The Korean Communications
in Statistics Vol. 8, No. 3, 2001
pp. 823-829

A Decision Tree-based Analysis for Paralysis Disease Data[1)]

Yangkyu Shin[2)]

Abstract

Even though a rapid development of modern medical science, paralysis disease is a highly dangerous and murderous disease. Shin et al. (1998) constructed the diagnosis expert system which identify a type of the paralysis disease from symptoms of a paralysis disease patients by using the canonical discriminant analysis. The decision tree-based analysis, however, has advantages over the method used in Shin et al. (1998), such as it does not need assumptions - linearity and normality, and suggest appropriate diagnosis procedure which is easily explained. In this paper, we applied the decision tree to construct the model which identify a type of the paralysis disease.

Keywords : decision tree, diagnosis model, paralysis disease data

1. 서론

중풍(혹은 뇌졸중)은 현대의학의 발전에도 불구하고 아직도 높은 사망률과 신체장애를 유발하는 질환으로 한국인의 질병 사망원인에 있어 암에 이어 2위를 차지하고 있다. 그러므로 중풍환자의 정확한 진단과 이에 대한 치료는 매우 중요하다. 우리 나라에서는 많은 중풍환자들이 한방 치료를 선호하고 있다. 한방에서 중풍환자의 치료에 대한 우수한 임상효과가 있는데도 불구하고 결과분석 및 결론도출에 이르는 과정에서 객관적이며 타당한 방법 적용의 결여로 그 효용성을 극대화하지 못하고 있다. 그러나 1976년부터 1992년까지의 우리 나라 한의학박사학위 논문의 87.5%가 실험논문(박종운 (1993))으로 이루어지고 있듯이 최근 한의학분야에서는 한의학의 이론을 객관적으로 검증하고 심화시키기 위한 요구가 점차로 높아지고 있다. 중풍과 관련하여 강효신 외 4인 (1997)은 중풍의 임상적 결과에 대한 병리학적 연구 및 통계적 방법을 이용한 연구를 하였다. 신양규 외 4인 (1998)은 정준판별함수를 이용한 진단모형에 기초하여 중풍의 증형을 자동으로 진단하는 중풍진단전문가시스템을 구축하였다. 한의학에 있어서 진단은 환자로부터 주어진 정보(설명변수라 하기로 한다)에 의하여 증형(목표변수라 하기로 한다)을 변별하는 것이며 치료는 변별된 증형에 따라 치료법을 결정하고 처치를 하는 과정이다.(김완희 외 1인(1996)) 중풍의 주요 증형은 한열을 근거로 한증에 속하는 습담, 기혈구허와 열증에 속하는 화열, 담화, 음허양항으로 구분된다. 신양규 외 4인 (1998)의 연구에 사용한 분석법은 설명변수가 연속형 변수이고 증형별로 동일한 공분

1) This research was supported by a Grant from Kyungsan University, Kirin foundation, in 2000
2) Associate Professor, Faculty of Information and Science, Kyungsan University, Kyungsan Kyungpook, 712-240
E-mail : yks@kyungsan.ac.kr

산 행렬을 가져야 의미가 있으며 다변량 정규분포의 가정이 필요하기도 한다. 그러나 진단모형구축을 위한 자료분석에 사용된 37종류의 설명변수의 측도를 살펴보면 6종류만이 연속형 변수이고 나머지는 21종류는 순서형 변수 그리고 10종류는 명목형 변수로 31종류가 범주형 변수에 속함을 알 수 있다. 그러므로 본 연구에서는 많은 변수들이 범주형인 경우에 효과가 있는 의사결정나무를 이용한 중풍의 증형 진단모형을 제안하고자 한다.

본 연구에 활용할 자료는 1996년-1998년 동안 경산대학교부속 한방병원에서 보건복지부 보건의료기술연구개발과제 수행을 위해 수집된 125건의 임상자료이다. 자료에 대한 분석은 SPSS AnswerTree 2.0을 이용하였다.

