The Korean Communications
in Statistics Vol. 9, No. 3, 2002
pp. 809-824

# A Study on Unbiased Methods in Constructing Classification Trees[1]

Yoon-Mo Lee[2] and Moon Sup Song[3]

## Abstract

we propose two methods which separate the variable selection step and the split-point selection step. We call these two algorithms as CHITES method and F&CHITES method. They adapted some of the best characteristics of CART, CHAID, and QUEST. In the first step the variable, which is most significant to predict the target class values, is selected. In the second step, the exhaustive search method is applied to find the splitting point based on the selected variable in the first step. We compared the proposed methods, CART, and QUEST in terms of variable selection bias and power, error rates, and training times. The proposed methods are not only unbiased in the null case, but also powerful for selecting correct variables in non-null cases.

*Keywords* : variable selection bias, variable selection power, exhaustive search method

## 1. Introduction

A classification tree is constructed by recursively partitioning the training sample of data in which the class labels of the target variable are known. A splitting rule decides how to partition the cases (or objects) at each node. Any splitting rule can be characterized by its variable selection step and split-point selection step. Some algorithms such as CART (Breiman et al., 1984) and CHAID (Kass, 1980) do not separate these two steps, and some algorithms such as QUEST (Loh and Shih, 1997) and CRUISE(Kim and Loh, 2001) separate these two steps.

Many splitting rules have been proposed as a part of classification tree algorithms. Classification tree algorithms can be characterized by the number of subnodes they produce at each split. Some, such as CART and QUEST, yield binary splits. Others allow multiway splits. In C4.5 (Quinlan, 1993) a split on a continuous predictor variable gives two subnodes but a split on a categorical predictor yields one subnode for each predictor value. The CHAID

[1] This research was supported in part by the Brain Korea 21 Project.

[2] Consulting Division/Director, Systembusiness Co., Ltd., Seoul 150-872, Korea
E-mail : deonian@systembusiness.co.kr

[3] Professor, Department of Statistics, Seoul National University, Seoul 151-742, Korea
Email : songms@plaza.snu.ac.kr

algorithm uses the same approach, but tries to merge some of the subnodes.

One of the most popular algorithms to construct a classification tree is CART, which uses the exhaustive search method to find the best splitting rule at each node. The exhaustive search method can be summarized as follows.

If $X$ is an ordered variable, this approach searches over all possible values $c$ for splits of the form

$$X \leq c, \quad \forall c \in (-\infty, \infty). \tag{1}$$

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

$$X \in A,$$

where $A$ is a non-empty subset of the set of values taken by $X$. However, it is well known that the exhaustive search method of the CART tends to be biased towards selecting variables that afford more splits. As a result, such trees should be interpreted with caution.

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 categorical variables, the order of computations increases exponentially with the number of categories, being $(2^{M-1}-1)$ for a variable with $M$ values.

(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.

The FACT algorithm (Loh and Vanichsetakul, 1988) employs a computationally simpler approach. Instead of combining the problem of variable selection with that of split-point selection, 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 is applied to it to find the split point $c$. 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.

The QUEST algorithm made some improvements on FACT. For example, QUEST employs quadratic discriminant analysis method instead of linear discriminant analysis method of FACT in split-point selection step. According to the results of Loh and Shih (1997), the QUEST (i) has negligible variable selection bias, (ii) retains the computational simplicity of FACT, and (iii) includes pruning as an option.

A method is called biased in variable selection, if the predictor variables do not have the same chance to be selected for splitting even when they are independent of the target variable. In this paper, we propose two unbiased methods which separate the variable selection step and the split-point selection step. To select the most significant predictor variable in the first step, two approaches are proposed. The first is to use the chi-square tests of independence between the predictor variables and the target variable. The second is to use $F$ -tests on continuous predictors and chi-square tests on categorical predictors. The proposed methods are compared with CART and QUEST. The results show that the proposed methods are compatible with CART and QUEST, and sometimes better.

## 2. Proposed Algorithms

In order to solve the bias problems in the exhaustive search algorithm such as CART, we propose two methods which separate the variable selection step and the split-point selection step. In the first step the variable, which is most significant to predict the target class values, is selected. In the second step, the exhaustive search method is applied to find the splitting point based on the selected variable in the first step.

To select the most significant predictor variable in the first step, two approaches are proposed. The first method is to use the chi-square tests of independence between the predictor variables and the target variable. The continuous variables must be discretized before applying the chi-square test. The variable with the smallest $p$-value will be selected as a splitting variable in the variable selection step.

