데이터마이닝 기법을 이용한 사상체질 판별함수에 관한 연구

김규곤* · 김종원[1] · 이의주[2] · 김종열[3] · 최선미[3]

동의대학교 정보통계학과, 1 : 동의대학교 한의과대학 사상체질과, 2 : 경희대학교 한의과대학 사상체질과, 3 : 한국한의학연구원

Study on Classification Function into Sasang Constitution Using Data Mining Techniques

Kyu Kon Kim*, Jong Won Kim[1], Eui Ju Lee[2], Jong Yeol Kim[3], Sun-Mi Choi[3]

Department of Information Statistics, Dongeui University,
1 : Department of Sasang Constitutional Medicine, College of Oriental Medicine, Dongeui University,
2 : Department of Sasang Constitutional Medicine, College of Oriental Medicine, Kyung-Hee University,
3 : Korea Institute of Oriental Medicine

In this study, when we make a diagnosis of constitution using QSCC II(Questionnaire of Sasang Constitution Classification), data mining techniques are applied to seek the classification function for improving the accuracy. Data used in the analysis are the questionnaires of 1051 patients who had been treated in Dong Eui Oriental Medical Hospital and Kyung Hee Oriental Medical Hospital. The criteria for data cleansing are the response pattern in the opposite questionnaires and the positive proportion of specific questionnaires in each constitution. And the criteria for variable selection are the test of homogeneity in frequency analysis and the coefficients in the linear discriminant function. Discriminant analysis model and decision tree model are applied to seek the classification function into Sasang constitution. The accuracy in learning sample is similar in two models, the higher accuracy in test sample is obtained in discriminant analysis model.

Key words : Sasang constitution, classification function, data mining, frequency analysis, discriminant analysis model, decision tree model

서 론

사상체질의학은 東武 李濟馬가 완성한 1880년의 <格致藁>[1], 1894년의 <東醫壽世保元>[2]에서 정립한 한의학의 한 분야이다. 사상체질의학에서는 인간의 체질을 太陰人, 少陰人, 少陽人, 太陽人의 四象으로 定義하고 있으며[3], 각 체질적 특성에 따라 性質才幹, 容貌詞氣, 心性心慾, 生理病理 및 사회적 養生法 등에 있어 차이가 생긴다고 한다[4].

사상의학의 우수성은 많은 임상의사들로부터 환영을 받으면서도 체질진단의 난이성과 객관성에 많은 문제가 제기되고 있으며, 이러한 어려움을 극복하기 위한 방법의 일환으로 체질판별의 객관화를 위하여 많은 연구결과가 꾸준히 제시되고 있다[5]. 많은 연구자들은 사상체질을 분류하기 위하여 설문 결과를 기초자료로 사용하고 있는데, 사상체질 진단에 설문지를 이용한 방법을 처음으로 사용하기 시작한 것은 고병희 등[6,7]에 의해서이다. 그 이후 김선호 등[8]과 이정찬 등[9]은 체질별로 새로운 설문문항을 만들어 사상체질분류검사지 I 을 구성하였고, 다시 체질분석에 부적합하거나 수검자의 반응빈도가 낮은 문항을 삭제하는 등 설문지의 신뢰도를 높이고자 설문지의 보완작업을 진행하여 사상체질분류검사지 II를 개발하여 사용하고 있다[10,11]. 한편 이의주 등[12]은 사상변증내용 설문조사지(I)의 타당화 연구에서 고병희 등[6,7]의 방법과 달리 판별식을 이용하여 분석하는 것이 더욱 사상체질을 진단하는데 도움이 된다고 하였다.

본 연구에서는 사상체질분류검사 설문지(QSCC II : Questionnaire of Sasang Constitution Classification)를 이용한 체질진단에서 그 정확도를 향상시키기 위한 새로운 판별함수를 구하기 위하여 데이터마이닝 기법을 이용하고자 한다.

* 교신저자 : 김규곤, 부산시 진구 가야동 산24 동의대학교 정보통계학과
· E-mail : kkkim@deu.ac.kr, · Tel : 042-229-6805
· 접수 : 2004/07/13 · 수정 : 2004/08/14 · 채택 : 2004/09/15

대상 및 방법

1. 연구 대상

1) 설문지 구성

본 연구에서 사용하는 사상체질분류검사 설문지는 인적사항 11개 문항과 함께 사상체질을 판별하는데 사용되는 설문은 121개 문항으로서 2~4개 항목 중 1개 항목을 선택하는 설문이 15개, ○×로 체크하는 설문이 106개로 구성되어 있다. 전자의 경우 15개의 설문들은 특정 체질인 사람은 1개의 항목만을 선택할 수 있기 때문에 사실상 QSCC II는 157개의 ○× 설문으로 구성된 체질검사지라고 할 수 있다.

2) 데이터 수집

본 연구에서 사용하는 데이터는 경희대학교 강남경희한방병원과 동의대학교 부속한방병원에서 치료를 받은 환자들 중 각 병원의 사상체질전문의로부터 체질진단을 받고 최소한 4주 이상 사상체질 처방을 사용한 후 주 증상이 전반적으로 호전되어 체질이 확인된 환자 1051명을 대상으로 하고 있다.

3) 데이터 정제 및 전처리

(1) 데이터 정제(1차) : 입력오류 수정

데이터 파일에는 종종 일관성이 없고 불완전하며 오류가 있는 데이터가 존재할 수 있다. 따라서 데이터 입력 과정에서 발생한 오류를 수정하기 위하여 데이터 정제(data cleansing) 과정을 거침으로서 데이터의 질을 보장할 수 있다.

본 연구에서는 1번~15번 설문의 응답결과를 입력할 때 예를 들어 ①②③④ 중 1개의 문항을 선택해야 하는데도 불구하고 [0]이나 [5], [6] 등으로 입력되어 있는 경우는 모두 결측값(missing data)으로 처리한다. 16번~121번 설문의 응답결과를 입력할 때 체크(√)하면 [1], 아니면 [0]이어야 하는데도 불구하고 [.]이나 [2] 등으로 입력되어 있는 경우는 모두 0으로 수정한다.

