Proceedings of the Autumn Conference in 2001
and the 30th Anniversary of the Korean Statistical Society

Robust Variable Selection in Classification Tree

Jeong Yee Jang[1], Kwang Mo Jeong[2]

Abstract

In this study we focus on variable selection in decision tree growing structure. Some of the splitting rules and variable selection algorithms are discussed. We propose a competitive variable selection method based on Kruskal-Wallis test, which is a nonparametric version of ANOVA F-test. Through a Monte Carlo study we note that CART has serious bias in variable selection towards categorical variables having many values, and also QUEST using F-test is not so powerful to select informative variables under heavy tailed distributions.

Key Words : Classification tree; Robust variable selection; CART; QUEST

1. Introduction

Researches on classification tree have been develped over the past two decades and it has been applied to several areas such as marketing, direct mail, credit approval, customer segmentation, fraud detection, maunfacturing, health care and so on. Standard classification tree algorithms recursively partition the data space with the aim of making the distribution of the class variable as homogeneous as possible within each partition. The main steps for constructing tree are data partitioning, tree simplification, and model fitting. Various tree growing algorithms such as CHAID(Kass, 1980), CART(Breiman, Friedman, Olshen and Stone, 1984), FACT(Loh and Vanichsetakul, 1988), C4.5(Quinlan, 1993), QUEST(Loh and Shih, 1997), and CRUISE(Kim and Loh, 2001) have been proposed.

An algorithm to select the split variable and the split point is necessary in the data partitioning. Variable selection is a very important step in constructing classification trees and we especially focus on variable selection procedure. According to researches up to now QUEST has negligible bias compared to the exhaustive

[1]Graduate student, Dept. of Statistics, Pusan National University, Pusan, 609-735, Korea
[2]Professor, Dept. of Statistics, Pusan National University, Pusan, 609-735, Korea

search method. If a predictor has different sample mean in each class, F-test could select the predictor but if it has different distributions having very similar means, the F-test may fail to select the proper predictor even though the predictor is informative. This is a defect of QUEST algortithm and hence we suggest a robust variable selection method based on nonparametric Kruskal-Wallis(KW) test, which is good both in bias and power under heavy tailed distributions. We compare two algorithms through a Monte Carlo simulation study.

2. Splitting Rules

Suppose that there is a candidate split for any node t which divides it into t_L and t_R. The cases in node t anwsering, for example, 'yes' go to the left descentdant node t_L and those answering 'no' to the right descendant t_R.

We briefly review three splitting rules; entropy measure, Gini index and χ^2-statistic, which are widely used in many tree growing softwares.

(i) Entropy Measure : $i(t) = -\sum_j p(j|t)log_2(p(j|t))$
We choose the split that minimizes entropy measure. This has been adopted by C4.5(Quinlan, 1993).

(ii) Gini Index : $i(t) = \sum_{i \neq j} p(i|t)p(j|t)$
We choose the split that minimizes Gini index of impurity measure. The Gini index is frequently used in classificaion tree such as CART(Breiman et al., 1984).

(iii) χ^2-statistic : $\chi^2 = \sum_{i,j} \frac{(f_{ij}-e_{ij})^2}{e_{ij}}$
Here e_{ij} is an expected frequency and f_{ij} is an observed frequency with i and j denoting category numbers of preditor variable. We select the split variable and the point which has the smallest p-value based on χ^2-statistic.

3. Variable Selection Algorithms

3.1 QUEST Algorithm

The exhaustive search method such as CART searches all possible candidate splits and chooses the variable which minimizes the impurity measure at each

node. In the exhaustive search method, a continuous predictor with n distinct values gives n-1 candidate splits, and a categorical predictor with k categories gives 2^{k-1}-1 candidates. Then, if one predictor variable is continuous and the other is categorical variable with many categories, the categorical variable has greater chance to be selected and the exhaustive search method is biased toward the categorical variable.

On the other hand the idea of QUEST algorithm is to compute p-values both from ANOVA F-test for continuous variables and from χ^2-test for categorical vaiables. We select the variable which has the smallest p-value for all variables. Lastly the Levene's F-test is performed to to deal unequal variances between target groups. We repeat the QUEST algoritm in the following. Hereafter we assume that $X_1, \ldots, X_{K_1}$ are continuous variables and $X_{K_1+1}, \ldots, X_K$ are categorical variables.

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 $\hat{\alpha}_1 = Pr\left\{F_{J_t-1,N(t)-J_t} > F_{k_1}\right\}$, 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 $\hat{\beta}(k)$ for the χ^2-test of independence for each categorical variables X_k, $k = K_1 + 1, \ldots, K$. Let k_2 be the smallest integer such that $\hat{\beta}(k_2) = min\left\{\hat{\beta}(k) : k = K_1 + 1, \ldots, K\right\}$ 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 $\min(\hat{\alpha}_1, \hat{\alpha}_2) < \alpha/K$, where α is a pre-specified significance level, 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-statistic $F_k{}^{(z)}(k = 1, \ldots, K_1)$ for the ordered variables based on the absolute deviations $z_{ij}^{(j)} = \mid x_{ik}^{(j)} - \bar{x}_k^{(j)} \mid$, where $\bar{x}_k^{(j)} = N_j(t)^{-1} \sum_{i=1}^{N_j(t)} x_{ik}^{(j)}$.
   (b) Compute $\tilde{\alpha} = Pr\left\{F_{J_t-1,N(t)-J_t} > F_{k''}{}^{(z)}\right\}$, where k'' denotes the smallest index having maximum of $F_k{}^{(z)}$. If $\tilde{\alpha} < \alpha/(K + K_1)$, select variable k'' to split the node. Otherwise, select variable $X_{k'}$.