In order to discretize continuous predictors, equal proportion discretization is perhaps the simplest method. It involves sorting the observed values of continuous variable and dividing the range of observed values into $M$ equal proportion bins, where $M$ is a parameter supplied by users. Lee (2002) compared the equal proportion method with the entropy-based method. But, since the difference between the results of two methods is not significant, we choose the equal proportion method which is much simpler.

The idea of the second method is to use the variable selection method of QUEST. According to the simulation study by Song and Yoon (2000), the QUEST algorithm shows no serious bias in variable selection. We thus adopt the basic ideas in QUEST for selecting the most significant variable. That is, we compute $p$-values from the ANOVA $F$-tests for numerically ordered predictors and the chi-square tests for categorical predictors. Again we choose the variable with the samllest $p$-value as the splitting variable.

We assume that $X_1, \cdots, X_{K_1}$ are numerically ordered variables and $X_{K_1+1}, \cdots, X_K$ are categorical variables. At each node $t$, let $M_k(t)$ be the number of distinct categories of the $k$th predictor variable, $J_t$ be the number of classes of the target variable, and $N(t)$ be the total number of cases. Then, at node $t$, the variable selection step based on chi-square test can be described as follows.

**Variable Selection Algorithm : Chi-square Test Method**

1. If $K_1 \geq 1$, categorize the numerically ordered variables using the equal proportion method at the root node.
2. For each $k(1 \leq k \leq K)$, compute the $p$-value, $\hat{\alpha}(k)$, of the contingency table chi-square test of independence between the categorical or categorized predictor $X_k$ and the target variable. The degrees of freedom of the chi-square statistic are $(J_t - 1)(M_k(t) - 1)$.
3. Let $k'$ be the value of $k$ such that

$$\hat{\alpha}(k') = \min_{1 \leq k \leq K} \hat{\alpha}(k)$$

Then select $X_{k'}$ as the splitting variable.

The variable selection step based on $F$ and chi-square tests can be described as follows.

**Variable Selection Algorithm : $F$ & Chi-square Test Method**

Let $\alpha \in (0, 1)$ be a pre-specified level of significance.

1.(a) For each continuous variable $X_k(1 \leq k \leq K_1)$, compute the $p$-value, $\hat{\alpha}(k)$, of the one-way ANOVA $F$-test on $X_k$ by treating the target classes as treatment labels. The degrees of freedom of the $F$-statistic are $J_t - 1$ and $N(t) - J_t$. Let $k_1$ be the value of $k$ such that

$$\hat{\alpha}(k_1) = \min_{1 \leq k \leq K_1} \hat{\alpha}(k)$$

(b) For each categorical variables $X_k(K_1 + 1 \leq k \leq K)$, compute the $p$-value, $\hat{\alpha}(k)$, of the contingency table chi-square test of independence between $X_k$ and the target variable. The degrees of freedom of the chi-square statistic are $(J_t - 1)(M_k(t) - 1)$. Let $k_2$ be the value of $k$ such that

$$\hat{\alpha}(k_{2)} = \min_{K_1 + 1 \leq k \leq K} \hat{\alpha}(k).$$

2. Define $k' = k_1$ if $\hat{\alpha}(k_1) \leq \hat{\alpha}(k_2)$, otherwise define $k' = k_2$.
3. If $\min(\hat{\alpha}(k_1), \hat{\alpha}(k_2)) < \alpha/K$, select $X_{k'}$ to split the node.
4. Otherwise, if $\min(\hat{\alpha}(k_1), \hat{\alpha}(k_2)) > \alpha/K$, then

(1) Compute the one-way ANOVA $F$-test for the ordered variables based on the absolute deviations $z_{ik}^{(j)} = |x_{ik}^{(j)} - \bar{x}_k^{(j)}|$, where $x_{ik}^{(j)}$ is the $i$th observation of $X_k$ among class $j$ cases and $\bar{x}_k^{(j)} = N_j(t)^{-1} \sum_{i=1}^{N_j(t)} x_{ik}^{(j)}$ with $N_j(t)$ the number of class $j$ cases at node $t$. Let

$$\alpha(k'') = \min_{1 \le k \le K_1}\{p(k)\},$$

where $p(k)$ is the $p$-value of $F$-test based on $z_{ik}$.

(2) If $\tilde{\alpha}(k'') < \alpha/(K+K_1)$, select variable $X_{k''}$. Otherwise, select variable $X_{k'}$.

If $X_{k'}$ is selected as a splitting variable in the variable selection step, then apply the exhaustive search method of CART based on $X_{k'}$. We use the Gini index as the impurity measure. Thus, at node $t$ we want to find the split point to minimize

$$i(t) = \sum_{i \neq j} p(i|t)p(j|t), \quad i,j = 1,\cdots,J_t\ ,$$

where $p(j|t)$ is the proportion of cases belonging to class $j$.

