Proceedings of Joint Conference of Korean Data And Information Science Society
and The Korean Data Analysis Society,
April 30 - May 1, 2004, pp. 61-68

한의학에서의 사상체질판별함수 개발에 관한 연구 (I)
- 크론박 알파 계수에 의한 변수선택 -

김규곤[1] · 최승배[2]

요 약

본 논문에서는 한방병원에서 사상체질분류검사설문지를 이용하여 사상체질을 진단할 때 진단의 정확도를 향상시키기 위한 사상체질분류함수를 개발하기 위하여 데이터마이닝에서의 판별분석모형을 이용한다. 데이터 정제 과정에서 불성실한 응답자를 제거시키기 위한 기준은 상반되는 설문의 응답 패턴과 체질별 설문의 응답 비율을 이용하며, 변수선택의 기준은 상관분석의 크론박 알파 계수와 선형판별함수의 계수를 이용한다.

주요용어 : 사상체질분류함수, 데이터마이닝, 판별분석모형, 데이터 정제, 크론박 알파.

1. 서론

사상체질의학은 東武 李濟馬가 완성한 1880년의 '格致藁'(李濟馬, 2000), 1894년의 '東醫壽世保元'(李濟馬, 1986)에서 정립한 한의학의 한 분야이다. 사상체질의학에서는 인간의 체질을 太陽人, 少陽人, 太陰人, 少陰人의 四象으로 定義하고 있으며(이태호, 1961), 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(송일병, 1993). 본 연구의 목적은 한방병원에서 사상체질분류검사설문지(QSCC II : Questionnaire of Sasang Constitution Classification)를 이용하여 사상체질을 진단할 때 진단의 정확도를 향상시키기 위한 새로운 사상체질분류함수를 개발하는 것이다. 이를 위하여 본 연구에서는 데이터마이닝 기법 중에서 판별분석모형을 이용한다.

2. 연구대상

2.1 설문지 구성

본 연구에서 사용하는 사상체질분류검사설문지는 인적사항 11개 문항과 함께 사상체질을 분류

[1](교신저자) 614-714 부산시 부산진구 가야동 산24, 동의대학교 정보통계학과 교수.
E-mail : kkkim@deu.ac.kr
[2]614-714 부산시 부산진구 가야동 산24, 동의대학교 정보통계학과 전임강사. E-mail : csb4851@deu.ac.kr

하는데 사용되는 121개 설문으로 구성되어 있다. 특히 사상체질을 분류하기 위한 121개 설문 중 15개 설문은 2개, 3개, 4개의 항목 중에서 1개를 선택하는 문항이고, 나머지 106개 설문은 '그렇다'(○) 또는 '아니다'(×) 중에서 어느 한가지에 체크하는 이항 문항이다. 여기서 전자의 15개 설문에 대해서는 특정 체질인 사람은 1개의 항목만을 선택할 수 있기 때문에, 이들을 이항 문항으로 표현하면 15개 설문은 항목 수만큼의 이항 문항으로 대체된다. 따라서 QSCC II는 ○× 중에서 한 가지를 선택하는 157개의 이항 문항으로 구성된 사상체질검사설문지라고 할 수 있다.

2.2 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.

2.3 데이터 정제 및 전처리

1) 데이터 정제 : 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 ⓪이나 ⑤, ⑥ 등으로 입력되어 있는 경우는 모두 결측값(missing data)으로 처리한다. 16번~121번 설문의 응답결과를 입력할 때 체크(√)하면 1, 아니면 0이어야 하는데도 불구하고 ⊙이나 ② 등으로 입력되어 있는 경우는 모두 0으로 수정한다.

2) 데이터 탐색 및 변환

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 설문은 ①②③ 세 항목 중에서 1개를 선택하는 문항인데, 이 경우 만약 ①에 체크하면 q1_1=1, q1_2=0, q1_3=0으로 하고, ②에 체크하면 q1_1=0, q1_2=1, q1_3=0으로 하고, ③에 체크하면 q1_1=0, q1_2=0, q1_3=1로 변환시킨다. 이 경우 q1이 본래 결측값이었던 응답자에 대해서는 q1_1, q1_2, q1_3을 모두 0으로 수정한다.

3. 사상체질분류함수

데이터마이닝의 기법에는 일반적으로 통계학에서의 다변량 분류 기법들을 포함하여 연관성(associations), 군집분석(clustering), 의사결정나무(decision trees), 신경망모형(neural networks)과 같은 기법들이 있다(Berry와 Linoff, 1997; Breiman 등, 1984). 본 논문에서는 데이터마이닝 기법 중 판별분석모형을 적용하여 사상체질분류함수를 추정한다(김규곤, 2003a; 2003b).