(2) 데이터 탐색 및 변환

데이터마이닝 기법에서는 데이터의 탐색(exploration) 과정을 통하여 수많은 변수들의 관계를 살펴보고 정보화할 수 있는 기반을 잡는다. 이 과정을 통하여 수십 개 이상의 많은 변수들이 실제 중요한 정보를 주는 소수의 변수로 축소시킬 수 있다. 그리고 데이터의 변환(transformation) 과정을 통하여 생성 또는 수정된 변수를 차후 모형화(modeling)에서 아주 중요한 정보로 활용하기 위하여 준비한다[13]. 본 연구에서는 전체 데이터를 탐색한 결과 본래의 121개의 설문은 157개의 ○× 설문으로 파악되었고, 이들은 다시 4개의 체질 집단으로 축소될 수 있다. 따라서 데이터 변환 과정에서는 16번~121번 설문의 응답이 [0,1]이므로 1번~15번 설문의 응답도 [0,1]로 통일시킨다. 예를 들어 1번 설문의 응답에서 ①에 체크하면 q1_1=1, ②③에 체크하면 q1_1=0으로 하고, ②에 체크하면 q1_2=1, ①③에 체크하면 q1_2=0으로 하고, ③에 체크하면 q1_3=1, ①②에 체크하면 q1_3=0으로 변환시킨다. 이 경우 q1이 본래 결측값이었던 응답자에 대해서는 q1_1, q1_2, q1_3 모두 0으로 수정된다.

(3) 데이터 정제(2차) : 불량응답자 1차 제거

불성실한 응답자들은 설문지의 응답에 일관성이 없는 경우가 많기 때문에, 이러한 데이터를 제거시키는 데이터 정제 과정을 다시 한번 거침으로서 양질의 데이터를 확보할 수 있다. 여기서 선택된 응답자들은 모형 설정의 대상으로 활용하고, 제거된 응답자들은 설정된 모형의 타당성을 평가하기 위한 대상으로 활용한다.

예를 들어 설문 26번 "개인적인 일보다 사회적인 일에 열심이다"와 설문 27번 "사회적인 일보다 개인적인 일에 열심이다"는 서로 상반되는 설문이다. 이런 경우 26번에 체크했다면 27번에는 체크하지 않아야 하는데도 불구하고 만약 두 설문에 모두 체크한 응답자는 불량응답자로 간주하여 사상체질판별함수를 구할 때까지의 분석에서는 제외시킨다. 사상체질분류검사 설문지에서 서로 상반되는 설문으로 판단되는 것은 q25↔q30, q26↔q27, q28↔q29, q61↔q101 등으로서 그 내용은 다음과 같다.

Table 1 reverse questionnaires

25. 사람을 사귈 때 이것저것 따지지 않고 쉽게 잘 사귄다.	↔	30. 사람을 사귈 때 이것저것 따져서 쉽게 사귀지 못하는 편이다
26. 개인적인 일보다 사회적인 일에 열심이다.	↔	27. 사회적인 일보다 개인적인 일에 열심이다.
28. 내면적인 것보다 외면적인 것을 더 중요시한다.	↔	29. 외면적인 것보다 내면적인 것을 더 중요시한다.
61. 남성적인 면이 많고 여성적인 면이 적다.	↔	101. 여성적인 면이 많고 남성적인 면이 적다.

이상의 데이터 정제 및 전처리 과정을 거치면서 불량응답자를 1차 제거시킨 후에 얻은 표본은 1051명 중 422명이었다.

2. 학습표본

이제 422명의 표본을 대상으로 먼저 도수분석을 하여 변수를 선택하고, 다음으로 선택된 변수에 대하여 결측값(missing data)의 비율이 크면 또다시 불량응답자로 간주하여 제거시킨 후 양질의 학습표본(learning sample)을 선택한다[14-16].

1) 도수분석을 이용한 변수선택

(1) 변수선택의 기준

도수분석 결과를 이용하여 변수선택을 하기 위한 기준은 다음과 같다. 첫째, 각 체질별 설문의 체크여부 [0,1]간의 비율차이 검정을 위한 일원도수분석을 이용한다. 본 연구의 대상인 데이터는 각 체질별 응답자 수가 다르기 때문에 특정한 체질의 설문에 그 체질인 사람이 체크한 비율이 50% 이상이면 그 체질에 관련된 설문으로 볼 수 있다. 유의한 경우는 <부록 표 1>의 각 셀에서 [*]표로 표시하였다. 둘째, 사상체질간의 비율차이검정을 위한 이원도수분석을 이용한다. 설문의 체크여부 [0,1]과 사상체질과의 교차표에서 특정한 체질의 설문에 그 체질인 사람이 체크한 비율이 50% 이상이고, 다른 체질인 사람이 체크한 비율이 50% 미만이면 그 체질에 관련된 문항으로 볼 수 있다. 유의한 경우는 <부록 표 1>의 마지막 열에서 p-값으로 표시하였다.

이와 같은 기준은 체질 구분이 참(true)이라는 가정 하에서 옳은 것이다. 본 연구에서의 체질은 사상체질전문의가 확진한 결과이므로 위와 같은 기준의 설정은 가능하다고 할 수 있다.

(2) 변수제거의 예

예 1) 해당 체질 체크 50% 미만인 경우 : 설문 q1_3은 소음인

이 체크하는 문항이지만, 소음인임에도 불구하고 체크한 사람의 비율이 40.6%로서 50%를 넘지 못하므로 분석에서 제외시킨다. 또한 체질간 비율차이검정에서 유의한 결과를 얻었다고 할지라도 소음인을 찾아낼 변별력이 없는 문항이다.

예 2) 해당 체질 체크 50% 이상이지만 다른 체질의 체크 비율보다 적은 경우 : 설문 q8_1은 소음인이 체크하는 문항이지만, 소음인 중 체크한 사람의 비율이 53.3%로서 50%를 넘고 있지만 체크하지 않은 사람과의 비율차이검정에서 유의한 결과를 얻지 못하였으며, 또한 태양인 중 체크한 사람의 비율이 60.0%로서 소음인 보다 오히려 더 많은 사람이 체크하고 있으므로 소음인과 태양인을 판별할 수 없는 문항이므로 분석에서 제외시킨다.

