered variables with normal distributions.

		Number of classes (J)			
K	N	2	3	5	10
2	300	8.0	7.6	7.4	8.0
	900	21.3	20.2	20.1	21.0
	3000	85.4	73.8	75.0	76.0
	9000	273.0	249.0	254.0	254.8
10	300	9.4	9.5	9.6	10.2
	900	25.6	25.8	26.2	26.4
	3000	92.7	90.5	92.4	94.2
	9000	291.2	289.7	297.4	297.7
20	300	9.6	9.7	9.8	10.4
	900	28.9	29.5	29.2	30.3
	3000	96.0	98.1	96.4	98.2
	9000	285.5	294.3	302.1	310.7

6. Computational Speed

The FACT and QUEST split selection procedures tend to be equally fast, except when there are categorical variables. FACT is slower than QUEST in the latter case because it needs to carry out the CRIMCOORD transformation for every categorical variable whereas QUEST only performs the CRIMCOORD transformation on the categorical variable that is selected. On the other hand, the exhaustive search method is expected to require much more computations than the QUEST or FACT methods. To obtain an indication of how the computational speed of QUEST relative to exhaustive search scales with increasing values of K, N, and J, the computational times of the two methods were measured on simulated normal and discrete uniformly distributed data. Table 4 gives the ratios of computation times of exhaustive search to QUEST for normally distributed data. The most striking pattern is that the relative speed increases roughly linearly with the sample size N but is fairly constant across the values of J and K.

Table 5 shows the corresponding results for uniformly distributed categorical variables and $N = 300$. Similar results were obtained for $N = 3000$. The main conclusions are:

1. The exhaustive search method is faster than QUEST when $J = 2$. This is due to a short-cut algorithm that reduces the number of splits searched from $(2^{M-1} - 1)$ to M (Breiman et al. (1984), Theorem 4.5). This short-cut is only applicable when $J = 2$.
2. For $J > 2$ and $M > 4$, QUEST is faster than exhaustive search, with relative speed increasing exponentially with M and linearly with K.
3. The relative speed does not vary much with J for $J \geq 3$.

Table 5. Computational speed of QUEST relative to exhaustive search method for categorical variables distributed as U_M. Total learning sample size is 300; class sizes are equal.

K	M	$J = 2$	$J = 3$	$J = 5$	$J = 10$
2	4	0.39	0.5	0.6	0.6
	10	0.12	4.2	4.3	4.4
	15	0.05	60.5	59.7	61.3
	20	0.02	1,050.2	1,044.1	1,070.5
10	4	0.68	1.1	1.0	1.1
	10	0.29	16.2	16.1	16.5
	15	0.15	265.8	259.4	263.8
	20	0.09	4,943.4	4,844.6	4,984.3
20	4	0.79	1.2	1.2	1.2
	10	0.46	25.2	25.4	24.9
	15	0.26	447.8	446.7	443.3
	20	0.16	9,103.4	9,084.7	8,904.8

7. Two Examples

The results so far have been restricted to node-wise comparisons. We now use two examples to compare the size and accuracy of the trees constructed by the two split selection methods when each is employed with the cross-validation pruning method of CART.

7.1. Waveform simulation example

Our first comparison uses the waveform simulation example in Breiman et al. (1984). There are 3 classes and 21 predictor variables, with each class being a random combination of two triangular waveforms with noise added. We carried out 30 simulation trials. In each trial, 1,000 learning samples were simulated using equal class prior probabilities. Trees were constructed using the QUEST method (with univariate and linear combination splits, and with $\alpha = 0.05$) and

the exhaustive search method. As in CART, the trees were pruned with 10-fold cross-validation and the 1-SE rule. The error rate of each tree was estimated with an independent test sample of size 5,000.

Results on the number of terminal nodes, the estimated error rates, and the computational times on a SUN SPARCstation 20/50 are shown in Figure 7. The QUEST method using linear combination splits is clearly best in terms of accuracy and size of the trees. Its computational times are also much lower than those for the exhaustive search method, although they are about four times that for QUEST using univariate splits. The latter method has similar accuracy as the exhaustive search method, although it tends to produce trees that are slightly larger.