**Split-Point Selection Algorithm : Exhaustive Search Method**

1. (a) If $X_{k'}$ is numerically ordered, let the sorted observed values be $x_{(1)} \le \cdots \le x_{(n)}$ at node $t$, where $n = N(t)$. For each $j = 1,\cdots,n-1$, send the cases associated with $x_{(1)},\cdots,x_{(j)}$ to the left subnode, and the cases with $x_{(j+1)},\cdots,x_{(n)}$ to the right subnode, and compute the impurity $i_j(t)$. Let

$$i^*(t) = \min_j \ i_j(t).$$

(b) If $X_{k'}$ is categorical, let $\{c_1,\cdots,c_M\}$ be the set of categories of $X_{k'}$ at node $t$. For each subset $A$ of $\{c_1,\cdots,c_M\}$, send the cases associated with $A$ to the left subnode and the others to the right subnode, and compute the impurity $i_A(t)$. Let

$$i^*(t) = \min_A \ i_A(t).$$

2. Perform the split corresponding to $i^*(t)$.

In the case of two-class problem, the exhaustive search algorithm on categorical variable can be reduced to the algorithm on numerically ordered variable, by Theorem 4.5 of Breiman et al. (1984). Let $p_{c_i}(t) = P[\text{ class}=1\,|\,t,\ X_k = c_i\,]$, $i = 1,\cdots,M$, and let $p_{c_{(1)}} \le \cdots \le p_{c_{(M)}}$ be ordered $p_{c_i}(t)$'s. Then the exhaustive search algorithm 1(a) can be applied on $c_{(1)},\cdots,c_{(M)}$, assuming they are ordered.

The splitting process is repeated at every node until some stopping criteria are satisfied. In our algorithm, the splitting process at a node will stop if one of the following conditions is satisfied.

1. There is only one class in the node.
2. The number of cases in the node is less than the preassigned value.
3. The error rate in the node does not decrease by splitting.

In this paper we proposed two algorithms for constructing a classification tree. We call these two algorithms as CHITES (CHI-square Test and Exhaustive Search) method, F&CHITES ( $F$ & CHI-square Test and Exhaustive Search) method. CHITES applies the chi-square test to select the most significant variable in variable selection step and applies the exhaustive search method to find the split point using the selected variable in the first step. F&CHITES applies $F$-test to numerically ordered variables and chi-square test to categorical variables in the variable selection step. Assuming normality, F&CHITES is expected to be efficient in selecting the most significant variable, meanwhile CHITES is expected to be robust in selecting significant variables.

## 3. A Comparative Study of the Proposed Algorithms

In this section, the results of simulation studies are discussed. We also want to confirm that the proposed methods, CHITES and F&CHITES, are not seriously biased in the null case. Here, the null case means that predictor variables and the target variable are all mutually independent and thus each predictor variable has equal chance to be chosen as a selection variable.

We are interested not only in unbiasedness in the null case but also in power for selecting correct variables in non-null cases. To compare the selection power, we generated some training data sets in which the target variable and predictor variables are correlated.

The proposed methods, CART, and QUEST are also compared in terms of error rates on simulated data sets and also on real data sets. The real data sets are drawn from the UCI (University of California, Irvine) Repository (Blake and Merz, 1998). We briefly describe the nine data sets used in Table 1. The details of these data sets are appeared in http://www.ics.uci.edu/~mlearn/MLRepository.html.

**Table 1.** Characteristics of the real data

| | | number | number of variables | |
|---|---|---|---|---|
| data set name | size | of classes | numeric | categorical |
| balance | 625 | 3 | 4 | 0 |
| bupa | 345 | 2 | 6 | 0 |
| german | 1000 | 2 | 7 | 13 |
| glass | 214 | 6 | 10 | 0 |
| iris | 150 | 3 | 4 | 0 |
| ionosphere | 351 | 2 | 34 | 0 |
| new-thyroid | 215 | 3 | 6 | 0 |
| pima indian | 768 | 2 | 8 | 0 |
| wine | 178 | 3 | 14 | 0 |

### 3.1 Comparison in Variable Selection Bias

A Monte Carlo simulation study was performed to compare the variable selection bias of the exhaustive search method of CART and the proposed methods CHITES and F&CHITES.