3.1 모든 변수 모든 환자에 대한 판별함수의 추정

판별함수를 추정하는 가장 첫 번째 단계는 모든 변수 157개와 모든 환자 1051명에 대한 응답 데이터를 이용한다. 각 체질별 변수의 개수와 환자 수는 <표 3.1-1>과 같다. <표 3.1-1>에서 체질별 변수의 합계가 158인 이유는 157개의 설문 중에는 1개의 문항이 두 개 이상의 체질에 중복하여 속하는 것도 있으며, 어떠한 체질에도 포함되지 않는 것도 있기 때문이다.

<표 3.1-1> 체질별 변수의 개수와 환자 수

	태양	소양	태음	소음	합계
체질별 변수의 개수	34	38	41	45	158
환자 수	34 3.24	254 24.17	389 37.01	374 35.59	1051 100.0

모든 변수 157개를 사용하여 모든 환자 1051명을 사상체질의 4집단으로 분류하기 위하여 판별분석모형을 적합한 결과는 <표 3.1-2>와 같으며, 정분류율(accuracy)이 74.47%, 오분류율(error rate)이 25.53%로서 정분류율이 다소 낮아 만족스럽지 못한 것으로 나타났다.

그리고 정분류율을 체질별로 살펴보면 소음인 78.34%, 소양인 75.59%, 태양인 73.53%, 태음인 70.44%의 순으로 높게 나타났다.

<표 3.1-2> 모든 변수 · 모든 환자인 경우 사상체질분류표 <도수/열%>

	태양	소양	태음	소음	합계
정분류율	25 73.53	192 75.59	274 70.44	293 78.34	774 74.47
오분류율	9 26.47	62 24.41	115 29.56	81 21.66	267 25.53

3.2 상반되는 설문의 응답패턴으로 불량응답자 1차 제거 후 판별함수의 추정

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 설정을 위한 학습표본으로 활용하고, 제거된 응답자들은 추정된 모형의 타당성을 평가하기 위한 검정표본으로 활용할 수도 있다(김규곤과 조민형, 2004).

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 체크하거나 모두 체크하지 않은 응답자는 불량응답자로 간주하여 사상체질분류함수를 구할 때까지의 분석에서는 제외시킨다. 사상체질분류검사설문지에서 서로 상반되는 설문으로 판단되는 것은 q25↔q30, q26↔q27, q28↔q29, q61↔q101 등으로서 그 내용은 <표 3.2-1>과 같다.

<표 3.2-1> 상반되는 설문

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 중요시 한다.	↔	29. 외면적인 것보다 내면적인 것을 더 중요시 한다.
61. 남성적인 면이 많고 여성적인 면이 적다.	↔	101. 여성적인 면이 많고 남성적인 면이 적다.

위와 같이 상반되는 설문의 응답패턴으로 불량응답자를 1차 제거시킨 후에 얻은 표본은 1051명 중 422명이었다. 두 번째 단계로서 모든 변수 157개를 사용하여 선택된 환자 422명을 사상체질의 4집단으로 분류하기 위하여 판별분석모형을 적합한 결과는 <표 3.2-2>와 같으며, 정분류율(accuracy)이 87.44%, 오분류율(error rate)이 12.56%로서 정분류율이 앞의 첫 번째 단계보다는 향상되었으나 역시 다소 낮은 것으로 나타났다.

그리고 정분류율을 체질별로 살펴보면 소양인 92.31%, 소음인 91.52%, 태양인 85.00%, 태음인 80.14%의 순으로 높게 나타났다.

<표 3.2-2> 모든 변수 · 422명 환자인 경우 사상체질분류표 <도수/열%>

	태양	소양	태음	소음	합계
정분류율	17 85.00	84 92.31	117 80.14	151 91.52	369 87.44
오분류율	3 15.00	7 7.69	29 19.86	14 8.48	53 12.56

3.3 크론박 알파 계수와 선형판별함수의 계수를 이용한 변수선택 후 판별함수의 추정

여기서는 상관분석(correlation analysis)에서의 크론박 알파 계수(Cronbach alpha coefficient)와 선형판별함수의 계수를 이용하여 변수를 선택한 후 앞에서 선택된 환자 422명을 대상으로 판별함수를 추정한다(김규곤, 1999; 전란희 등, 1999; 김규곤 등, 1999).

1) 크론박 알파 계수를 이용한 변수선택 기준

각 체질별 설문에 대하여 체질별로 구한 크론박 알파 계수를 기준으로 변수선택을 한다. 크론박

알파 계수를 구하는 절차를 단 한번 실행하는 것으로는 양질의 설문만을 찾아내기 어렵기 때문에, 더 이상 제거되는 설문이 없을 때까지 크론박 알파 계수를 구하는 절차를 계속적으로 반복하는 소위 변수소거법(backward elimination)을 적용하여 변수를 선택한다. 각 체질별 반복회수는 태양체질 설문은 7회, 소양체질 설문은 5회, 태음체질 설문은 4회, 소음체질 설문은 4회였다.