3.2 Nonparametric Test-Based Variable Selection

As a competitive to QUEST algorithm we propose a nonparametric version of F-test to improve the power of selecting informative variables when the predictors have heavy tailed distributions. The idea is to compute p-values from nonparametric KW-test for continuous variables, and χ^2-test for categorical variables. We select the variable with the smallest p-value among all variables. We don't need the last step of Levene's test in QUEST algorithm and hence the computational burden is alleviated.

A nonparametric variable selection algorithm based on KW-test can be written as following.

1. If $K_1 \geq 1$, compute the KW statistic KW_k for each X_k, $k = 1, \ldots, K_1$. Let k_1 be the smallest integer such that $KW_{k_1} = max\{KW_k : k = 1, \ldots, K_1\}$ and define $\hat{\alpha}_1 = Pr\left\{KW_{J_t-1,N(t)-J_t} > KW_{k_1}\right\}$, where KW_{ν_1,ν_2} denotes the KW 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 χ^2 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 $\hat{\beta}(k_2) = min\left\{\hat{\beta}(k) : k = K_1 + 1, \ldots, K\right\}$ and define $\hat{\alpha}_2 = \hat{\beta}(k_2)$.

3. Define $k' = k_1$ if $\hat{\alpha}_1 \leq \hat{\alpha}_2$; otherwise define $k' = k_2$.

3.3 A Monte Carlo study

Table 3.1 Estimated probabilities of variable seletion under Null

# of class J	N	X_i	Distribution	QUEST	*The proposed method*
2	400	x_1	$N(0,1)$	0.097	0.102
		x_2	$U(0,1)$	0.100	0.100
		x_3	$Exp(1)$	0.120	0.100
		x_4	$Poi(1)$	0.100	0.092
		x_5	$Ord.U_4$	0.107	0.107
		x_6	$Cat.U_2$	0.095	0.102
		x_7	$Cat.U_4$	0.098	0.101
		x_8	$Cat.U_6$	0.092	0.096
		x_9	$Cat.U_8$	0.097	0.101
		x_{10}	$Cat.U_{10}$	0.095	1.000
4	400	x_1	$N(0,1)$	0.089	0.096
		x_2	$U(0,1)$	0.104	0.103
		x_3	$Exp(1)$	0.112	0.097
		x_4	$Poi(1)$	0.104	0.097
		x_5	$Ord.U_4$	0.089	0.087
		x_6	$Cat.U_2$	0.105	0.109
		x_7	$Cat.U_4$	0.092	0.099
		x_8	$Cat.U_6$	0.108	0.110
		x_9	$Cat.U_8$	0.100	0.101
		x_{10}	$Cat.U_{10}$	0.099	0.103

Table 3.2 Estimated probabilities of variable seletion under Alternative

# of class J	N	*Distribution of* X_0	QUEST	*The proposed method*
2	400	$N(.3,1)$, $N(0,1)$	0.834	0.830
		$N(.5,1)$, $Cauchy(1)$	0.602	0.927
		$Exp(1)$, $Cauchy(1)$	0.592	1.000
		$U(0,1)$, $Cauchy(1)$	0.614	1.000
		$Ord.U_4$, $Cauchy(1)$	0.592	1.000
4	400	$N(.5,1)$, $Exp(1)$, $U(0,1)$, $Cauchy(1)$	0.817	0.994
		$N(.3,1)$, t_2, $U(0,1)$, $Cauchy(1)$	0.795	0.965
		$N(0,1)$, $N(1,3)$, $Exp(1)$, $Cauchy(1)$	0.753	1.000

4. Conclusions

The results show that the proposed method is good both in bias and power when the predictors have heavy tailed distributions, and also we alleviate computational burden by dropping the step of Levene's test required in QUEST algorithm. QUEST is originally designed to detect unequal class means in the continuous variables, and so QUEST based on the ANOVA F-test and Levene's test has less power than KW-test. The proposed method isn't behind QUEST with respect to the power under normal distributions, moreover it is more powerful than QUEST under heavy tailed distributions.

REFERENCES

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

[2] Quinlan, J.(1993). *C4.5: Programs for Machind Learning*, Morgan Kaufmann, San Mateo.

[3] Loh W.Y. and Shih Y.S.(1997). Split Selection Methods for Classification Trees, *Statistica Sinica*, **7**, 815-840.

[4] Loh, W.Y. and Vanichsetakul, N.(1998). Tree-Structured Classification via Generalized Discriminant Analysis, *Journal of the American Statistical Association*, **83**, 715-728.

[5] Rajeev R. and Shim K.S.(1998). PUBLIC: A Decision Tree Classifier that Integrates Building and Pruning, *Data Mining And Knowledge Discovery Journal*, **4**(4).

[6] Shih, Y.S.(1999). Families of Splitting Criteria for Classification Trees, *Statistics and Computing*, **96**, 309-315.

[7] Shih, Y.S.(2001). Selecting the Best Splits for Classification Trees with Categorical Variables, *Statistics and Probability Letters*, In press.

[8] Kim, H. and Loh, W.-Y.(2001). Classification Trees with Unbiased Multiway Splits, *Journal of the American Statistical Association*, **96**, 589-604.