예 3) 해당 체질 체크 50% 이상이고 유의한 차이가 있으나 다른 체질에서도 50% 이상이고 유의한 차이가 있는 경우 : 설문 q11_2는 소음인이 체크하는 문항이지만, 소음인 중 체크한 사람의 비율이 59.4%로서 50% 이상이고 체크하지 않은 사람과의 비율차이검정에서 유의한 결과를 얻었지만, 동시에 소양인 중 체크한 사람의 비율도 50% 이상이고 체크하지 않은 사람과의 비율차이검정에서 유의한 결과를 얻었기 때문에 소음인과 소양인을 판별할 수 없는 문항이므로 분석에서 제외시킨다.

이상의 결과를 정리한 Table 2를 보면 선택된 설문은 태양체질 4개, 소양체질 5개, 태음체질 6개, 소음체질 14개 등 총 29개 변수가 선택되었으며, 이 변수들은 서로 다른 체질에 겹치지 않는다.

Table 2. Result of variable selection and removal

	태양	소양	태음	소음
선택된 설문 (해당 체질 체크 50% 이상 이면서 유의한 차이가 있고 다른 체질에서는 50% 미만인 경우)			q2_1 q8_2	q5_2 q12_4 q30 q77
선택된 설문 (해당 체질 체크 50% 이상이면서 유의한 차이가 있지만 다른 체질에서는 50% 이상이면서 유의한 차이가 없는 경우)	q40 (예외) q97	q1_2 q2_2 q69		q23 q56 q59 q101 q111
선택된 설문 (해당 체질 체크 50% 이상이면서 비율이 가장 높지만 유의한 차이가 없고 다른 체질에서는 50% 미만인 경우)	q15_1	q5_3 q57	q5_1 q6_1 q80	q6_2 q9_2 q10_2 q51 q112
선택된 설문 (해당 체질 체크 50% 이상이면서 비율이 가장 높지만 유의한 차이가 없고 다른 체질에서도 50% 이상이면서 유의한 차이가 없는 경우)	q44		q114 (예외)	
제외된 설문 (해당 체질 체크 50% 미만인 경우)	q1_1 q2_3 q3_1 q4_1 q6_4 q7_3 q9_4 q10_3 q12_1 q13_1 q14_1 q26 q28 q61 q62 q74 q78 q83 q90 q92 q102 q108 q113	q3_1 q4_2 q6_3 q7_4 q8_2 q9_3 q10_3 q12_2 q13_2 q14_2 q15_2 q20 q26 q28 q31 q36 q49 q54 q55 q58 q63 q75 q79 q84 q93 q98 q103 q106 q109 q119	q1_1 q3_2 q4_3 q7_2 q9_1 q10_1 q11_1 q12_3 q13_3 q14_3 q15_3 q37 q42 q46 q64 q76 q85 q88 q91 q99 q104 q110 q115 q117 q121	q1_3 q2_3 q3_2 q3_3 q4_4 q7_1 q13_4 q15_4 q43 q47 q65 q72 q81 q86 q89 q95 q96 q105 q107 q116 q120
제외된 설문 (해당 체질 체크 50% 이상이지만 다른 체질의 체크 비율보다 적은 경우)	q17 q18 q25 q34 q35 q48 q68	q19 q45	q21 q22 q53 q94 q118	q8_1 q38
제외된 설문 (해당 체질 체크 50% 이상이고 유의한 차이가 있으나 다른 체질에서도 50% 이상이고 유의한 차이가 있는 경우)		q41	q27 q29 q50 q70 q71	q11_2 q14_4 q24 q27 q29 q33 q52 q100
합계	34	38	41	45

2) 학습표본의 선택

앞에서 선택된 29개의 변수에 대하여 422명의 설문결과가 각 체질별로 결측값(missing data)의 비율이 크면 또다시 불량응답자로 간주하여 2차 제거시킨 후 학습표본을 선택한다.

(1) 불량응답자의 기준

첫째, 어떤 특정한 체질에 속하는 사람은 그 체질 설문의 평균값이 0.5보다 작으면 불량응답자이다. 예를 들어, 태양체질인 사람은 태양체질 4개 설문 중 2개 설문에 체크했다면 평균값이 0.5이고 체크한 비율은 50%이다. 따라서 평균값이 0.5보다 작으면 불량응답자이다. 둘째, 어떤 특정한 체질에 속하는 사람은 그 체질 설문의 평균값이 다른 체질 설문의 평균값보다 작으면 불량응답자이다. 예를 들어, 태양체질인 사람은 태양체질 설문에 50% 이상 체크하였지만 다른 체질 설문에는 그 이상 체크했다면 불량응답자이다. 위와 같은 불량응답자를 제거시킨 후 얻은 학습표본(learning sample)은 205명이고, 1차 제거된 629명을 검정표본(test sample)으로 하고, 2차 제거된 217명은 완전히 제거시키며, 데이터 구조는 Table 3과 같다.

Table 3. Constitution of Data <Frequency/%>

	Taeyang	Soyang	Taeum	Soum	Total
Total	34 3.24	254 24.17	389 37.01	374 35.59	1051 100.0
1st removal (test sample)	14 2.23	163 25.91	243 38.63	209 33.23	629 59.85
2nd removal (complete removal)	10 4.61	41 18.89	88 40.55	78 35.94	217 20.65
learning sample	10 4.88	50 24.39	58 28.29	87 42.44	205 19.51

결과 및 고찰

1. 사상체질판별함수

데이터마이닝에는 특정 문제에 적용하는 기법이 정해져 있는 것은 아니다. 또한 특정 기법이 적용된다고 해서 모든 문제가 해결되는 것도 아니다. 얻고자 하는 결과나 데이터의 상태 등에 따라 적용할 수 있는 기법들은 다를 수가 있다. 그러므로 기법들에 대해 어느 정도의 이해가 수반되면 문제를 해결하는데 좀 더 최적의 접근으로, 보다 효과적이고 적극적인 데이터마이닝을 수행할 수 있을 것이다. 데이터마이닝의 기법에는 일반적으로 통계학에서의 다변량 분류 기법들을 포함하여 연관성(associations), 군집분석(clustering), 의사결정나무(decision trees), 신경망모형(neural networks)과 같은 기법들이 있다[18].