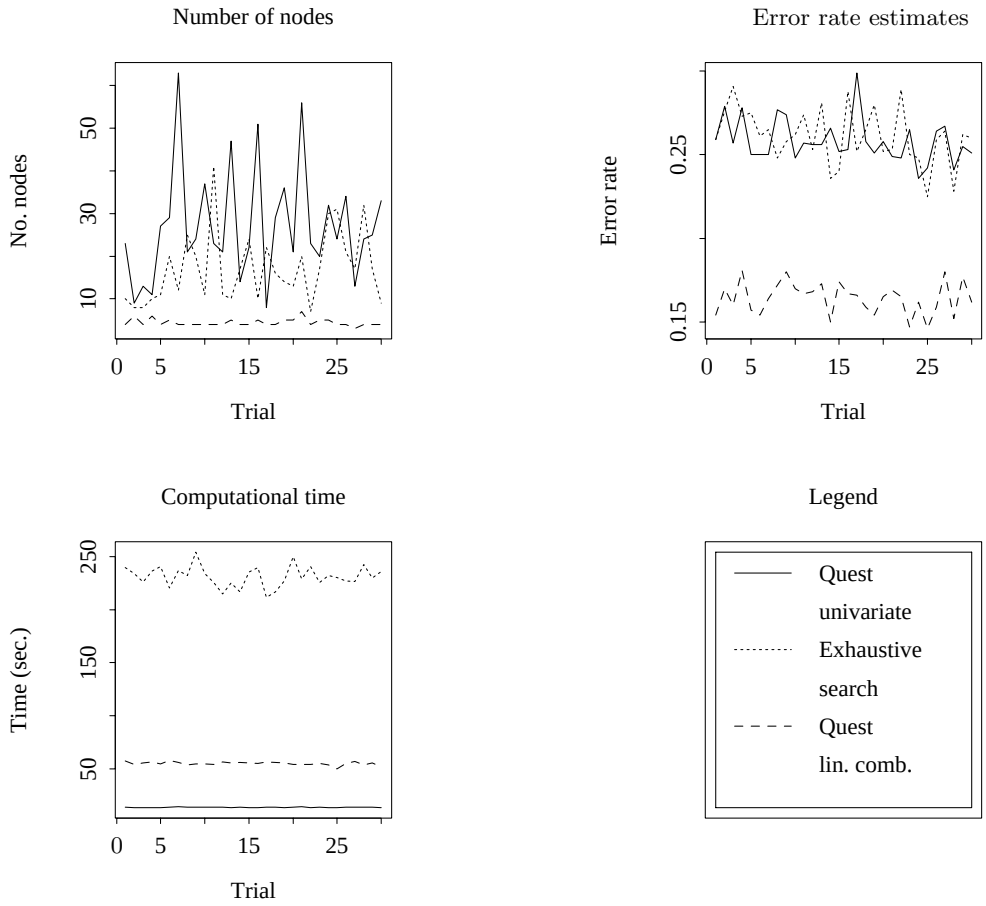


Figure 7. Results for 30 trials of the waveform simulation example on a SUN SPARCstation 20/50.

7.2. Real data example

Our second example employs a real data set. The data consist of evaluations of teaching performance over three regular semesters and two summer semesters of 151 teaching assistant (TA) assignments at the Statistics Department of the University of Wisconsin–Madison. The scores were divided into 3 roughly equal-sized categories ("low", "medium", and "high") to form the class variable. The predictor variables were

1. Whether or not the TA is a native English speaker (binary).
2. Course instructor (25 categories).
3. Course (26 categories).
4. Summer or regular semester (binary).
5. Class size (ordered).

Table 6 gives the P-values for each predictor variable. The highly significant P-value for the Semester variable is expected because summer teaching assignments are normally awarded to the better TAs.

The QUEST and exhaustive search methods were applied to the data using 10-fold cross-validation pruning and the 1-SE rule on an IBM RS/6000 workstation. To study the effect of the choice of cross-validation samples, the procedures were repeated 12 times using different random number seeds. Figure 8 plots the sizes of the 12 pairs of trees. In contrast to the waveform example, the QUEST method tends to yield shorter trees than the exhaustive search method here. Because the Semester variable is most significant, all the QUEST trees split on this variable first. On the other hand, all the trees based on exhaustive search split first on the Course variable. The difference is probably due to variable selection bias because the Course variable generates the most splits ($2^{25} - 1 \approx 34 \times 10^6$ splits versus 1 for the Semester variable).

The average computation times for the exhaustive search and QUEST methods for this data set were 30.5 CPU hours and 1 CPU second, respectively. Owing to the exceedingly long times, 10-fold cross-validation error estimates were obtained for only 2 of the 12 pairs of trees (the 1st and last pairs in Figure 8). The exhaustive search method took about 10 days of CPU time to obtain the error estimates for each tree while QUEST took 17 seconds. The error estimates were not significantly different.