2. 의사결정나무를 이용한 중풍의 증형 진단

2.1 자료

본 연구에 사용된 자료는 1996년부터 1998년 동안 경산대학교부속 한방병원에서 중풍진단전문가시스템구축을 위하여 중풍전문의에 의해 선정되어 수집되고 검증된 125명의 환자들의 자료로 목표변수인 증형과 37종류의 설명변수로 구성되어 있다. 목표변수는 명목형 변수로 (화열, 음허양항, 담화, 습담, 기혈구허)의 범주로 나뉘어져있고 설명변수는 6종류의 연속형 변수(신장, 몸무게, 나이, 고혈압, 저혈압, 맥박 수), 21종류의 순서형 변수(두통, 현운, 면색, 체형, 수면, 변조, 의식, 담성, 대변, 소변색, 소변기능, 한출, 설질1, 설질2, 설태색, 설태질1, 설태질2, 맥상1, 맥상2, 맥상3, 맥상4) 그리고 10종류의 명목형 변수 (성별, 가족력, 갈음, 신열, 사지냉, 수족심열, 오한발열, 골절통, 양상, 기운)로 이루어져있다. 순서형 변수와 명목형 변수들의 범주는 2 - 5개로 설정되어있다.

2.2 의사결정나무를 이용한 증형진단모형

의사결정나무는 의사결정규칙을 나무구조로 도표화하여 목표변수에 대한 분류와 예측을 수행하는 분석방법으로 나무구조에 의한 귀납적 추론 과정을 나무구조에 의해 표현하므로 분석과정이 쉽게 이해되고 설명할 수 있다는 장점을 가지고 있다.(최종후 외 3인 (1998), Choi et al. (2001) SPSS Inc. (1998)) 의사결정나무는 CHAID, CART, QUEST등의 알고리즘을 이용하여 의사결정나무의 구조를 형성하는데 이 때 분석의 목적과 자료의 구조에 따라 적절한 알고리즘을 선택한다. 본 연구에서는 목표변수의 범주가 5개이고, 범주가 2개 이상인 설명변수들이 많으므로 부모마디에서 자식마디들이 형성될 때 2개 이상의 분리가 일어날 수 있는 방법이 적합하다. 그러므로 다지분리(multiway split)를 수행하는 알고리즘인 CHAID(Chi-squared Automatic Interaction Detection)(SPSS Inc.(1998))를 이용하여 진단모형을 설정하고자 한다. 목표변수가 범주형이므로 카이제곱통계량을 분리기준으로 사용한다. 즉, 설명변수의 각 범주들의 병합에 대하여 자유도에 대한 카이제곱통계량의 값에 따라 p-값을 구한다. 그런데 이때 설명변수의 범주가 2개 이상인 경우에는 구하여진 p-값이 정확한 p-값보다 작게된다. 범주가 2개 이상인 변수를 포함하고 있는 본 연구에서는 정확한 p-값을 구하기 위하여 구하여진 p-값에 대하여 Bonferroni조정을 하였다.(

Bigg et al. (1991)) 목표변수의 범주가 5개이므로 r개의 범주로 나뉘는 순서형 설명변수에 대해서는 $_{5-1}C_{r-1}$을 명목형 설명변수에 대하여는

$$\sum_{i=0}^{r-1}(-1)^i \frac{(r-1)^4}{i!(r-i)!}$$

을 각 설명변수에 대하여 최종적인 병합에 의하여 얻어진 p-값에 곱하여 얻은 값을 조정된 p-값으로 한다. 위의 과정을 되풀이하여 모든 설명변수에 대하여 최적분리를 찾아내고 다시 이들 중에서 조정된 p-값이 가장 작은 설명변수의 최적분리에 의하여 자식마디를 결정한다.