2) 선형판별함수의 계수를 이용한 변수선택 기준

첫째, 선형판별함수의 계수가 무한대 값인 변수를 제거한다. 이와 같은 변수는 사상체질분류함수 값에 너무 큰 영향을 주어 이러한 변수들의 체크 여부가 체질분류에 결정적 역할을 하기 때문이다.

둘째, 선형판별함수의 계수가 해당 체질에서 음수인 변수를 제거한다. 이와 같은 변수는 해당 체질인 사람이 체크하면 오히려 역작용을 하여 함수 값을 작게 함으로서 그 체질이 아니라는 결정을 하기 때문이다.

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. 이상의 결과를 정리한 <표 3.3-1>을 보면 선택된 설문은 태양체질 6개, 소양체질 8개, 태음체질 14개, 소음체질 15개로 총 43개 변수가 선택되었으며, 이 변수들은 서로 다른 체질간에 겹치지 않는다.

<표 3.3-1> 변수선택의 결과

	태양	소양	태음	소음
선택된 설문	q15_1 q17 q18 q34 q35 q83	q12_2 q15_2 q19 q20 q28 q55 q58 q69	q1_1 q2_1 q5_1 q6_1 q11_1 q13_3 q22 q50 q70 q76 q80 q88 q114 q118	q1_3 q2_3 q5_2 q10_2 q23 q24 q33 q52 q56 q59 q77 q81 q96 q111 q112
설문 수	6	8	14	15

세 번째 단계로서 크론박 알파 계수와 선형판별함수의 계수를 이용하여 선택된 변수 43개를 사용하여 앞에서 선택된 환자 422명을 사상체질의 4집단으로 분류하기 위하여 판별분석모형을 적합한 결과는 <표 3.3-2>와 같으며, 정분류율(accuracy)이 71.06%, 오분류율(error rate)이 28.94%로서 정분류율이 앞의 두 가지 경우보다 더 낮아진 것으로 나타났다.

그리고 정분류율을 체질별로 살펴보면 소음인 78.18%, 태음인 71.23%, 태양인 70.00%, 소양인 64.84%의 순으로 높게 나타났다.

<표 3.3-2> 43개 변수 · 422명 환자인 경우 사상체질분류표 <도수/열%>

	태양	소양	태음	소음	합계
정분류율	14 70.00	59 64.84	104 71.23	129 78.18	306 71.06
오분류율	6 30.00	32 35.16	42 28.77	36 21.82	116 28.94

3.4 선택된 변수와 모든 환자에 대한 판별함수의 추정

네 번째 단계로서 선택된 변수 43개를 사용하여 모든 환자 1051명을 사상체질의 4집단으로 분류하기 위하여 판별분석모형을 적합한 결과는 <표 3.4-1>과 같으며, 정분류율(accuracy)이 62.51%, 오분류율(error rate)이 37.49%로서 이제까지의 세 가지 경우보다 더 낮아진 것으로 나타났다.

그리고 정분류율을 체질별로 살펴보면 소음인 68.45%, 태양인 64.71%, 태음인 63.75%, 소양인 53.15%의 순으로 높게 나타났다.

<표 3.4-1> 43개 변수 · 1051명 환자인 경우 사상체질분류표 <도수/열%>

	태양	소양	태음	소음	합계
정분류율	22 64.71	135 53.15	248 63.75	256 68.45	661 62.51
오분류율	12 35.29	119 46.85	141 36.25	118 31.55	390 37.49

3.5 체질별 설문의 응답비율로서 불량응답자 2차 제거 후 판별함수의 추정

앞에서 선택된 43개의 변수에 대하여 422명의 응답결과가 각 체질별로 결측값(missing data)의 비율이 크면 또다시 불량응답자로 간주하여 2차 제거시킨 후 판별함수를 추정한다. 2차로 제거시킬 불량응답자의 기준은 다음과 같다.

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 응답비율은 50%이다. 따라서 응답비율이 50%보다 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.

이와 같이 체질별 설문의 응답비율로서 불량응답자를 2차 제거시킨 후에 선택된 환자는 422명 중 188명이었다. 다섯 번째 단계로서 앞에서 선택된 변수 43개를 사용하여 선택된 환자 188명을 사상체질의 4집단으로 분류하기 위하여 판별분석모형을 적합한 결과는 <표 3.5-1>과 같으며, 정분류율(accuracy)이 98.40%, 오분류율(error rate)이 1.60%로서 정분류율이 이제까지의 모든 4가지 경우보다 가장 높은 것으로 나타났다(김규곤과 조민형, 2004).

그리고 정분류율을 체질별로 살펴보면 태양인과 소양인 100.0%, 소음인 98.26%, 태음인 98.0%의 순으로 높게 나타났다.

