The Korean Communications
in Statistics Vol. 10 No. 3, 2003
pp. 971-980

A Study on Decision Tree for Multiple Binary Responses

Seong Keon Lee[1)]

Abstract

The tree method can be extended to multivariate responses, such as repeated measure and longitudinal data, by modifying the split function so as to accommodate multiple responses. Recently, some decision trees for multiple responses have been constructed by Segal (1992) and Zhang (1998). Segal suggested a tree can analyze continuous longitudinal response using Mahalanobis distance for within node homogeneity measures and Zhang suggested a tree can analyze multiple binary responses using generalized entropy criterion which is proportional to maximum likelihood of joint distribution of multiple binary responses. In this paper, we will modify CART procedure and suggest a new tree-based method that can analyze multiple binary responses using similarity measures.

Keywords : Decision Tree, Multiple binary responses, Similarity measures

1 Introduction

Due to the fast development of computer technology, a huge amount of data has been rapidly accumulated. This entails a new problem of data analysis for such data sets. Analysts and researchers are trying to find important "knowledge" from large databases using data mining. Decision tree as one of many data mining techniques is a popular approach for segmentation, classification and prediction by applying a series of simple rules. It has the advantage that researchers can easily understand and explain the results, since it is expressed by a tree structured diagram as a final output.

The landmark work of decision tree is the methodology of Breiman, Friedman, Olshen, and Stone (1984), who introduced classification tree for a univariate discrete/continuous response. There are various competing approaches to the work of Breiman et al. (1984), such as that of Hawkins and Kass (1982) and Quinlan (1992). These approaches are focused on the single response. But, in many clinical trials and marketing research problems, multiple responses are often observed on individual subject. For example, disease-relate symptoms and customer-patterns are usually correlated. But, most tree-based algorithms handle only one target variable at a time. So, if more than one correlated target responses are observed, they

1) PhD Student, Department of Mathematics, Chuo University, 1-13-27, Kasuga, Bunkyo-Ku, Tokyo, 112-8551, Japan.
E-mail : sklee@grad.math.chuo-u.ac.jp

can not give reasonable result to analysts.

Recently, some decision trees for multiple responses have been constructed by Segal (1992) and Zhang (1998). Segal (1992) suggested a tree can analyze continuous longitudinal response using Mahalanobis distance for within node homogeneity measures. Zhang (1998) suggested a tree can analyze multiple binary response using generalized entropy criterion which is proportional to maximum likelihood of joint distribution of multiple binary responses (Cox, 1972; Zhao and Prentice, 1990).

In this paper, we will show the decision tree method can be extended to multiple responses, by modifying the split function so as to accommodate multiple responses. In section 2, we will introduce and review multivariate trees for multiple responses. Then, we will suggest a new tree-based method that can analyze multiple binary responses using similarity measure, in section 3. This measure is a new homogeneous type for binary multiple target variable. Finally, using a well known data set, we will compare the performance of the homogeneous measures.

2 Decision tree for multiple responses

In this section, we will introduce some tree approaches for multiple responses. The basic idea underlying these approaches is to define new homogeneity/impurity measures for response vector in each node. In univariate case, for example CART, we usually use a gini-index or variance as impurity measures.

2.1 Generalized Entropy Index (Zhang, 1998)

Zhang (1998) proposed the generalized entropy index as a homogeneity measure. He defined a new splitting function and cost-complexity in order to extend classification trees for the analysis of multiple discrete responses. First, he shows how to generalize the univariate entropy criterion to the present situation making use of the log-linear model. He assumed that joint distribution of responses Y depends on the linear term and the sum of the second-order products of its components only. That is, he assumed that the joint probability distribution of Y is

$$f(y;\Psi,\theta) = \exp(\Psi' y + \theta w - A(\Psi,\theta)),$$

where

$\Psi = (\psi_{i1}, \cdots, \psi_{ik})', \Omega = (\omega_{i12}, \omega_{i13}, \cdots, \omega_{ik-1k}, \cdots, \omega_{i12\cdots k})'$: *Canonical parameter,*

$w = (y_{i12}, y_{i13}, \cdots, y_{ik-1k}, \cdots, y_{i12\cdots k})'$: *vector of two* and *higher way cross product of* y,

$A(\Psi, \Omega)$: *Normalizing constant,*

$\exp(A(\Psi, \Omega)) = \sum \exp(\Psi_i' y_i + \Omega_i' w_i)$: *Sum of over all* 2^n *all possible values of* y_i.