To examine the selection bias at a node, we considered 5 mutually independent predictor variables and a binary target variable which is also independent of the predictor variables. Thus in this null case each predictor is supposed to be selected with probability 0.2. The variables $X_1$ and $X_2$ are continuous with normal and exponential distributions, respectively. The variable $X_3$ is an ordered predictor which has a uniform distribution on integers {1, 2, 3, 4}. Variables $X_4$ and $X_5$ are categorical such that $X_4$ has a uniform distribution on {1, 2}, and $X_5$ on {1, 2, ⋯, M}, where $M$ is the number of categories of $X_5$. In our simulation study we considered two cases: $M=5$ and $M=15$. The continuous variables are discretized using 5 equal-proportion bins. Target values of the first one half cases are assigned the value 1 and the other half the value 2. The sample sizes examined are 200 and 500.

The data were generated using the SAS program, and the individual algorithms were applied to these simulated data sets. All the algorithms are implemented in S-Plus. The selection probability of a variable is estimated by the number of selections of the variable at the root node. The selection probability of each $X_i$ is estimated based on 300 Monte Carlo iterations. Thus the standard error is 0.023.

The results of Monte Carlo simulation study to compare the biasness of the proposed methods with CART are summarized in Table 2. The estimated probabilities in the table are the proportion of being selected as a splitting variable at the root node.

**Table 2.** Estimated probabilities of variable selection. An unbiased method selects each variable with probability 0.2. Estimates are based on 300 replicates and training samples of size $N$ = 200 and 500. $M$ is the number of categories in $X_5$. Standard errors of estimates are about 0.023.

| Algorithms | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | $M=5$, $N=200$ | | | | | $M=5$, $N=500$ | | | | |
| CART | .387 | .343 | .060 | .013 | .197 | .387 | .383 | .033 | .030 | .167 |
| CHITES | .197 | .160 | .233 | .197 | .213 | .193 | .260 | .203 | .153 | .190 |
| F&CHITES | .213 | .163 | .244 | .160 | .220 | .187 | .250 | .183 | .167 | .213 |
| | $M=15$, $N=200$ | | | | | $M=15$, $N=500$ | | | | |
| CART | .103 | .120 | .017 | .003 | .757 | .077 | .137 | .007 | .003 | .776 |
| CHITES | .207 | .220 | .193 | .167 | .213 | .170 | .210 | .240 | .153 | .227 |
| F&CHITES | .227 | .206 | .247 | .167 | .153 | .183 | .210 | .237 | .150 | .220 |

When $M=5$, CART shows some bias with respect to the continuous variables $X_1$ and $X_2$. When $M=15$, the many-valued categorical variable $X_5$ is more likely to be selected, and as a result CART is seriously biased toward $X_5$. The results of CHITES and F&CHITES are similar, and they show no serious bias in the null case. So CHITES and F&CHITES are better than CART in terms of variable selection bias.

### 3.2 Comparison in Variable Selection Power

To investigate the variable selection power of each method in selecting the informative variables, we considered 3 simulation models. Those are the shifted mean model, the correlated model with binary variable, and the shifted mean model in contaminated normal.

#### Shifted Mean Model

To form a correlation structure between a predictor variable and the target variable, $X_1$ variable is shifted for class 1 as follows.

$$X_1 \sim \begin{cases} N(0.5,1), & \text{for Class 1} \\ N(0,1), & \text{for Class 2.} \end{cases}$$

All other variables are the same as in the null case of Table 2.

**Table 3.** Estimated probabilities of correct variable selection when $X_1$ variable has discriminatory power. Estimates are based on 300 replicates and training samples of size $N$ = 200 and 500. $M$ is the number of categories in $X_5$.

| Algorithms | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | $M=5$, | | $N=200$ | | | $M=5$, | | $N=500$ | | |
| CART | .950 | .037 | .000 | .000 | .013 | 1.000 | .000 | .000 | .000 | .000 |
| CHITES | .880 | .037 | .033 | .030 | .020 | 1.000 | .000 | .000 | .000 | .000 |
| F&CHITES | .963 | .010 | .013 | .007 | .007 | 1.000 | .000 | .000 | .000 | .000 |
| | $M=15$, | | $N=200$ | | | $M=15$, | | $N=500$ | | |
| CART | .753 | .010 | .003 | .000 | .234 | .983 | .000 | .000 | .000 | .017 |
| CHITES | .873 | .037 | .037 | .006 | .047 | .993 | .003 | .000 | .000 | .003 |
| F&CHITES | .950 | .010 | .013 | .007 | .020 | 1.000 | .000 | .000 | .000 | .000 |

Table 3 shows the estimated probabilities that the informative variable is selected. In the case of $M=5$ and $N=200$, the CART method detects the informative variable fairly well. However, considering that CART is already biased toward the continuous variable $X_1$, the selection power may not be enough. When $M=15$ and $N=200$, the CART method is biased toward the many-valued categorical variable $X_5$ and therefore it loses power in selecting the informative variable $X_1$. As the sample size increases, the discriminatory power also significantly increases in all cases. F&CHITES is, in general, better than CHITES in these shifted mean models.