본 논문에서는 29개의 선택된 변수에 대하여 데이터마이닝 기법 중 판별분석모형, 의사결정나무모형을 적용하여 사상체질판별함수를 구한다[13,15]. 학습표본으로 사용할 데이터는 205명이고, 검정표본으로 사용할 데이터는 1차 제거된 불량응답자 629명이다.

1) 판별분석모형

판별분석(discriminant analysis)은 다변량자료분석의 한 분야로서 집단간의 차이를 식별하는데 사용되는 여러 개의 서로 상관되어 있는 판별변수(discriminant variable)와 사전에 정의된 하나의 집단변수(group variable)를 가지고 있는 다변량자료를 그 대상으로, 집단간의 분리 정도에 관한 해석과 각 개체를 특정 집단에 분류하는데 필요한 적정분류기준의 설정 및 판별변수에 관한 구조분석, 그리고 이에 따른 분류방법과 관련된 통계적 기법을 총괄적으로 포함하고 있다[19].

(1) 판별함수의 추정

학습표본 205명을 사상체질의 4집단으로 분류하기 위하여 학습표본에 판별분석모형을 적합한 결과로 얻은 판별함수는 Table 4와 같다.

앞에서 구한 판별함수를 205명의 학습표본에 적용한 결과는 Table 5, Table 6과 같으며, 정분류율(accuracy)이 90.71%, 오분류율(error rate)이 9.29%로 나타났다. 그리고 정분류율을 체질별로 살펴보면 태음인 98.28%, 소양인 98.0%, 소음인 96.55%, 태양인 70.0%의 순으로 높게 나타났으며, 태음인, 소양인, 소음인의 세 가지 체질에서 90% 이상의 정분류율을 보였다.

(2) 검정표본에 적용한 결과

적합된 판별분석모형인 사상체질판별함수를 사용하여 629명의 검정표본을 체질 분류한 결과는 Table 7, Table 8과 같으며, 정분류율이 55.80%, 오분류율이 44.20%로 나타났다. 그리고 정분류율을 체질별로 살펴보면 소양인 62.58%, 소음인 55.02%, 태음인 54.32%, 태양인 14.29%의 순으로 나타났다.

Table 4. Linear Discriminant Function for sasang of Sasang constitution

Variable	Taeyang	Soyang	Taeum	Soum
Constant	-43.29545	-43.24482	-52.23113	-41.75998
q40	1.44023	0.36968	-0.77456	0.92941
q97	3.06339	1.02056	0.56519	-0.07669
q15_1	9.13294	3.66976	3.81095	3.05561
q44	1.60849	-0.79411	-0.24965	-1.13290
q1_2	4.50716	6.97749	3.18611	2.68541
q2_2	1.63215	2.72309	2.23128	2.18045
q69	0.56018	2.59295	-0.81068	0.84823
q5_3	59.39728	66.65804	56.50937	56.72126
q57	3.69921	5.65431	5.72414	1.77795
q2_1	8.63789	5.79988	24.30723	7.70649
q8_2	3.88126	3.97549	4.80238	2.54496
q5_1	57.22485	57.35666	61.55609	51.12490
q6_1	1.02868	0.86880	5.76755	0.17089
q80	-0.17745	-1.34416	1.36060	-0.06884
q114	2.72860	3.36946	4.55589	2.11627
q5_2	55.67364	58.38012	53.28038	56.22136
q12_4	1.97536	1.75750	2.59812	3.75595
q30	2.47335	0.72899	1.26116	4.06618
q77	1.58708	0.45791	2.60942	1.04486
q23	-2.60527	-0.34768	-1.27650	1.70942
q56	0.79939	2.18714	0.76735	2.52360
q59	0.08731	1.62386	0.15056	0.96303
q101	1.66320	-1.91804	0.61532	1.82790
q111	0.60402	-0.36055	0.22944	0.37986
q6_2	5.30539	3.38282	3.61851	5.99510
q9_2	0.15335	0.97132	1.89513	2.43079
q10_2	1.44791	1.15385	-0.47050	2.84415
q51	2.19598	0.41067	3.13915	2.83063
q112	0.45196	0.38359	1.12260	1.78307

Table 5. Number of Observations and Percent Classified into Sasang (Frequency/%)

		Actual category				
		Taeyang	Soyang	Taeum	Soum	Total
Predicted category	Taeyang	7 70.00	1 2.00	1 1.72	2 2.30	11 5.37
	Soyang	1 10.00	49 98.00	0 0.00	0 0.00	50 24.39
	Taeum	2 20.00	0 0.00	57 98.28	1 1.15	60 29.27
	Soum	0 0.00	0 0.00	0 0.00	84 96.55	84 40.98
	Total	10 100.0	50 100.0	58 100.0	87 100.0	205 100.0

Table 6. Error Count Estimates for Sasang (Frequency/%)

	Taeyang	Soyang	Taeum	Soum	Total
Error rate	3 30.00	1 2.00	1 1.72	3 3.45	8 9.29
Accuracy	7 70.00	49 98.00	57 98.28	84 96.55	197 90.71

Table 7. Misclassification table in discriminant analysis model (Frequency/%)

		Actual category				
		Taeyang	Soyang	Taeum	Soum	Total
Predicted category	Taeyang	2 14.29	15 9.20	22 9.05	16 7.66	55 8.74
	Soyang	9 64.29	102 62.58	66 27.16	59 28.23	236 37.52
	Taeum	0 0.00	15 9.20	132 54.32	19 9.09	166 26.39
	Soum	3 21.43	31 19.02	23 9.47	115 55.02	172 27.34
	Total	14 100.0	163 100.0	243 100.0	209 100.0	629 100.0

Table 8. Sasang constitutional Risk Estimate of Test sample (Frequency/%)

	Taeyang	Soyang	Taeum	Soum	Total
Error rate	12 85.71	61 37.42	111 45.68	94 44.98	278 44.20
Accuracy	2 14.29	102 62.58	132 54.32	115 55.02	351 55.80

2) 의사결정나무모형(decision tree model)