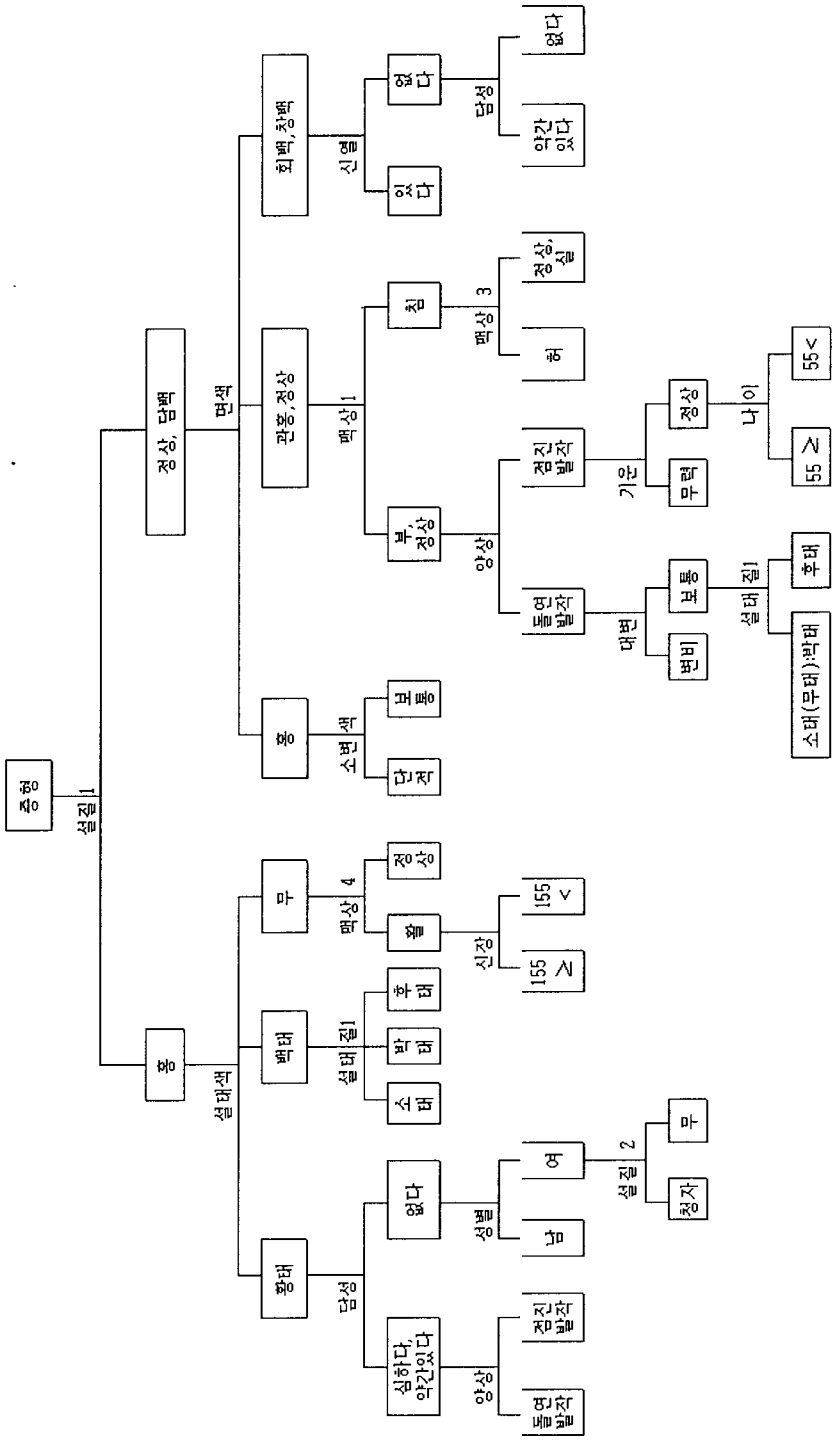
자료의 특성상 설명변수의 수가 많은 본 연구에서는 최적의 결과를 얻기 위하여 적절한 정지규칙의 설정이 필요하다. 의사결정나무에 의해 잘못 분류되거나 예측될 위험을 나타내는 위험 추정치(risk estimate (RE))를 기준으로 마디에 포함될 최소 관측자수와 자식마디가 형성될 때 각 자식마디에 포함될 최소 관측자수를 정하였다. (표 1)은 부모마디와 자식마디에 포함되는 최소 관측자수의 변화에 따른 위험 추정치 및 위험 추정치에 대한 표본오차이다.

부모마디	자식마디	RE	SE of RE
20	10	0.304	0.04114
20	5	0.304	0.04114
20	3	0.304	0.04114
20	1	0.304	0.04114
10	5	0.296	0.04083
10	3	0.264	0.03943
10	1	0.256	0.03903
5	3	0.224	0.03729
5	1	0.192	0.03523
3	1	0.192	0.03523

(표 1) 마디에 포함되는 관측자수에 따른 위험 추정치(RE)