### Correlated Model with Binary Variable

As we have seen in Table 2, the exhaustive search method of CART is more likely to select many-valued continuous or categorical variables. Thus the binary variable $X_4$ is rarely selected as a splitting variable by CART in Table 2. In this sense Table 2 model is favorable to CART, and therefore we consider an opposite case.

To generate a correlation structure between the binary predictor variable $X_4$ and the target variable, the values of $X_4$ are generated as follows.

$$\text{In case 1,}\quad X_4 = \begin{cases} 1, & \text{with probability } 0.55 \\ 2, & \text{with probability } 0.45 \end{cases}$$

$$\text{In case 2,}\quad X_4 = \begin{cases} 1, & \text{with probability } 0.45 \\ 2, & \text{with probability } 0.55 \end{cases}$$

All other variables and simulation designs are the same as in the null case of Table 2.

**Table 4**. Estimated probabilities of variable selection when $X_4$ variable has discriminatory power. Estimates are based on 300 replicates and training samples of size $N$ = 200 and 500. $M$ is the number of categories in $X_5$.

| Algorithms | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | $M=5$, | | $N=200$ | | | $M=5$, | | $N=500$ | | |
| CART | .413 | .393 | .050 | .000 | .143 | .297 | .240 | .017 | .320 | .127 |
| CHITES | .177 | .130 | .157 | .407 | .130 | .033 | .027 | .023 | .883 | .033 |
| F&CHITES | .157 | .147 | .157 | .413 | .127 | .033 | .077 | .037 | .820 | .033 |
| | $M=15$, | | $N=200$ | | | $M=15$, | | $N=500$ | | |
| CART | .097 | .113 | .013 | .000 | .773 | .090 | .087 | .003 | .047 | .773 |
| CHITES | .177 | .123 | .153 | .390 | .157 | .033 | .027 | .023 | .900 | .017 |
| F&CHITES | .173 | .143 | .147 | .390 | .147 | .037 | .077 | .037 | .833 | .017 |

Table 4. gives simulation results under the correlated model with binary variable $X_4$. As expected, the proposed methods are significantly better than the CART method. When $M=5$ and $N=200$, the estimated probability of CART selecting $X_4$ as a splitting variable is .000, while that of CHITES is .407. When $M=5$ and $N=500$, the estimated probability of CART selecting $X_4$ as a splitting variable is .320, while that of CHITES is .883. When $M=15$, the effect of biasness toward $X_5$ does not disappear, and therefore the selection power of CART are seriously lower than that of CHITES or F&CHITES. The behaviors of CHITES and F&CHITES are almost the same.

### Shifted Mean Model in Contaminated Normal

Table 3 shows that the F&CHITES method is better than CART or CHITES in discriminating power. But, since F&CHITES uses $F$-test for continuous variables to select the most significant variables, F&CHITES has some advantages. Algorithms in data mining need to be robust enough to cope with various non-normal data sets. We thus compare the three methods in non-normal case. The predictor variables $X_2, \cdots, X_5$ are generated from the contaminated normal distribution.

$$CN(0,\, 5^2,\, 0.2) = 0.8N(0,1) + 0.2N(0,5^2)$$

That is, 80% of the data are generated from $N(0,1)$ and 20% from $N(0,5^2)$. $X_1$ variable is shifted for class 1 as follows.

$$X_1 \sim \begin{cases} CN(.5,\, 5^2,\, 0.2), & \text{for Class 1} \\ CN(0,\, 5^2,\, 0.2), & \text{for Class 2} \end{cases}$$

Thus $X_1$ is the informative variable for prediction of the class of target variable. As before, the probabilities of selecting predictor variables as a splitting variable are estimated based on 300 Monte Carlo iterations.

**Table 5.** Estimated probabilities of variable selection when $X_1$ variable has discriminatory power. All variables are generated from contaminated normal distributions. Estimates are based on 300 replicates and training samples of size $N$ = 200 and 500.

| Algorithms | $N=200$ | | | | | $N=500$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ |
| CART | .720 | .103 | .057 | .057 | .063 | .457 | .127 | .150 | .113 | .153 |
| CHITES | .687 | .067 | .076 | .100 | .070 | .420 | .127 | .133 | .147 | .173 |
| F&CHITES | .567 | .140 | .116 | .097 | .080 | .340 | .150 | .173 | .167 | .170 |

Table 5 shows the estimated probabilities of selecting each variable. The results show that the F&CHITES method based on $F$-test has the lowest discriminatory power in both cases, $N=200$ and $N=500$. The result implies that the F&CHITES method is not robust enough in the case of heavy tailed distributions. CART is better than CHITES. This is because CART uses the exhaustive search method and it should not be biased in these equal distribution simulation designs.