의사결정나무(decision tree)는 의사결정규칙을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류하거나 예측하는 분석방법이다. 분석과정이 나무구조에 의해서 표현되기 때문에, 신경망, 판별분석, 회귀분석 등과 같은 방법들에 비해, 연구자가 분석과정을 쉽게 이해하고 설명할 수 있다는 장점을 가지고 있다. 의사결정나무는 분류 또는 예측을 목적으로 하는 어떤 경우에도 사용될 수 있으나, 분석의 정확도보다는 분석과정의 설명이 필요한 경우에 더 유용하게 사용된다(Berry와 Linoff, 1997). 의사결정나무모형의 알고리즘은 CHAID(chi-squared automatic interaction detection) 알고리즘과 CART(classification and regression trees) 알고리즘 등이 있으나 본 연구에서는 CART 알고리즘을 다룬다. 여기서 CART 알고리즘은 이산형 목표변수인 경우에 적용하는 불순도(impurity)를 측정하는 지니 지수(Gini index) 또는 연속형 목표변수인 경우에 적용하는 분산의 감소량을 이용하여 이지분리(binary split)를 수행하는 알고리즘이다(Breiman 등, 1984).

(1) 의사결정나무의 추정

학습표본 205명을 사상체질의 4집단으로 분류하기 위하여 학습표본에 의사결정나무모형을 적합한 결과는 Fig. 1과 같다. Fig. 1에서 원(○)은 뿌리 노드(root node)와 중간 노드(internal node)를 나타내고, 사각형(□)은 최종 노드(terminal node)를 나타낸다. 분리기준(split criterion)은 부모 노드(parent node)에서 5명, 정지규칙(stopping rule)은 자식 노드(child node)에서는 1명을 지정하였다. 그리고 분류오류(classification error)를 크게 할 위험(risk)이 높거나 부적절한 추론규칙(induction rule)을 가지고 있는 가지(branch)는 가지치기(pruning)하였다.

Fig. 1에서 뿌리 노드 [n=205, Q2_1(태음)]은 205명에 대하여 태음체질 설문인 Q2_1에 체크하였다면(Yes) 오른쪽 가지로 분리(split)되고, 체크하지 않았다면(No) 왼쪽 가지로 분리된다. 그리고 최종 노드를 집계하면 Table 9를 얻을 수 있다.

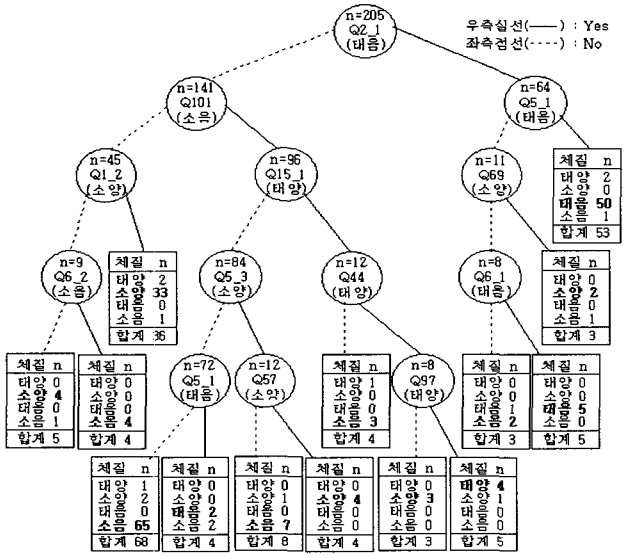

Fig. 1 Decision tree

CART 알고리즘을 적용한 의사결정나무를 학습표본에 적용한 결과는 Table 9, Table 10과 같으며, 정분류율이 91.71%, 오분류율이 8.29%로 나타났다. 그리고 정분류율을 체질별로 살펴보면 태음인 98.28%, 소음인 93.10%, 소양인 92.0%, 태양인 40.0%의 순으로 높게 나타났으며, 태음인, 소음인, 소양인의 세 가지 체질에서 90% 이상의 정분류율을 보였다.

Table 9. Number of Observations and Percent Classified into Sasang (Frequency/%)

		Actual category				
		Taeyang	Soyang	Taeum	Soum	Total
Predicted category	Taeyang	4 40.00	1 2.00	0 0.00	0 0.00	5 2.44
	Soyang	2 20.00	46 92.00	0 0.00	3 3.45	51 24.88
	Taeum	2 20.00	0 0.00	57 98.28	3 3.45	62 30.24
	Soum	2 20.00	3 6.00	1 1.72	81 93.10	87 42.44
	Total	10 100.0	50 100.0	58 100.0	87 100.0	205 100.0

Table 10. Error Count Estimates for Sasang (Frequency/%)

	Taeyang	Soyang	Taeum	Soum	Total
Error rate	6 60.00	4 8.00	1 1.72	6 6.90	17 8.29
Accuracy	4 40.00	46 92.00	57 98.28	81 93.10	188 91.71

(2) 검정표본에 적용한 결과

CART 알고리즘을 적용한 적합된 의사결정나무모형을 사용하여 629명의 검정표본을 체질 분류한 결과는 <Table 11>, <Table 12>와 같으며, 정분류율이 52.62%, 오분류율이 47.38%로 나타났다. 그리고 정분류율을 체질별로 살펴보면 소양인 69.94%, 태음인 53.09%, 소음인 42.11%, 태양인 0.0%의 순으로 나타났다.