These ideas are from the foundation of parametric models used to fit multiple binary

responses (Cox, 1972; Zhao and Prentice, 1990). This form of equation suggested by Cox (1972) and he assumed that the joint distribution of Y depends on the linear terms and the sum of the second order products of its component only.

Using above equation, Zhang define the homogeneity of node t, as the maximum of the log-likelihood derived from this distribution, which equals

$$h_{GE}(t) = \sum_{i \in t} [\hat{\Psi} y_i + \hat{\theta} w_i - A(\hat{\Psi}, \hat{\theta})],$$

where $\hat{\Psi}$ and $\hat{\theta}$ may be viewed as the maximum likelihood estimates of Ψ and θ, respectively. Obviously, the homogeneity of node t can be defined by analogy. The node impurity $i(t)$ can be chosen as $-h_{GE}(t)$ if you will.

2.2 Mahalanobis Distance (Segal, 1992)

Segal (1992) modified and used regression trees to model continuous type longitudinal data. If continuity is not a major concern, regression tree provide a useful tool to stratify growth curves.

For any node t, let V be the within-node covariance matrix of the longitudinal responses and $\bar{y}(t)$ the vector of within-node sample averages of the responses, where is a vector of parameters that may depend on the node. Then, an obvious within-node impurity as measured by the least squares is

$$h_{MD}(t) = \frac{1}{n_t} \sum_{i \in node\ t} (y_i - \bar{y}(t))' V^{-1} (y_i - \bar{y}(t)).$$

3 Decision tree for multiple responses using Similarity measure

Up to now, we have introduced some homogeneity measures in the previous researches. Now, let us consider a non-parametric approach but not parametric approach. In the cluster analysis, individuals are grouped into some clusters by the distance measure or similarity measure. Similarity measures for binary variables are given as in the Table 1.

So, we can define a following criterion as a homogeneity measure using the similarity coefficients ; that is,

$$h_S(t) = \frac{\sum_{i<j} S_{ij}}{{}_{n_t}C_2}, \quad i = 1,2,3,\cdots,n_t,$$

where $S_{ij} = (a+d)/(a+b+c+d)$. Suppose that there exist individuals n_t in a node. Then, the number of combination within individuals is $\binom{n_t}{2}$, and the similarity in the node is

<Table 1> Similarity measure for individuals

		ndividual I	
		1	0
individual J	1	a	b
	0	c	d

i) $(a + d)/(a + b + c + d)$ ii) $a/(a + b + c)$
iii) $2a/(2a + b + c)$ iv) $2(a + d)/(2(a + d) + b + c)$
v) $a/(a + 2(b + c))$ vi) $a/(a + b + c + d)$

defined as the average of the S_{ij}. Hence, the best split toward a child node is determined by maximizing the weighted average of h_S in each child node

$$Max\left(\frac{n_{t_L}}{n} h_S(t_L) + \frac{n_{t_R}}{n} h_S(t_R)\right).$$

4 Application

To construct our methodology, Fortran language was used and use it to analyze application data at Fujitsu workstation of Unix(Solaris) environment . In the program, *"missing together strategy"* is used for missing covariate and the minimum number of subjects can be splitted is setted 30, that is about 1% of total subjects.

4.1 BROCS Data

Building-related occupant compliant syndrome (BROCS) is a nonspecific set of related syndromes of discomfort reported by occupants of buildings. The most common symptoms of BROCS include irritation of the eyes, nose, and throat, and headache. The analysis of these data is difficult because of their high dimensionality and the strong correlation structure among numerous health symptoms that characterize BROCS.

To enhance the understanding of BROCS, We analyze a subset of the data from a 1989 survey of 3,400 employees of the Library of Congress (LOC) and headquarters of the Environmental Protection Agency (EPA) in the United States. This data contain 22 predictors as the factor of BROCS and six binary responses including several symptoms.

