Variable selection for decision trees with mixed responses

Gong, Yi-Hung
National Chung Cheng University, Department of Mathematics
Chiayi 621, Taiwan
E-mail: yihung0214@hotmail.com

Shih, Yu-Shan
National Chung Cheng University, Department of Mathematics
Chiayi 621, Taiwan
E-mail: yshih@math.ccu.edu.tw

Introduction

Classification and regression tree is a method for data mining. Its data analysis results are presented in a tree-structured format which is intuitively appealing. This nature makes the method a favor choice among practitioners.

Three basic elements for constructing classification and regression trees are split selection, model fitting and model selection. The usual approach selects the splitting variables using so called the exhaustive search method. This approach is implemented in methods, like CART (Breiman, Friedman, Olshen & Stone 1984) and C4.5 (Quinlan 1993). However, this exhaustive search approach has been demonstrated to have selection bias toward variables with more splitting points and/or missing values (see, for example, Loh & Shih (1997) and Loh (2002)). To avoid the selection bias, the approach which separates the issue of variable selection from that of split point selection is proposed. For model fitting methods at each node, constant models are mostly used (Breiman et al. 1984, Quinlan 1993). To increase the flexibility of tree models, simple linear or higher order linear models are considered at each node (Loh 2010). Methods, like QUEST (Loh & Shih 1997), CRUISE (Kim & Loh 2001), and GUIDE (Loh 2002, Loh 2009) are proved to be more reliable in assessing and explaining the resulting trees.

Most of the classification and regression trees deal with univariate response. The first attempt to develop regression tree method handling multivariate responses can be traced back to Gillo & Shelly (1974) who extended the AID method. Recently, several multivariate classification and regression tree methods emerge from the literature. For example, Zhang (1998), Siciliano & Mola (2000), Kim, Kim & Lee (2003) and Noh, Song & Park (2004) propose multivariate classification trees. On the other hand, De'ath (2002), Larsen & Speckman (2004), and Lee (2005) study multivariate regression trees. Most of them can be treated as multivariate versions of CART. Among them, the exhaustive search principle is still used as the default method of variable selection (Zhang 1998, Siciliano & Mola 2000, Kim et al. 2003, De'ath 2002, Larsen & Speckman 2004). At each node, they usually fit a piece-wise constant model to the multivariate data. Thus, they constantly ignore possible correlation between the responses. After data are partitioned recursively, a pruning method (model selection method) is decided to choose the best subtree. The aforementioned methods often adapt CART's pruning method (Breiman et al. 1984) or some ad hoc direct stopping rules. Moreover, Lee & Shih (2006) and Hsiao & Shih (2007) have showed that the exhaustive search method based on various criteria in the multivariate trees still selects variables with more splitting points.

Recently, Dine, Larocque & Bellavance (2009) propose a tree method for mixed responses. Their approach to variable selection follows exhaustive search principle. Therefore, we believe that the method itself has selection bias and we show the evidences in this study. A statistical approach which extends that of Lee & Shih (2006) and Hsiao & Shih (2007) is proposed. We show that our

Table 1: Distributions of X variables used in the simulation studies. Z, E, D_{20}, C_5, and C_{10} are mutually independent; Z is a standard normal variable; E is an exponential variable with mean one; D_{20} is a uniformly distributed variable on the set $\{1, 2, ..., 20\}$; C_m denotes a $m-$category variable taking values $\{1, 2, ..., m\}$ with equal probabilities; U is a uniform variable over (0,1).

	Independent	Weakly Dependent	Strongly Dependent
X_1	Z	$Z + E + D_{20}$	$E + 0.1Z$
X_2	E	E	E
X_3	D_{20}	D_{20}	D_{20}
X_4	C_5	$\lfloor UC_{10}/2 \rfloor + 1$	$\lfloor UC_{10}/2 \rfloor + 1$
X_5	C_{10}	C_{10}	C_{10}

proposed method is relatively unbiased in split selection and thus is more reliable in explaining the resulting trees.