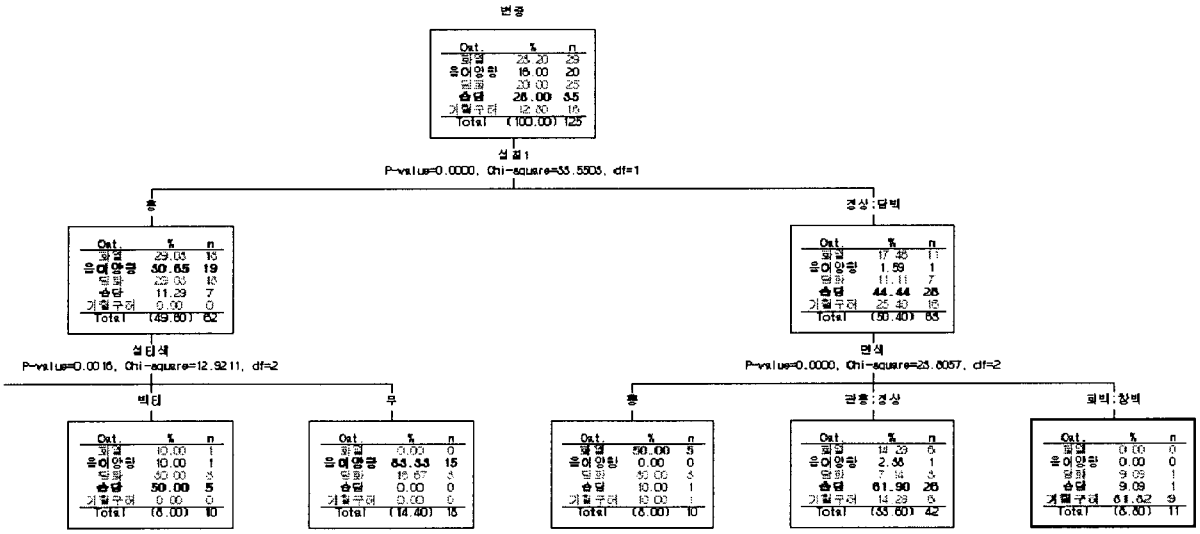
(표 1)에 의하면 부모마디와 자식마디에 포함되는 최소 관측자수는 5명과 1명인 경우가 위험 추정치 0.192로 가장 적합하다고 할 수 있다. 그러므로 (표 1)로부터 진단모형설정을 위한 CHAID 알고리즘에 적용할 부모 마디에 포함될 최소 관측자수는 5명, 자식마디가 형성될 때 각 자식마디에 포함될 최소 관측자수는 1명이다.

(그림 1)은 설정된 조건하에서 생성된 의사결정나무모형의 구조이다. (그림 1)에 의하면 증형 진단에 가장 중요한 기여를 하는 설명변수는 설질1 즉 설질의 색이고 다음으로 설태색, 면색, 그리고 담성, 설태질1, 맥상4, 소변색, 맥상1, 신열, 양상, 성별, 신장, 맥상3, 설질2, 대변, 기운, 나이이다. 37종류의 설명변수 중에서 나머지 20종류의 변수들은 증형 진단에 기여를 거의 못하고 있다고 할 수 있다.



(그림 1) 중풍증형진단 의사결정나무모형의구조

다음 (그림 2)는 생성된 의사결정나무의 일부인 첫 번째 분리결과와 두 번째 분리결과에 대한 출력결과이다.



(그림 2) 중풍증형진단에 대한 의사결정나무구조의 일부

		실 제 증 형					
		화 열	음허양항	담 화	습 담	기혈구허	전 체
예측증형	화 열	27	4	6	2	0	39
	음허양항	0	14	0	0	0	14
	담 화	2	2	19	2	1	26
	습 담	0	0	0	29	3	32
	기혈구허	0	0	0	2	12	14
	전 체	29	20	25	35	16	125
오 진 단 율		0.192					

(표 2) 오 진단 행렬 표

(그림 2)의 첫 번째 분리결과에 의하면 중풍의 증형 진단을 위한 첫 번째 변수는 설질1(설질색)이다. 설질의 색이 홍이면 음허양항증이 30.65%로 가장 두드러진 증형임을 볼 수 있다. 특히 열증에 속하는 화열, 담화, 음허양항의 증형이 설질의 색이 홍인 관측자 62명중 55명으로 전체의 88.71%를 차지하고 있음을 알 수 있다. 설질의 색이 정상이거나 담백하면 관측자의 63명중 44명

즉 69.84%가 한증에 속하는 습담, 기혈구허증으로 진단되고 특히 기혈구허증인 관측자 16명은 모두 설질이 정상이거나 담백으로 관측됨을 (표 2)의 첫 번째 마디로부터 볼 수 있다. 그러므로 설명변수 중에서 설질1이 증형 진단에 가장 큰 기여를 하며 특히 음허양항증과 기혈구허증의 진단에 기여를 한다고 볼 수 있다. 두 번째 분리 결과의 오른쪽 마디로부터 환자의 설질1이 정상이거나 담백하고 면색이 회백하거나 창백하다면 82%가 기혈구허증이라고 할 수 있다. 각 마디에 대하여 위와 같은 방법으로 증형 진단에 관한 추론을 한다.