4.2 Tree Construction for BROCS

In this study, we construct two trees. One is using generalized entropy index and the other is using similarity measure. When h_{GE}, generalized entropy index, is used as a node homogeneity, the initial tree has 63 nodes. Then, applying the cost function, we get a sequence of 26 nested optimal sub-trees. However, when h_S, the similarity measure, is

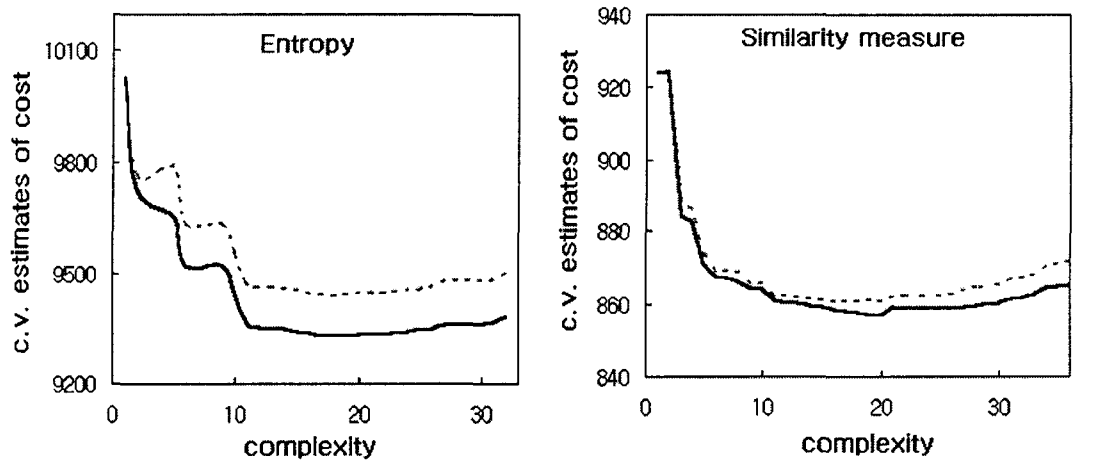
<Table 2> Predictors of BROCS

Variable	Question	Answer
X_1	Type of the working space	Enclosed office with door, cubicle without door, stacks, etc.
X_2	How is the working space shared?	Single, occupant, shared, etc.
X_3	Use of metal desk	Yes or no
X_4	Having new equipment	Yes or no
X_5	Allergic to pollen	Yes or no
X_6	Allergic to dust	Yes or no
X_7	Allergic to molds	Yes or no
X_8	Age	16-70 years old
X_9	Gender	Male or female
X_{10}	Is there too much air movement?	never, rarely, sometimes, often, always
X_{11}	Is there too little air movement?	the same as X_{10}
X_{12}	Is your work area too dry?	the same as X_{10}
X_{13}	Is the air too stuffy?	the same as X_{10}
X_{14}	Is your work area too noisy?	the same as X_{10}
X_{15}	Is your work area too dusty?	the same as X_{10}
X_{16}	Glare experience	No, sometimes, often, always
X_{17}	Comfortability of chair	Reasonably, somewhat, very uncomfortable, no one specific chair
X_{18}	Adjustability of chair	Yes, no, not adjustable
X_{19}	Influence over arranging the furniture	Very little, little, moderate, much, very much
X_{20}	Do you have children?	Yes or no
X_{21}	Do you have major childcare duties?	Yes or no
X_{22}	Type of job	Managerial, professional, technical, etc.

<Table 3> Responses and Symptoms of BROCS

Response	Cluster	Symptoms
y_1	CNS	Difficulty remembering/concentrating, dizziness, lightheadedness, depression, tension, nervousness
y_2	UA	Runny/stuffy nose, sneezing, cough, sore throat
y_3	PAIN	Aching muscles/joints, pain in back/shoulders/neck, pain
y_4	FLU	Nausea, chills, fever
y_5	EYES	Dry, itching, or tearing eyes; sore/strained eyes; blurry vision; burning eyes
y_6	LA	Wheezing in chest, shortness of breath, chest tightness

<Figure 1> Cost-complexity for two sequences of nested sub-trees.
Dashed line displays cross-validation estimates of cost, dotted line one SE above the estimated cost by cross-validation.



used, the initial tree has 71 nodes. Through the pruning procedure, we obtain a sequence of 33 nested sub-trees. To get a right sized tree, the sub-tree cost estimate and its standard error is derived from 10 repetitions of five-fold cross-validation, as given in figure 1. So, we can get the best sub-tree having 11 terminal nodes in the case of using generalized entropy index, h_{GE}, but 12 terminal nodes in the case of using similarity measure, h_S. We can see that similarity measure is more stable than generalized entropy index.

4.3 Results and Interpretations

Seeing the figure 2 and 3, we can find two approaches give us similar results. The tree result using generalized entropy index, figure 2, suggest that "Was air often too stuffy?" is the first split factor for symptoms and the number of terminal node is 10. The profiles of symptoms at terminal node are shown in the below of figure. They represent frequencies of symptoms.