Table 10. Misclassification table in decision tree model (Frequency/%)

		Actual category				
		Taeyang	Soyang	Taeum	Soum	Total
Predicted category	Taeyang	0 0.00	3 1.84	0 0.00	0 0.00	3 0.48
	Soyang	12 85.71	114 69.94	80 32.92	103 49.28	309 49.13
	Taeum	0 0.00	13 7.98	129 53.09	18 8.61	160 25.44
	Soum	2 14.29	33 20.25	34 13.99	88 42.11	157 24.96
	Total	14 100.0	163 100.0	243 100.0	209 100.0	629 100.0

Table 12. Sasang constitutional Risk Estimate of Test sample (Frequency/%)

	Taeyang	Soyang	Taeum	Soum	Total
Error rate	14 100.00	49 30.06	114 46.91	121 57.89	298 47.38
Accuracy	0 0.00	114 69.94	129 53.09	88 42.11	331 52.62

결 론

본 논문에서는 사상체질분류검사 설문지(QSCC II : Questionnaire of Sasang Constitution Classification)를 이용한 체질진단에서 그 정확도를 향상시키기 위한 새로운 판별함수를 구하기 위하여 데이터마이닝 기법을 이용하였다. 본 연구에서 사용하는 데이터는 경희대학교 강남경희한방병원과 동의대학교 부속한방병원에서 치료를 받은 환자들 중 각 병원의 사상체질전문의로부터 체질진단을 받고 최소한 4주 이상 사상체질 처방을 사용한 후 주 증상이 전반적으로 호전되어 체질이 확인된 환자 1051명을 대상으로 하고 있다. 데이터 정제 과정에서 양질의 데이터를 확보하기 위한 기준은 상반되는 설문의 응답 패턴과 체질별 설문의 응답 비율을 이용하였으며, 변수선택의 기준은 도수분석의 비율차이검정과 선형판별함수의 계수를 이용하였다. 사상체질판별함수를 구하기 위한 데이터마이닝 기법은 판별분석모형과 의사결정나무모형을 적용하였다. 판별함수를 학습표본에 적용한 결과 정분류율은 판별분석모형에서 90.71%, 의사결정나무모형에서 91.71%로서 비슷한 결과를 얻었다.

체질별로는 판별분석모형에서 태음인 98.28%, 소양인

98.0%, 소음인 96.55%, 태양인 70.0%의 순으로 높게 나타났으며, 태음인, 소양인, 소음인의 세 가지 체질에서 90% 이상의 정분류율을 보였다. 의사결정나무모형에서 태음인 98.28%, 소음인 93.10%, 소양인 92.0%, 태양인 40.0%의 순으로 높게 나타났으며, 태음인, 소음인, 소양인의 세 가지 체질에서 90% 이상의 정분류율을 보였다. 판별함수를 검정표본에 적용한 결과 정분류율은 판별분석모형에서 55.80%, 의사결정나무모형에서 52.62%로서 판별분석모형이 우수하였다. 체질별로는 판별분석모형에서 소양인 62.58%, 소음인 55.02%, 태음인 54.32%, 태양인 14.29%의 순으로 높게 나타났으며, 의사결정나무모형에서 소양인 69.94%, 태음인 53.09%, 소음인 42.11%, 태양인 0.0%의 순으로 높게 나타났다.

본 연구의 결과 학습표본의 정분류율은 만족할 수 있지만 검정표본의 정분류율은 매우 낮아 만족스럽지 못하였다. 그 이유는 2차 데이터 정제 과정에서 제거된 불성실한 응답자들을 검정표본으로 사용했기 때문이다.

따라서 향후 연구과제로서는 응답자들이 이해하기 쉽고 혼동을 피할 수 있는 설문을 개발함으로서 신뢰도를 향상시키는 것이 무엇보다 중요한 일이라고 할 수 있다.

참고문헌

1. 李濟馬. 格致藁. 청계출판, 서울: 10. 2000.
2. 李濟馬. 東醫壽世保元. 행림출판. 서울: 137-142. 1986.
3. 이태호. 실제적 동의사상진료의 비결. 행림서원. 서울: 31-47. 1961.
4. 송일병. 알기쉬운 사상의학, 사상사, 서울: 50-89. 1993.
5. 김영우. 사상체질진단을 위한 사상체질분류검사지Ⅱ(QSCCⅡ)의 연구. 동의대학교 대학원 한의학과 박사학위 논문. 2004
6. 고병희, 송일병. 사상체질변증에 관한 소고. 대한한의학회지. 6(1): 40-47, 1985.
7. 고병희, 송일병. 사상체질변증방법론 연구(제2보). 대한한의학회지. 8(1): 146-160, 1987.
8. 김선호, 고병희, 송일병 . 사상체질분류검사(QSCC)의 타당화 연구. 사상의학회지. 5(1): 67-85. 1993.
9. 이정찬, 고병희, 송일병 . 사상체질분류검사의 준거타당화 연구. 사상의학회지. 5(1): 87-104. 1993.
10. 김선호, 고병희, 송일병 . 사상체질분류검사지(QSCCⅡ)의 표준화 연구. 사상의학회지. 8(1): 187-246. 1996.
11. 이정찬, 고병희, 송일병 . 사상체질분류검사지(QSCCⅡ)의 타당화 연구. 사상의학회지. 8(1): 247-294. 1996.
12. 이의주, 고병희, 송일병 . 사상변증내용 설문조사지(Ⅰ)의 타당화 연구. 사상의학회지. 7(2): 89-100. 1995.
13. 김규곤. 데이터마이닝에서의 분류방법에 관한 연구. Journal of the Korean Data Analysis Society. 5(1): 101-112, 2003.
14. 김규곤. 이산 다변량 분석을 이용한 한방 진단 프로그램 개발 연구. Journal of The Korean Data Analysis Society. 1(1): 15-27, 1999.
15. 김규곤. 한방 통계분석방법에 관한 사례연구. Journal of the Korean Data Analysis Society. 5(4): 907-917. 2003.
16. 전란희, 이인선, 김규곤, 강창완. 한방 부인과 자료에서의 수량화분석. Journal of The Korean Data Analysis Society. 1(1): 53-63. 1999.
17. 김규곤, 강창완. 한의학에서의 변증점수개발에 대한 가중주성분분석의 응용. 응용통계연구. 12(1): 17-28, 1999.
18. BerryMJA, Linoff G. Data Mining Techniques. John Wiley & Sons, Inc. New York. 1997.
19. 김기영, 전명식 . SAS 판별 및 분류분석. 자유아카데미, 서울, 1997.
20. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees, Belmont, Wordsworth. 1984.