### 3.3 Comparison in Error Rate

In the previous two subsections, we investigated the biasness and power of variable selection at the root node when only one variable is correlated with the target variable. But, in real data set many variables are informative to predict the class of target variables. Moreover, to compare the misclassification error rates in simulated data sets, we have to generate data sets in which more than one variable are correlated with the target variable. Otherwise, there may be no informative variable left to proceed splitting procedures after the first split.

We thus considered a simulation data set which is a modified form of the shifted mean model in Table 3. Among 5 predictor variables, $X_1$, $X_3$, and $X_4$ are correlated with the target variable as follows.

$$X_1 \sim \begin{cases} N(.5,1), & \text{for Class 1} \\ N(0,1), & \text{for Class 2} \end{cases}$$

Values of $P(X_3=j)$, for $j=1,2,3,4$

| Class | $j=1$ | $j=2$ | $j=3$ | $j=4$ |
|---|---|---|---|---|
| 1 | 0.35 | 0.25 | 0.25 | 0.15 |
| 2 | 0.15 | 0.25 | 0.25 | 0.35 |

Values of $P(X_4=j)$, for $j=1,2$

| Class | $j=1$ | $j=2$ |
|---|---|---|
| 1 | 0.60 | 0.40 |
| 2 | 0.40 | 0.60 |

The other two variables, $X_2$ and $X_5$, are the same as in Table 3 and they are uncorrelated with the target variable. Thus, $X_2$ is generated from the exponential distribution, and $X_5$ is a categorical variable which has a uniform distribution on $\{1,2,\cdots,M\}$. The number of categories in $X_5$ are two cases: $M=5$ and $M=15$.

For these sets of simulated data, the variable selection powers are also computed using 300 replicates, and results are tabulated in Table 6. CHITES and F&CHITES select $X_1$ and $X_3$, and no other variables. But, CART selects $X_2$ and $X_5$ in some cases.

**Table 6.** Estimated probabilities of variable selection when the predictors $X_1$, $X_3$ and $X_4$ have discriminatory power. Estimates are based on 300 replicates and training samples of size $N$ = 200 and 500.

| Algorithms | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | $M=5$, | | $N=200$ | | | $M=5$, | | $N=500$ | | |
| CART | .767 | .013 | .217 | .000 | .003 | .580 | .000 | .420 | .000 | .000 |
| CHITES | .227 | .000 | .773 | .000 | .000 | .137 | .000 | .863 | .000 | .000 |
| F&CHITES | .350 | .000 | .650 | .000 | .000 | .270 | .000 | .730 | .000 | .000 |
| | $M=15$, | | $N=200$ | | | $M=15$, | | $N=500$ | | |
| CART | .680 | .007 | .143 | .000 | .170 | .563 | .000 | .437 | .000 | .000 |
| CHITES | .283 | .000 | .717 | .000 | .000 | .150 | .000 | .850 | .000 | .000 |
| F&CHITES | .373 | .000 | .627 | .000 | .000 | .283 | .000 | .717 | .000 | .000 |

### Comparison with Simulated Data

To compare the misclassification error rates, 20 data sets, for each case of $M=5$ and $M=15$ with sample sizes $N=200$ and $N=500$, are generated using the simulation model used in Table 6. The algorithms to be compared are CART, CHITES, F&CHITES and QUEST. To apply the same conditions on each algorithm, we used the following stopping rule: At each node if the number of objects is less than 5% of the total sample size, then the splitting procedure is stopped.

To compare the test sample errors, 10 data sets are used as training sample and another 10 data sets are used to compute test sample errors. The results are shown in Table 7 ($N=200$) and Table 8 ($N=500$), respectively. The order of average misclassification error rates can be summarized as follows.

| | | | | |
|---|---|---|---|---|
| $N=200$, $M=5$ : | QUEST | CHITES | F&CHITES | CART |
| $N=200$, $M=15$ : | CHITES | F&CHITES | QUEST | CART |
| $N=500$, $M=5$ : | CHITES | QUEST | F&CHITES | CART |
| $N=500$, $M=15$ : | QUEST | F&CHITES | CHITES | CART |

The unbiased methods CHITES, F&CHITES and QUEST are better than CART in these simulation data sets.