Selection methods

Dine et al. (2009) propose a tree-structured method for a mixture of continuous and categorical outcomes, MTMO. The method considers splits of the form $X \leq C$ for some constant C, if X is an ordered variable or $X \in A$ for some subset A, if X is categorical variable. Their method adapt the exhaustive search principle to choose the splitting variable. That is, it searches through all possible covariates (X) and their split points (C) or sets (A). The split which partitions t into t_L and t_R with the maximum $i(t) - i(t_L) - i(t_R)$ value is the one actually channels data into two children nodes. Function $i(t)$ denotes node impurity which is defined using the likelihood function of the responses at node t (Dine et al. 2009, p. 3797). Dine et al. (2009, p. 3798) show that some known impurity functions are special cases of the proposed function.

On the other hand, for multivariate continuous or categorical response, it has been shown that the methods which use the exhaustive search principle tend to select variables with more splits when the response is independent of the covariates (Lee & Shih 2006, Hsiao & Shih 2007). In other words, the methods have selection bias. In order to avoid this possible bias, we modify the algorithms of Lee & Shih (2006) and Hsiao & Shih (2007) to accommodate mixed responses. It utilizes conditional independence tests based on the hierarchical loglinear model for three way contingency tables for each covariate and we denote our method, CIT.

Simulation studies

In all the experiments, five covariates including three ordered variables (X_1, X_2, and X_3) and two categorical variables (X_4 and X_5) are considered and their distributions are shown in Table 1. For each method, the estimated probabilities of variable selection are recorded in 1,000 iterations with 500 random samples in each iteration. The response vector $\boldsymbol{Y}$ has two components Y_1 and Y_2 where Y_1 is a continuous random variable and Y_2 is a categorical variable.

In the following simulation study, we assume the response vector $\boldsymbol{Y} = (Y_1, Y_2)$ is independent of the X's covariates. Furthermore, let Y_2 be a binary random variable with $\Pr(Y_2 = 0) = \Pr(Y_2 = 1)$. Let Z_μ be a normal random variable with mean μ and variance 1 and Z_μ's are independent. The conditional distribution of Y_1 given Y_2 has the following distribution.

$$\text{(Distribution A)} \quad \begin{array}{l} Y_1|(Y_2 = 0) \sim (1-\omega_1)Z_0 + \omega_1 Z_1, \\ Y_1|(Y_2 = 1) \sim \omega_1 Z_0 + (1-\omega_1)Z_1, \end{array}$$

where the value of ω_1 is set so that the correlation coefficient between Y_1 and Y_2, ρ, is equal to $-0.4, 0.0$, or 0.4, respectively. Estimated probabilities of selecting each covariate are recorded for the CIT and the MTMO method. Since the covariates are independent of the response vector, each covariate shall have 1/5 chance of being selected. A selection method which processes this property is called an *unbiased* method.

Figure 1 shows bar graphs of the estimated selection probabilities for the two methods. We find that the MTMO method tends to select variables with more split points. On the other hand, the proposed method, CIT, selected each variable all within 3 standard error of 0.2. Thus, the CIT method is relatively unbiased.

We then conduct the following simulation study where the response vector is related to some covariates. An effective selection method shall be able to choose the correct variable(s) with higher probabilities. We simulate various dependent structures according to the models given in Table 2. The estimated probabilities of selecting the right covariate(s) are recorded for the CIT and MTMO method. By changing the values of parameter β or b while holding the value of the other parameter to be 0.05 , we obtain the power curves under various dependent models and the plots are shown in Figure 2 to Figure 5.

In all these Figures, we observe that the CIT method almost always has higher probability of selecting the correct variable(s) than the MTMO method. Thus, the CIT method is more powerful than the MTMO method in choosing the correct variable(s) in these models.

Conclusion

We propose a new variable selection method for decision trees with mixed responses. The method relies on the tests of conditional independence among three ways contingency tables. Through simulations, we demonstrate that the method is relatively unbiased when the responses are independent of the covariates. Furthermore, it has more power in selecting the correct variables when the responses are related to some covariates.

Table 2: Models for power studies of the variable selection methods. Y_1 is a continuous variable and Y_2 is a binary random variable. The models are generated by using the mean function plus a standard normal error. The X's follow the independent structure in Table 1. The generated variables are $I_1 = I(X_3 > 10)$; $I_2 = I(Y_2 = 1)$; $I_3 = I(X_4 \in \{1, 2\})$; $I_4 = \text{sgn}(X_3 - 10.5)$; and $W = X_1 + I_1 + X_1 I_1$.