Figure 3 is the decision tree result using similarity measure. The first classifier is "Was air often too stuffy?", too, but, figure 3 has differences with figure 2 at depth 3. And the terminal node profiles are also shown in the below.

Interpreting a part of results, let us consider the node number 12. In node number 12 has 153 respondents and suggest that if "Air of building was often too stuffy" and "Respondent had allergic to molds" and "Work area is not too dusty", then number 2 symptom and number 3 symptom are often occurred. Other nodes also can be interpreted like this.

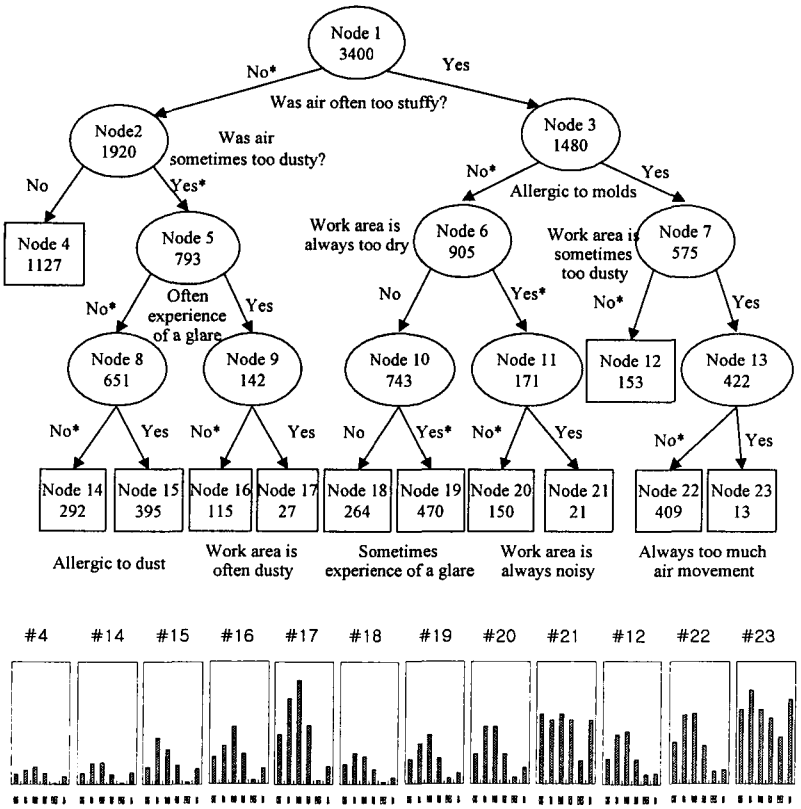
Though we have constructed trees applying different method, two approaches give us

<Figure 2> Decision tree using generalized entropy index and response rates.

<Table 4> Frequencies in the terminal nodes of the tree constructed by generalized entropy index

Terminal node No.	Node size	Cluster of symptoms					
		CNS	UA	Pain	Flu	Eyes	LA
8	667	48	79	96	64	7	55
9	460	46	54	62	34	4	16
10	651	76	186	154	81	11	74
13	76	11	28	25	15	4	25
14	634	190	321	322	177	58	60
15	73	23	46	34	22	11	29
16	71	21	30	42	20	0	12
17	71	16	25	35	22	5	7
18	305	62	75	95	42	12	8
20	295	52	103	124	61	11	22
21	97	28	30	41	28	0	11