**Table 7.** Test sample error rates of ten test data sets. The sample size of each test sample is 200. The averages of error rates are given in the last row.

| $M=5$ | | | | $M=15$ | | | |
|---|---|---|---|---|---|---|---|
| CART | CHITES | F&CHITES | QUEST | CART | CHITES | F&CHITES | QUEST |
| .405 | .400 | .380 | .355 | .430 | .405 | .395 | .400 |
| .415 | .450 | .420 | .370 | .425 | .385 | .370 | .360 |
| .395 | .340 | .445 | .330 | .410 | .365 | .415 | .485 |
| .350 | .435 | .480 | .465 | .475 | .430 | .385 | .375 |
| .485 | .390 | .400 | .425 | .430 | .380 | .370 | .335 |
| .460 | .365 | .370 | .405 | .440 | .425 | .415 | .410 |
| .500 | .475 | .405 | .430 | .415 | .345 | .395 | .410 |
| .395 | .390 | .400 | .380 | .365 | .370 | .415 | .390 |
| .425 | .445 | .465 | .370 | .400 | .380 | .320 | .385 |
| .430 | .350 | .365 | .385 | .355 | .380 | .405 | .395 |
| .426 | .404 | .413 | .392 | .415 | .387 | .389 | .395 |

**Table 8.** Test sample error rates of ten test data sets. The sample size of each test sample is 500. The averages of error rates are given in the last row.

| $M=5$ | | | | $M=15$ | | | |
|---|---|---|---|---|---|---|---|
| CART | CHITES | F&CHITES | QUEST | CART | CHITES | F&CHITES | QUEST |
| .390 | .380 | .364 | .368 | .440 | .362 | .398 | .368 |
| .380 | .348 | .376 | .338 | .402 | .384 | .396 | .358 |
| .380 | .380 | .362 | .366 | .398 | .372 | .366 | .364 |
| .368 | .386 | .368 | .348 | .376 | .390 | .366 | .356 |
| .380 | .352 | .360 | .364 | .442 | .370 | .358 | .364 |
| .414 | .418 | .452 | .380 | .410 | .400 | .418 | .368 |
| .390 | .352 | .344 | .430 | .412 | .430 | .378 | .432 |
| .384 | .372 | .366 | .382 | .440 | .344 | .348 | .382 |
| .400 | .364 | .392 | .394 | .412 | .416 | .408 | .386 |
| .408 | .362 | .378 | .372 | .388 | .400 | .350 | .372 |
| .389 | .371 | .376 | .374 | .412 | .387 | .379 | .375 |

### Comparison with Real Data

The four algorithms, CART, CHITES, F&CHITES and QUEST, are also compared using real data sets in Table 1. These algorithms use stopping rules and do not perform any pruning. The objects with missing values are deleted from the data sets. 10-fold cross-validation method is used to estimate the misclassification error rates. The results are summarized in Table 9.

Table 9. 10-fold cross-validation error rates of CART, CHITES, F&CHITES and QUEST. The figures indicate error rates ± 1 SE.

| data set name | CART | CHITES | F&CHITES | QUEST |
|---|---|---|---|---|
| balance | .235 ±.017 | .243 ±.017 | .237 ±.017 | .267 ±.018 |
| bupa | .395 ±.026 | .371 ±.026 | .368 ±.026 | .368 ±.026 |
| german | .341 ±.015 | .302 ±.015 | .342 ±.015 | .293 ±.014 |
| glass | .047 ±.014 | .042 ±.014 | .056 ±.016 | .051 ±.015 |
| iris | .127 ±.027 | .113 ±.026 | .120 ±.027 | .140 ±.028 |
| ionosphere | .177 ±.020 | .103 ±.016 | .131 ±.018 | .160 ±.020 |
| new-thyroid | .168 ±.025 | .149 ±.024 | .144 ±.024 | .135 ±.023 |
| pima indian | .361 ±.017 | .278 ±.016 | .275 ±.016 | .266 ±.016 |
| wine | .197 ±.030 | .214 ±.031 | .219 ±.031 | .270 ±.033 |

For the nine data sets, the order of misclassification error rates can be summarized as follows.

|  |  |  |
|---|---|---|
| balance | : | CART F&CHITES CHITES QUEST |
| bupa | : | F&CHITES QUEST CHITES CART |
| german | : | QUEST CHITES CART F&CHITES |
| glass | : | CHITES CART QUEST F&CHITES |
| iris | : | CHITES F&CHITES CART QUEST |
| ionosphere | : | CHITES F&CHITES QUEST CART |
| new-thyroid | : | QUEST F&CHITES CHITES CART |
| pima indian | : | QUEST F&CHITES CHITES CART |
| wine | : | CART CHITES F&CHITES QUEST |

As a whole, CHITES is very robust in terms of error rates. For example, in *german* data set CART and F&CHITES are significantly inferior to CHITES or QUEST, in *pima indian* data set CART is significantly worse than others, and in *wine* data set QUEST is significantly worse than others. But, CHITES is never significantly worse than other algorithms.