Model	$\text{E}(Y_1)$	$\text{E}(logit(Pr\{Y_2 = 1\}))$
I	$0.5 + bI_1$	$0.5 + \beta I_1$
II	$0.5 + bI_1 + 0.5I_2$	$0.5 + \beta I_1$
III	$0.5 + bI_3 + 0.5I_2$	$0.5 + \beta I_4$
IV	$0.5 + bW + 0.5I_2$	$0.5 + \beta I_4$

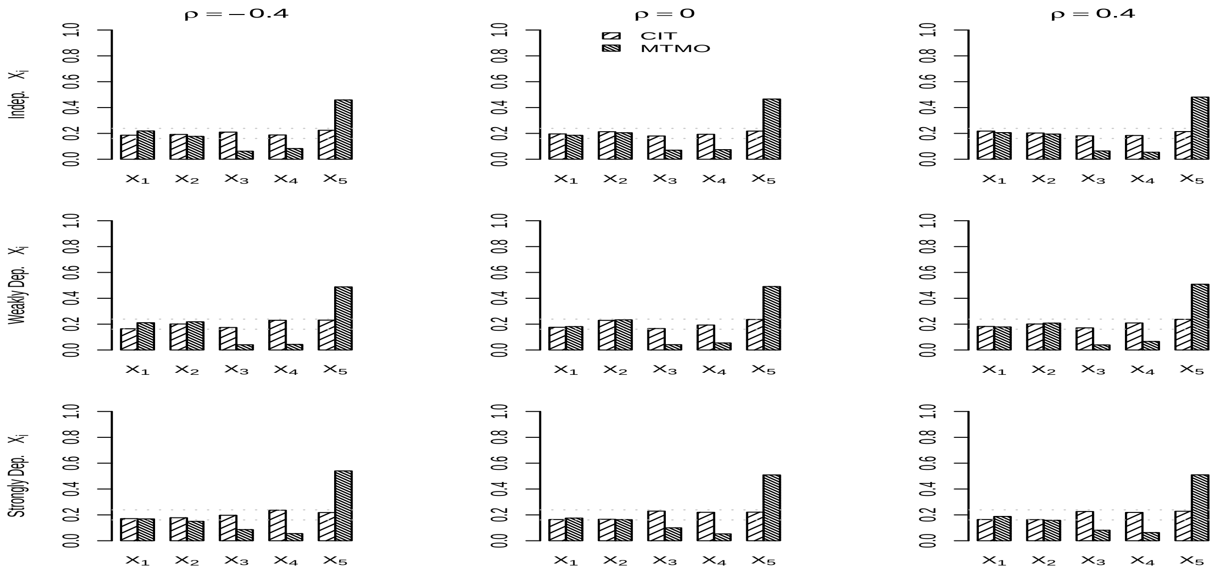

Figure 1: Estimated probabilities of variable selection for the CIT and MTMO method for constant fit where $\boldsymbol{Y}$ is independent of the X's. The distribution of $\boldsymbol{Y}$ follows Distribution A. The distributions of X's are given in Table 1. The simulation standard errors is about 0.013.