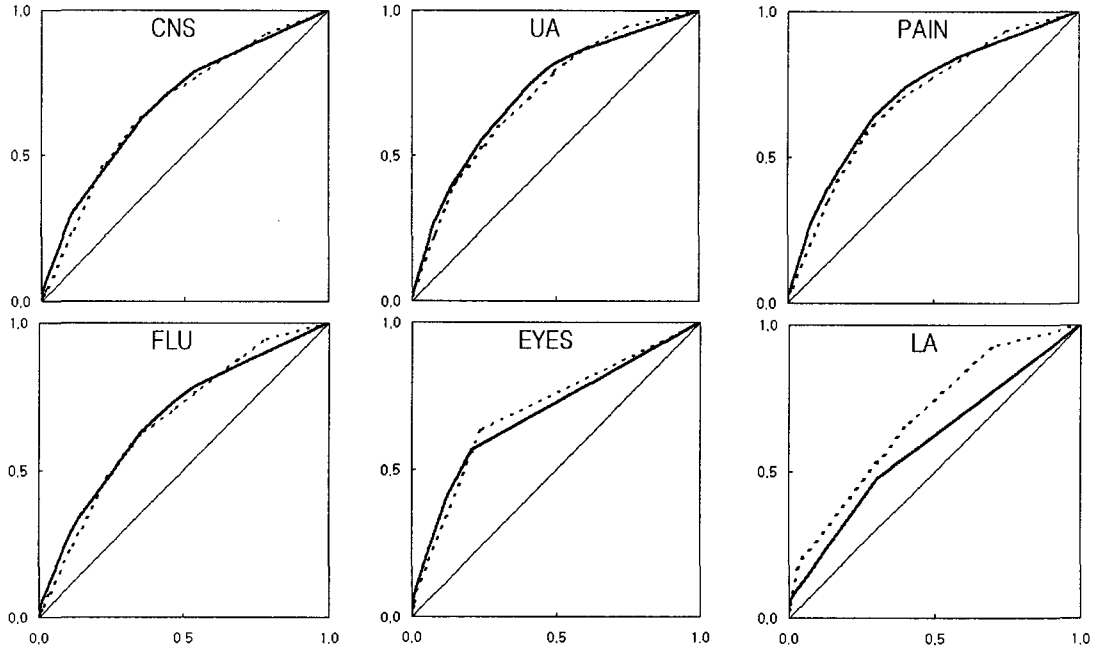
<Figure 3> Decision tree using similarity measure and response rates



<Table 5> Frequencies in the terminal nodes of the tree constructed by similarity measure

Terminal node No.	Node size	Cluster of symptoms					
		CNS	UA	Pain	Flu	Eyes	LA
4	1127	94	133	158	98	11	71
12	153	32	62	66	31	12	13
14	292	25	49	52	24	3	27
15	359	51	137	102	57	8	47
16	115	26	36	54	29	4	15
17	27	11	19	23	13	1	4
18	264	44	66	59	32	5	15
19	470	94	152	189	98	21	41
20	150	37	70	70	37	8	20
21	21	12	11	12	11	4	11
22	409	139	232	237	129	41	46
23	13	8	10	8	7	5	9

<Figure 4> ROC curves of each responses. Solid line indicates ROC curve of the tree constructed by similarity measure, and dotted line indicated that by generalized entropy index.



similar results. So, we will apply a traditional method, ROC (Receiver Operating Characteristic curves), to compare the predictive performance of two trees given before. A good predictions rule yields an ROC curve toward the northwest. Figure 4 displays good predictions for the two trees and presents similar effects in predicting the outcomes. Seeing the figure, in some case, the similarity measure is better, in others, the generalized entropy index is better.

5 Conclusion

In this work, we have introduced and presented several tree-based methods for the analysis of multiple binary responses. Comparing the similarity measure with the generalized entropy index using BROCS data, we had useful and reasonable results.

For the parametric method, difficulties arise when the dataset is large in both the number of observations and the number of variables, for example, the implementation for likelihood approaches requires severe computational effort. In such a sense, the similarity measure are competitive. Since this non-parametric approach is easily understandable, usable without assuming distribution and stable by the property of non-parametric method, we can construct a efficient tree for multiple binary responses.

REFERENCES

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

[2] Cox, D. R. (1972), The Analysis of Multivariate Binary Data, *Applied Statistics*, 21, 113-120.

[3] Diggle, P. J., Liang, K. Y., and Zeger, S. L. (1994), *Analysis of Longitudinal Data*, NewYork: Oxford Science Publications.

[4] Fizmaurice, G., and Laird, N. M. (1993), A Likelihood-Based Method for analyzing Longitudinal Binary Responses, *Biometrika*, 80, 141-151.

[5] Hawkins, D.M. and Kass, G.V., (1982), *Automatic Interaction Detection. In Topics in Applied Multivariate Analysis Hawkins, D. H., Ed.*; Cambridge University Press, pp.269-302.

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

[7] Segal, M. R. (1992), Tree-Structured Methods for Longitudinal Data, *Journal of the American Statistical Association*, 87, 407-418.

[8] Zhang, H. P., Holford, T., and Bracken, M. B. (1996), A Tree-Based Method of analysis for Prospective Studies, *Statistics in Medicine*, 15, 37-49.

[9] Zhang, H. P. (1998), Classification Trees for Multiple Binary Responses, *Journal of the American Statistical Association*, 93, 180-193.

[10] Zhao, L. P., and Prentice, R. L. (1990), Correlated Binary Regression Using a Quadratic Exponential Model, *Biometrika*, 77, 642-648.

[Received April 2003, Accepted October 2003]