위의 의사결정나무를 이용한 진단모형에 대한 오진단결과가 (표 2)이다. (표 2)에 의하면 실제 증형과 일치하게 진단된 관측자수는 125명중 117명으로 오진 율은 19.2%이다. 각 증형 별 오진 율은 화열증에 대하여 6.9%, 음허양항증에 대하여 30%, 담화증에 대하여 17.1%, 습담증에 대하여 11%, 기혈구허증에 대하여 25% 이다. 특히 (표 2)에서 보면 오 진단은 화열증이 담화증으로 음허양항증이 화열증이나 담화증등과 같이 인접한 범주로 대부분 오 진단됨을 알 수 있다. 즉 열증은 열증내에서 한증은 거의 한증내에서 오 진단이 이루어짐을 알 수 있다. 병증을 열증(화열, 음허양항, 담화)과 한증(습담, 기혈구허)으로 분류한다면 3명만이 오 진단되어 오진 율은 2.4%이다.

3. 결론 및 고찰

본 연구에서는 의사결정나무를 이용한 중풍의 증형진단모형을 제안하였다. 의사결정나무를 이용한 진단은 중풍의 증형이 진단되는 과정을 나무구조로 도표화하여 표현함으로서 진단과정을 쉽게 이해시키고 설명할 수 있고, 자료의 특성상 많은 설명변수를 포함하는 진단과정에서 유용한 설명변수를 자동으로 찾아주므로 수리적 지식이 부족한 중풍전문의교육에 유용하게 사용될 수 있고 특히 소수의 설명변수로 증형의 진단이 가능하므로 빠른 처치를 필요로 하는 응급환자의 진단에 유용하게 활용될 수 있다. 물론 해석의 용이함이 진단모형의 가장 중요한 특성은 아니다. 진단모형의 가장 중요한 특징은 정확한 진단을 하는데 있다. 분석자료에 대한 정준판별함수를 이용한 진단모형의 오진 율 36%와 본 연구의 CHAID 알고리즘을 이용한 의사결정나무에 의한 진단모형의 오진 율 19.2%을 비교하면 의사결정나무를 이용한 진단모형의 정확도가 다른 통계적 방법을 이용한 진단모형의 정확도에 비하여 우수하다고 할 수 있다. 특히 의사결정나무는 비모수적 방법이므로 자료에 대하여 선형성과 정규성을 가정하기 힘든 한의학분야의 자료분석에 유용하게 사용될 수 있으리라 생각된다.

참고문헌

[1] 강효신, 권영규, 박창국, 신양규, 김상철 (1996). 중풍임상자료에 대한 통계적 분석방법연구, 대한한의학회지, 제 17권 1호, 302-328.

[2] 김완희, 김광중 (1996). 장부학의 이론과 임상, 일중사, 서울.

[3] 박종운 (1993). 한의학 학위논문의 내용에 대한 조사 연구, 대한원전의사학회지, 제7권 167-197.

[4] 신양규, 강효신, 권영규, 박창국, 김상철 (1998). 전문가시스템을 이용한 한의진단의 객관화에

관한 연구, 연구과제최종보고서, 보건복지부.
[5] 최종후, 한상태, 강현철, 김은석 (1998). 데이터마이닝 의사결정나무분석, 자유아카데미, 서울.
[6] David Biggs, Barry de Ville and Ed Suen (1991). A method of choosing multiway partitions for classification and decision trees, *Journal of Applied Statistics*, Vol. 18, No 1, 49-62.
[7] Yong-Seok Choi, Jong-Geoun Kim and Jong-Hee Lee (2001), A Comparison of Capabilities of Data Mining Tools, *The Korean Communications in Statistics,* Vol. 8, No 2, 531-541.
[8] SPSS Inc. (1998). *AnswerTree 2.0 User's Guide*, SPSS Inc., Chicago.