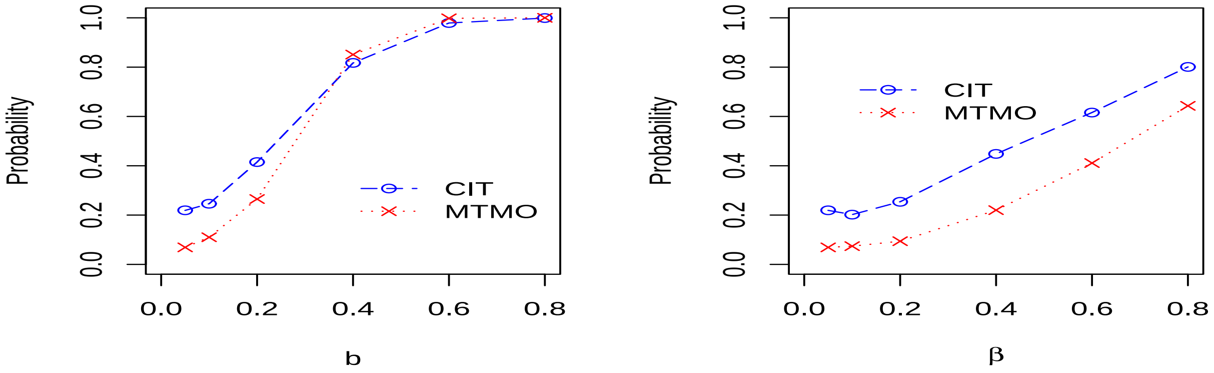

Figure 2: Estimated probability of X_3 is selected for the CIT and the MTMO method under Model I where one parameter (b or β) varies and the other is fixed at 0.05.

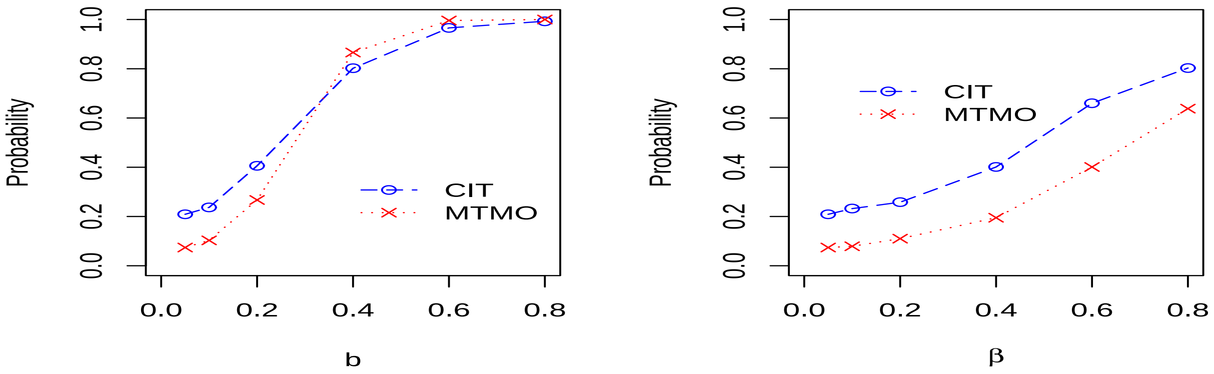

Figure 3: Estimated probability of X_3 is selected for the CIT and the MTMO method under Model II where one parameter (b or β) varies and the other is fixed at 0.05 .

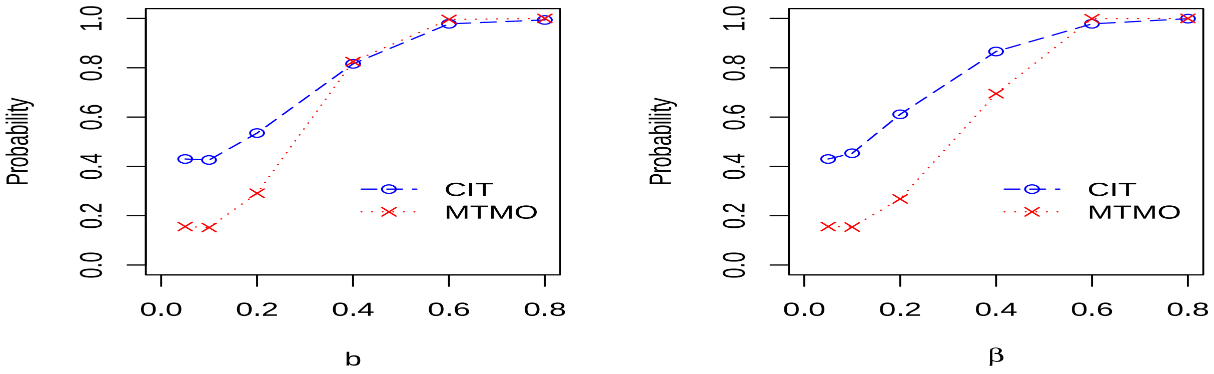

Figure 4: Estimated probability of X_3 or X_4 is selected for the CIT and the MTMO method under Model III where one parameter (b or β) varies and the other is fixed at 0.05.

