Selecting variables for evidence-diagnosis of paralysis disease using CHAID algorithm

Yankyu Shin[1)]

Abstract

Variable selection in oriental medical research is considered. Decision tree analysis algorithms such as CHAID, CART, C4.5 and QUEST have been successfully applied to a medical research. Paralysis disease is a highly dangerous and murderous disease which accompanied with a great deal of severe physical handicap. In this paper, we explore the use of CHAID algorithm for selecting variables for evidence-diagnosis of paralysis disease. Empirical results comparing our proposed method to the method using Wilks λ are given.

keywords : selecting variable, CHAID algorithm, paralysis disease

1. INTRODUCTION

Decision tree analysis(SPSS Inc.(1998)) is a method for approximating discrete-valued target functions, in which the learned function is represented by a decision tree. Learned trees can also be re-represented as sets of if -then rules to improve human readability. Decision tree analysis is one of the most widely used and practical methods for inductive inference. Decision tree analysis is generally best suited to problems with the following characteristics: Instances are represented by attribute-value pairs. Disjunctive descriptions may be required. The training data may contain errors. The training data may contain missing attribute values.(Quinlan(1988)) Many oriental medical data is fit for these characteristics. We can therefore apply the decision tree analysis for classifying oriental medical patients by their evidence of disease. Since most oriental medical data contains many exploratory variables, the problem of variable selection is a particularly important problem in oriental medical research. A few important variables can result in a rapid proliferation of states, which in situations of limited sample information can cause difficulty in estimation. Weiner and Dunn(1966) discussed the methods of variable selection.

In this paper, we propose a variable selection method based on decision tree analysis algorithm for oriental medical data.

---

1) Associate Professor, Faculty of Information and Science, Kyungsan University, Kyungsan Kyungpook, 712-240

2. Main result and conclusion

We focus on paralysis disease data. The decision tree algorithm uses the CHAID algorithm. The proposed method are illustrated by using the paralysis disease patients data from Shin. etc.(1998) which have been collected to construct the expert system for the evidence-diagnosis of paralysis disease. Of a total of 125 cases, five evidence classes(A, B, C, D, E) were classified based on 37 exploratory variables.

From a stepwise variable selection based on discriminant analysis using wilks λ, 6 exploratory variables are selected from 6 steps. <Table 1> describes the misclassification rate based on above method under. The 45 cases are misclassified.

		Actual Category					
		A	B	C	D	E	Total
Predicted Category	A	21	3	3	2	0	29
	B	0	16	5	4	0	25
	C	0	0	10	3	1	14
	D	7	0	6	19	1	33
	E	1	1	1	7	14	24
	Total	29	20	25	35	16	125

<Table 1> Misclassification Matrix

<Table 2> gives the estimated risk and its standard error of misclassification based on the method using CHAID algorithm under 6 maximum tree depth. The 24 cases are misclassified.

Misclassification Matrix							
		Actual Category					
		A	B	C	D	E	Total
Predicted Category	A	27	4	6	2	0	39
	B	0	14	0	0	0	14
	C	2	2	19	2	1	26
	D	0	0	0	29	3	32
	E	0	0	0	2	12	14
	Total	29	20	25	35	16	125
		Resubstitution					
Risk Estimate		0.192					
SE of Risk Estimate		0.0352291					

<Table 2> Risk Estimate for 6th split

The method using the variable based on CHAID algorithm shows the smaller

misclassification rate than those based on Wilks λ. Furthermore, the decision tree analysis is robust to out-lier.

Reference

[1] Quinlan, J. R (1988). Decision trees and multi-valued attributes. In Hayes, Michie, & richards(Eds.), *Machine Intelligence 11*, 305-318. Oxford, England: Oxford University Press.

[2] SPSS Inc. (1988). *AnswerTree 2.0 User's Guide*, SPSS Inc., Chicago.

[3] Weiner, J. and O. J. Dunn, (1966). "Elimination of variates in linear discrimination problems," *Biometrics*, 22, 268-288.

[4] 신양규, 강효신, 권영규, 박창규, 김상철 (1998). 전문가시스템을 이용한 한의진단의 객관화에 관한 연구, 연구과제최종보고서, 보건복지부.