<부록 표 1> 도수분석 결과

설문번호	체질(본래것)	1태양	2소양	3태음	4소음	비율차이
q1__1	1-태양 3-태음	20.0	14.3	43.8	6.7	0.0001
q1__2 *	2-소양	65.0	63.7**	43.8	52.7	0.0165
q1__3	4-소음	15.0	22.0	11.6	40.6	0.0001
q2__1 *	3-태음	10.0	8.8	58.9*	6.1	0.0001
q2__2 *	2-소양	65.0	70.3***	39.0	51.5	0.0001
q2__3	1-태양 4-소음	25.0	20.9	2.1	42.4	0.0001
q3__1	1-태양 2-소양	15.0	36.3	34.3	16.4	0.0003
q3__2	3-태음 4-소음	40.0	22.0	24.0	33.9	0.0691
q3__3	4-소음???	45.0	41.8	41.8	49.7	0.4811
q4__1	1-태양	15.0	15.4	13.0	14.6	0.9601
q4__2	2-소양	20.0	30.8	26.0	18.8	0.1539
q4__3	3-태음	5.0	13.2	15.1	10.3	0.4407
q4__4	4-소음	60.0	31.9	32.2	44.2	0.0157
q5__1 *	3-태음	15.0	9.9	56.2	8.5	0.0001
q5__2 *	4-소음	40.0	37.4	16.4	69.7***	0.0001
q5__3 *	2-소양	45.0	50.6	25.3	20.0	0.0001
q6__1 *	3-태음	10.0	11.0	50.7	9.1	0.0001
q6__2 *	4-소음	50.0	34.1	13.7	52.7	0.0001
q6__3	2-소양	20.0	31.9	18.5	11.5	0.0011
q6__4	1-태양	20.0	20.9	16.4	21.8	0.6748
q7__1	4-소음	25.0	16.5	15.1	22.4	0.3099
q7__2	3-태음	55.0	50.6	45.2	49.1	0.7663
q7__3	1-태양	10.0	24.2	26.7	24.9	0.4470
q7__4	2-소양	10.0	8.8	12.3	3.6	0.0436
q8__1	4-소음	60.0	52.8	39.7	53.3	0.0517
q8__2 *	2-소양 3-태음	40.0	47.3	58.9*	46.1	0.0824
q9__1	3-태음	5.0	18.7	36.3	13.9	0.0001
q9__2 *	4-소음	15.0	26.4	26.7	52.7	0.0001
q9__3	2-소양	50.0	40.7	27.4	24.2	0.0082
q9__4	1-태양	30.0	14.3	9.6	9.1	0.0298
q10__1	3-태음	15.0	20.9	34.3	12.1	0.0001
q10__2 *	4-소음	45.0	30.8	28.8	54.6	0.0001
q10__3	1-태양 2-소양???	40.0	47.3	37.0	30.3	0.0611
q11__1	3-태음	45.0	38.5	43.2	40.0	0.8680
q11__2	4-소음	55.0	60.4*	56.9	59.4*	0.9276
q12__1	1-태양	15.0	16.5	11.0	5.5	0.0354
q12__2	2-소양	35.0	26.4	17.1	4.9	0.0001
q12__3	3-태음	25.0	22.0	32.2	30.9	0.3428
q12__4 *	4-소음	20.0	34.1	39.7	58.8*	0.0001
q13__1	1-태양	25.0	6.6	11.6	10.9	0.1158
q13__2	2-소양	20.0	19.8	16.4	28.5	0.0727
q13__3	3-태음	20.0	40.7	36.3	32.1	0.2644
q13__4	4-소음	35.0	23.1	24.7	17.6	0.2116
q13__4	4-소음	35.0	23.1	24.7	17.6	0.2116
q14__1	1-태양	0.0	5.5	11.0	6.1	0.1547

설문번호	체질(본래것)	1태양	2소양	3태음	4소음	비율차이
q14__2	2-소양	20.0	14.3	8.2	10.9	0.2881
q14__3	3-태음	5.0	2.2	3.4	4.2	0.8386
q14__4	4-소음	75.0*	75.8***	76.0***	77.6***	0.9819
q15__1 *	1-태양	55.0	36.3	30.8	13.9	0.0001
q15__2	2-소양	20.0	33.0	21.2	12.7	0.0019
q15__3	3-태음	25.0	17.6	34.3	50.3	0.0001
q15__4	4-소음	0.0	12.1	13.7	22.4	0.0153
q16	무작위문항	30.0	51.7	43.2	27.9	0.0009
q17	1-태양	55.0	54.9	46.6	33.9	0.0056
q18	1-태양	55.0	64.8**	48.6	28.5	0.0001
q19	2-소양	55.0	53.9	40.4	16.4	0.0001
q20	2-소양	20.0	36.3	25.3	15.2	0.0018
q21	3-태음	75.0*	52.8	57.5	52.7	0.2493
q22	3-태음	55.0	42.9	50.7	63.6***	0.0099
q23 *	4-소음	45.0	50.6	56.9	78.2***	0.0001
q24	4-소음	60.0	55.0	76.7***	83.6***	0.0001
q25	1-태양	70.0	78.0***	62.3**	40.0	0.0001
q26	1-태양 2-소양???	30.0	37.4	37.0	24.2	0.0578
q27	3-태음??? 4-소음	70.0	62.6*	63.0**	75.8***	0.0578
q28	1-태양 2-소양???	40.0	27.5	31.5	20.6	0.0830
q29	3-태음??? 4-소음	60.0	72.5***	68.5***	79.4**	0.0830
q30 *	4-소음	30.0	22.0	37.7	60.0*	0.0001
q31	2-소양	35.0	45.1	26.7	10.9	0.0001
q32	무작위	70.0	48.4	55.5	64.2***	0.0531
q33	4-소음	60.0	52.8	67.1***	79.4***	0.0001
q34	1-태양	60.0	67.0**	41.8	17.6	0.0001
q35	1-태양	50.0	74.7***	78.8***	52.7	0.0001
q36	2-소양	45.0	37.4	35.6	39.4	0.8208
q37	3-태음	70.0	59.3	49.3	47.3	0.0922
q38	4-소음	45.0	62.6*	65.1***	54.6	0.1285
q39	무작위	40.0	42.9	47.3	20.6	0.0001
q40 *	1-태양	75.0*	56.0	66.4***	67.3***	0.2044
q41	2-소양	75.0*	60.4*	63.0**	64.2***	0.6678
q42	3-태음	5.0	9.9	13.0	14.6	0.5219
q43	4-소음	50.0	46.2	41.1	44.9	0.7988
q44 *	1-태양	55.0	42.9	51.4	47.3	0.5613
q45	2-소양	60.0	52.8	60.3*	61.2**	0.5911
q46	3-태음	35.0	25.3	26.0	22.4	0.6284
q47	4-소음	35.0	36.3	34.9	34.6	0.9943
q48	1-태양	55.0	76.9***	55.5	42.4	0.0001
q49	2-소양	30.0	44.0	35.6	24.9	0.0152
q50	3-태음	45.0	48.4	60.3*	74.6***	0.0001
q51 *	4-소음	35.0	30.8	38.4	57.6	0.0001
q52	4-소음	85.0**	58.2	58.9*	80.0***	0.0001
q53	3-태음	75.0*	45.1	54.8	58.8*	0.0497
q54	2-소양	25.0	22.0	16.4	17.6	0.6156
q55	2-소양	30.0	39.6	24.7	20.6	0.0103
q56 *	4-소음	65.0	57.1	53.4	68.5***	0.0447
q57 *	2-소양	50.0	58.2	54.1	44.2	0.1388
q58	2-소양	50.0	42.9	34.3	37.6	0.3900
q59 *	4-소음	65.0	57.1	53.4	61.8**	0.4450
q60	무작위	50.0	47.3	62.3**	72.1***	0.0008
q61	1-태양	45.0	64.8**	52.7	23.0	0.0001
q62	1-태양	50.0	49.5	45.9	35.8	0.1128
q63	2-소양	5.0	18.7	14.4	9.1	0.0990
q64	3-태음	25.0	20.9	29.5	41.2	0.0059
q65	4-소음	15.0	9.9	13.7	24.2	0.0144
q66	무작위	50.0	52.8	50.0	63.0***	0.1068