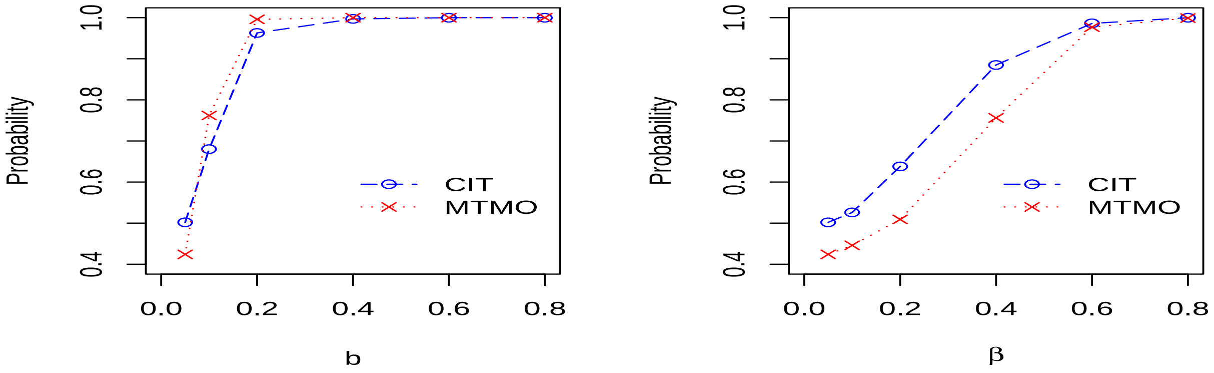

Figure 5: Estimated probability of X_1 or X_3 is selected for the CIT and the MTMO method under Model IV where one parameter (b or β) varies and the other is fixed at 0.05.

References

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

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

Dine, A., Larocque, D. & Bellavance, F. (2009). Multivariate trees for mixed outcomes, *Computational Statistics & Data Analysis* **53**: 3795–3804.

Gillo, M. & Shelly, M. W. (1974). Predictive modeling of multivariable and multivariate data, *Journal of The American Statistical Association* **69**: 646–653.

Hsiao, W.-C. & Shih, Y.-S. (2007). Splitting variable selection for multivariate regression trees, *Statistics and Probability Letters* **77**: 265–271.

Kim, H. J. & Loh, W.-Y. (2001). Classification trees with unbiased multiway splits, *Journal of The American Statistical Association* **96**: 598–604.

Kim, S. J., Kim, H. J. & Lee, K. B. (2003). On tree-based classifications with multi-Response variables, *International Journal of Industrial Enginerring–Theory, applications and practice* **10**: 427–434.

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

Lee, S. K. (2005). On generalized multivariate decision tree by using GEE, *Computational Statistics & Data Analysis* **49**: 1105–1119.

Lee, T.-W. & Shih, Y.-S. (2006). Unbiased variable selection for classification trees with multivariate responses, *Computational Statistics and Data Analysis* **45**: 457–466.

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

Loh, W.-Y. (2009). Improving the precision of classification trees, *Annals of Applied Statistics* **3**: 1710–1737.

Loh, W.-Y. (2010). Classification and regression trees, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* . In press.

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

Noh, H. G., Song, M. S. & Park, S. H. (2004). An unbiased method for constructing multilable classification trees, *Computational Statistics and Data Analysis* **47**: 149–164.

Quinlan, J. R. (1993). *C4.5 : Program for Machine Learning*, Morgan Kaufmann, Los Altos, California.

Siciliano, R. & Mola, F. (2000). Multivariate data analysis and modeling through classification and regressions, *Computational Statistics and Data Analysis* **32**: 285–301.

Zhang, H. (1998). Classification trees for multiple binary response, *Journal of The American Statistical Association* **93**: 180–193.

RÉSUMÉ (ABSTRACT)

We propose a variable selection method for constructing decision trees with a mixture of categorical and continuous responses. Compared with other selection methods, our method is relatively unbiased and is more powerful in selecting the correct split variables. Moreover, Our method is computational efficient. Simulation results are given to demonstrate the strength of our method.