Table 6. P-values for TA evaluation example

Predictor	P-value
English speaker	0.0026
Instructor	0.0174
Course	0.0091
Semester	0.0018
Class size	0.3930

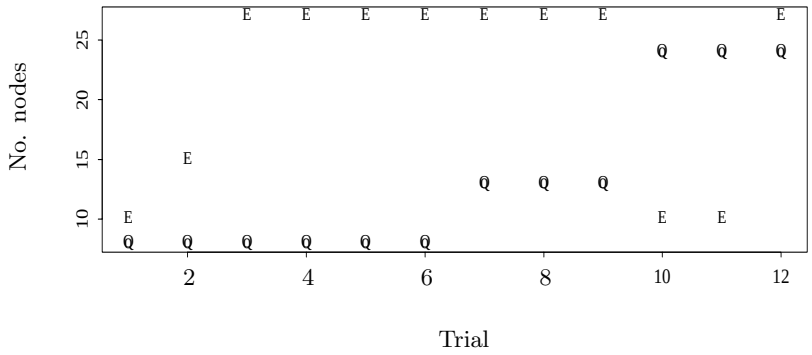


Figure 8. Effect of random number seed on number of terminal nodes in 12 univariate trees for TA evaluation data. The letters "E" and "Q" refer to the exhaustive search and QUEST methods, respectively. The trials are arranged according to increasing values of "Q".

8. Conclusion

The main appeal of classification trees is the insight thought to be provided by the splits. We demonstrated that exhaustive search tends to select variables that afford more splits. Therefore such trees should be interpreted with caution.

We showed that the FACT approach is also not free of selection bias. Fortunately, its basic strategy of using statistical tests to guide variable selection is sound, and we modified it to remove the bias. Additional modifications were made to ensure binary splits. This makes it feasible to compare methods in terms of split point variability and tree size.

In terms of classification accuracy, variability of split points, and tree size, our results show that there is no clear winner when univariate splits are used. Sometimes QUEST is better and other times exhaustive search is better. However, QUEST trees based on linear combination splits are usually shorter and more accurate than the same trees based on univariate splits.

The QUEST computer program is many programs in one. Besides implementing the QUEST split selection approach, it has an option for exhaustive search. It can be used with CART-style pruning or FACT-style direct stopping, and it accepts user-specified class priors and misclassification costs. All the comparisons in this paper were performed with this computer program. The Fortran source code, user guide, and compiled binaries for the MSDOS and OS/2 operating systems are available from `http://www.stat.wisc.edu/~loh/loh.html`.

Some of the ideas described here have been extended to tree-structured function estimation in Ahn and Loh (1994), Chaudhuri, Huang, Loh and Yao (1994), Chaudhuri, Lo, Loh and Yang (1995), and Yan (1995).

Acknowledgements

Loh's research was supported in part by U. S. Army Research Office grants DAAL03-91-G-0111 and DAAH04-94-G-0042 and National Science Foundation grant DMS-9304378. Shih's research was supported in part by by Republic of China National Science Council grant 83-0208-M-194-024. The authors are grateful to an Associate Editor for many helpful comments and suggestions.

References

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

Breiman, L. (1996a). Bagging predictors. *Machine Learning*. To appear.

Breiman, L. (1996b). Bias, variance, and arcing classifiers, Technical Report 460, Department of Statistics, University of California, Berkeley.

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

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

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

Chou, P. A. (1991). Optimal partitioning for classification and regression trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13**, 340-354.

Doyle, P. (1973). The use of Automatic Interaction Detector and similar search procedures. *Oper. Res. Quarterly* **24**, 465-467.

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

Galambos, J. (1978). *The Asymptotic Theory of Extreme Order Statistics*. Wiley, New York.

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. Wiley, New York.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm 136. A k-means clustering algorithm. *Appl. Statist.* **28**, 100.

Levene, H. (1960). Robust tests for equality of variances. In *Contributions to Probability and Statistics* (Edited by I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow and H. B. Mann), 278-292. Stanford University Press, Palo Alto.

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

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press, London.

Math Works (1991). *MATLAB User's Guide*, The MathWorks, Inc., Cochituate Place, 24 Prime Park Way, Natick, MA 01760.

Morgan, J. N. and Messenger, R. C. (1973). THAID: A sequential analysis program for the analysis of nominal scale dependent variables, Technical report, Institute for Social Research, University of Michigan, Ann Arbor.

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

Murthy, S. K., Kasif, S. and Salzberg, S. (1994). A system for induction of oblique decision trees. *J. Artificial Intelligence Research* **2**, 1-33.

Quinlan, J. R. and Cameron-Jones, R. M. (1995). Oversearching and layered search in empirical learning. In *Proceedings of the* 14*th International Joint Conference on Artificial Intelligence, Montréal,* Vol. 2 (Edited by Morgan Kaufman), 1019-1024.

Vanichsetakul, N. (1986). *Tree-Structured Classification via Recursive Discriminant Analysis*. PhD thesis, Department of Statistics, University of Wisconsin, Madison.

Yan, C. (1995). *Regression Trees and Nonlinear Time Series Modeling*. PhD thesis, Department of Statistics, University of Wisconsin, Madison.

Department of Statistics, University of Wisconsin, Madison, U.S.A.

E-mail: loh@stat.wisc.edu

Department of Mathematics, National Chung Cheng University, Minghsiung Chiayi 621, Taiwan.

E-mail: yshih@math.ccu.edu.tw

(Received March 1994; accepted December 1996)