설문번호	체질(본래것)	1태양	2소양	3태음	4소음	비율차이
q67	무작위	25.0	37.4	50.0	58.8*	0.0012
q68	1-태양	60.0	57.1	63.7***	61.2**	0.7951
q69 *	2-소양	60.0	63.7**	48.6	52.1	0.1282
q70	3-태음	55.0	58.2	68.5***	71.5***	0.1066
q71	3-태음	55.0	55.0	59.6*	61.8**	0.7278
q72	4-소음	25.0	24.2	26.0	27.9	0.9304
q73	무작위	45.0	50.1	65.1***	67.9***	0.0146
q74	1-태양	0.0	8.8	8.2	5.5	0.4023
q75	2-소양	20.0	27.5	14.4	12.1	0.0126
q76	3-태음	25.0	16.5	40.4	38.2	0.0006
q77 *	4-소음	35.0	41.8	45.9	59.4*	0.0113
q78	1-태양	15.0	19.8	12.3	10.3	0.1888
q79	2-소양	50.0	31.9	26.7	14.6	0.0003
q80 *	3-태음	40.0	37.4	50.7	46.7	0.2275
q81	4-소음	35.0	29.7	30.1	35.8	0.6686
q82	무작위	40.0	47.3	40.4	29.7	0.0361
q83	1-태양	30.0	38.5	34.3	20.0	0.0065
q84	2-소양	25.0	30.8	22.6	32.1	0.2714
q85	3-태음	20.0	40.7	44.5	37.0	0.1570
q86	4-소음	0.0	6.6	13.0	15.8	0.0564
q87	무작위	20.0	28.6	32.2	24.2	0.3794
q88	3-태음	10.0	24.2	41.8	55.2	0.0001
q89	4-소음	15.0	19.8	31.5	41.8	0.0011
q90	1-태양	15.0	8.8	8.9	12.7	0.5846
q91	3-태음	25.0	22.0	15.8	18.8	0.5728
q92	1-태양	10.0	17.6	14.4	13.3	0.7508
q93	2-소양	40.0	40.7	34.9	44.2	0.4210
q94	3-태음	40.0	50.6	50.7	52.1	0.7891
q95	4-소음	20.0	17.6	26.0	40.6	0.0005
q96	4-소음	45.0	29.7	29.5	41.2	0.0786
q97 *	1-태양	80.0**	46.2	51.4	49.7	0.0525
q98	2-소양	15.0	28.6	21.2	15.2	0.0715
q99	3-태음	45.0	34.1	28.1	33.9	0.3943
q100	4-소음	65.0	60.4*	63.0**	58.2*	0.8198
q101 *	4-소음	55.0	35.2	47.3	77.0***	0.0001
q102	1-태양	15.0	7.7	8.2	11.5	0.5635
q103	2-소양	60.0	40.7	50.0	45.5	0.3211
q104	3-태음	30.0	26.4	22.6	17.0	0.2369
q105	4-소음	30.0	22.0	24.7	22.4	0.8488
q106	2-소양	15.0	4.4	11.6	8.5	0.2094
q107	4-소음	0.0	0.0	3.4	3.6	0.2567
q108	1-태양	40.0	17.6	26.7	23.0	0.1360
q109	2-소양	40.0	19.8	26.7	29.1	0.2117
q110	3-태음	5.0	30.8	32.9	18.2	0.0022
q111 *	4-소음	65.0	55.0	57.5	68.5***	0.1055
q112 *	4-소음	40.0	39.6	43.2	55.8	0.0392
q113	1-태양	40.0	39.6	44.5	37.6	0.6594
q114 *	3-태음	55.0	52.8	54.1	53.9	0.9962
q115	3-태음	35.0	36.3	31.5	27.3	0.4908
q116	4-소음	30.0	38.5	37.0	28.5	0.2883
q117	3-태음	30.0	24.2	39.7	30.9	0.0858
q118	3-태음	60.0	53.9	56.2	52.1	0.8520
q119	2-소양	50.0	35.2	45.2	38.2	0.3165
q120	4-소음	45.0	34.1	30.8	32.1	0.6345
q121	3-태음	15.0	6.6	9.6	4.9	0.2225

※위 표에서 설문번호 뒤의 *는 선택된 변수이고 *가 없는 설문은 제거된 문항이다.
※각 셀의 *는 [0,1]간 비율차이검정에서 5% 수준에서 유의한 경우는 *, 1% 수준에서 유의한 경우는 **, 0.1% 수준에서 유의한 경우는 ***이다.