### 3.4 Comparison in Terms of Training Time

One of the disadvantages of using the exhaustive search method is that it takes too much time in computing if there are huge split candidates. The proposed methods separate the variable selection step and split-point selection step, and therefore they are expected to save computing times.

The four algorithms, CART, CHITES, F&CHITES and QUEST are compared in terms of training time for the nine data sets in Table 1. The function *date*() of S-Plus is used to count the training time. The results are listed in Table 10.

The data set *ionosphere* has 34 numeric variables and for this data set CART is significantly slow than other methods. On the average CART is the slowest, and the proposed methods are faster than others.

**Table 10.** S-PLUS 2000-equivalent training time and relative times of the algorithms. 's' and 'm' denote seconds and minutes, respectively.

| data set name | CART | CHITES | F&CHITES | QUEST |
|---|---|---|---|---|
| balance | 1.02m | 24s | 29s | 47s |
| bupa | 1.41m | 33s | 43s | 1.02m |
| german | 19.24m | 6.53m | 2.23m | 8.15m |
| glass | 1.07m | 17s | 15s | 12s |
| iris | 12s | 6s | 6s | 9s |
| ionosphere | 25.28m | 16s | 1.16m | 3.18m |
| new-thyroid | 40s | 12s | 13s | 15s |
| pima indian | 5.52m | 1.17m | 1.51m | 1.38m |
| wine | 1.55m | 30s | 25s | 27s |

## 4. Conclusion

In order to investigate the bias problem in exhaustive search algorithm such as CART (Breiman et al., 1984), we propose two methods which separate the variable selection step and the split-point selection step. We call these two algorithms as CHITES method and F&CHITES method. They adapted some of the best characteristics of CART, CHAID (Kass, 1980), and QUEST (Loh and Shih, 1997). In the first step the variable, which is most significant to predict the target class values, is selected. In the second step, the exhaustive search method is applied to find the splitting point based on the selected variable in the first step.

We compared the proposed methods, CART, and QUEST in terms of variable selection bias and power. The proposed methods are not only unbiased in the null case, but also powerful for selecting correct variables in non-null cases. The unbiasedness is achieved by separating the variable selection step and the split-point selection step.

The proposed methods, CART and QUEST are compared in terms of error rates on simulated data sets and also on real data sets. The accuracy of the tree is assessed either by test sample or cross-validation. The unbiased methods CHITES, F&CHITES and QUEST are better than CART in simulated data sets. In real data sets, as a whole, CHITES is very robust in terms of worst error rates. QUEST employs the quadratic discriminant method which heavily depends on the normality assumption. But the proposed methods apply the exhaustive search method based on Gini index which is a nonparametric method, once the most significant variable is selected. This seems to be the main reason why the proposed methods are robust in terms of error rates.

The proposed methods are also compared with CART and QUEST in terms of training time. The results show that the proposed methods are faster than other methods, in general.

In this paper, we confirmed that the proposed methods are compatible with CART and QUEST, even though they use simple methods in selecting split variables. Thus, it is worthy to improve the proposed algorithms CHITES and F&CHITES.

## References

[1] Blake, C.L. and Merz, C.J. (1998). UCI repository of machine learning databases, Department of Information and Computer Science, Irvine, CA: University of California, (http://www.ics.uci.edu/$\thicksim$ mlearn/MLRepository.html).

[2] Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and Regression Trees*, New York: Chapman and Hall.

[3] Kass, G.V. (1980). An Exploratory technique for investigating large quantities of categorical data, *Applied Statistics*, Vol. 29, 119-127.

[4] Kim, H. and Loh, W.Y. (2001). Classification trees with unbiased multiway splits, *Journal of the American Statistical Association,* Vol. 96, 589-604.

[5] Lee, Y.M. (2002). *A Study on Bias Problems in Constructing Classification Trees*, Ph.D. Thesis, Department of Statistics, Seoul National University

[6] Loh, W.Y. and Shih, Y.S. (1997). Split selection methods for classification trees, *Statistica Sinica,* Vol. 7, 815-840

[7] Loh, W.Y. and Vanichsetakul, N. (1988). Tree-structured classification via generalized discriminant analysis, *Journal of the American Statistical Association,* Vol. 83, 715-728.

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

[9] Song, M.S. and Yoon, Y.J. (2000). A comparable study on variable selection methods in data mining software packages, *Proceeding of the Tenth Japan and Korea Joint Conference of Statistics*, 125-130.

[ 2002년 7월 접수, 2002년 8월 채택 ]