<표 3.5-1> 43개 변수 · 188명 환자인 경우 사상체질분류표 <도수/열%>

		실제 체질				
		태양	소양	태음	소음	합계
예측된 체질	태양	8 100.00	0 0.00	1 2.00	2 1.74	11 5.85
	소양	0 0.00	15 100.00	0 0.00	0 0.00	15 7.98
	태음	0 0.00	0 0.00	49 98.00	0 0.00	49 26.06
	소음	0 0.00	0 0.00	0 0.00	113 98.26	113 60.11
	합계	8 100.0	15 100.0	50 100.00	115 100.00	188 100.00
정분류율		8 100.00	15 100.00	49 98.00	113 98.26	185 98.40
오분류율		0 0.00	0 0.00	1 2.00	2 1.74	3 1.60

4. 결론

본 논문에서는 사상체질분류검사설문지(QSCC II : Questionnaire of Sasang Constitution Classification)를 이용한 체질진단에서 그 정확도를 향상시키기 위한 새로운 분류함수를 구하기 위하여 데이터마이닝 기법 중 판별분석모형을 이용하였다. 본 연구에서 사용하는 데이터는 경희대학교 강남경희한방병원과 동의대학교 부속한방병원에서 치료를 받은 환자들 중 각 병원의 사상체질 전문의로부터 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 첫째로 상반되는 설문의 응답 패턴과 둘째로 체질별 설문의 응답 비율을 이용하였으며, 변수선택의 기준은 상관분석의 크론박 알파 계수와 선형판별함수의 계수를 이용하였다.

본 연구에서는 환자를 사상체질의 4집단으로 분류하기 위한 사상체질분류함수를 추정하기 위하여 5가지 단계로 나누어 고찰하였다.

1) 157개의 모든 변수를 사용하여 모든 환자 1051명에 대한 사상체질분류함수는 정분류율(accuracy)이 74.47%로서 다소 낮았다. 체질별로는 소음인 78.34%, 소양인 75.59%, 태양인 73.53%, 태음인 70.44%의 순으로 높게 나타났다.

2) 상반되는 설문의 응답 패턴으로 불량응답자 1차 제거 후 157개의 모든 변수와 422명의 선택된 환자에 대한 사상체질분류함수는 정분류율(accuracy)이 87.44%로서 정분류율이 1)의 경우보다는 향상되었으나 역시 다소 낮은 것으로 나타났다. 체질별로는 소양인 92.31%, 소음인 91.52%, 태양인 85.00%, 태음인 80.14%의 순으로 높게 나타났다.

3) 크론박 알파 계수와 선형판별함수의 계수를 이용한 변수선택 후 선택된 변수 43개와 선택된

환자 422명에 대한 사상체질분류함수는 정분류율(accuracy)이 71.06%로서 정분류율이 앞의 두 가지 1), 2)의 경우보다 더 낮아진 것으로 나타났다. 체질별로는 소음인 78.18%, 태음인 71.23%, 태양인 70.00%, 소양인 64.84%의 순으로 높게 나타났다.

4) 선택된 변수 43개와 모든 환자 1051명에 대한 사상체질분류함수는 정분류율(accuracy)이 62.51%로서 앞의 세 가지 1), 2), 3)의 경우보다 더 낮아진 것으로 나타났다. 체질별로는 소음인 68.45%, 태양인 64.71%, 태음인 63.75%, 소양인 53.15%의 순으로 높게 나타났다.

5) 체질별 설문의 응답비율로서 불량응답자 2차 제거 후 선택된 변수 43개와 선택된 환자 188명에 대한 사상체질분류함수는 정분류율(accuracy)이 98.40%로서 앞의 모든 4가지 경우보다 정분류율이 가장 높은 것으로 나타났다. 체질별로는 태양인과 소양인 100.0%, 소음인 98.26%, 태음인 98.0%의 순으로 높게 나타났으며, 태양인, 소양인, 소음인, 태음양인의 모든 체질에서 90% 이상이었다.

본 연구의 결과 판별분석모형을 이용하여 사상체질분류함수를 추정하고자 하는 경우 모든 변수와 모든 환자의 데이터를 사용하기보다는 양질의 응답 데이터를 확보하기 위한 적절한 데이터 정제 방법과 함께 적절한 변수선택 방법을 사용해야 한다.

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

참고문헌

[1] 김규곤 (1999). 이산 다변량 분석을 이용한 한방 진단 프로그램 개발 연구, *Journal of The Korean Data Analysis Society*, Vol. 1, No. 1, pp. 15-27.

[2] 김규곤 (2003a). 데이터마이닝에서의 분류방법에 관한 연구, *Journal of the Korean Data Analysis Society*, Vol. 5, No. 1, pp. 101-112.

[3] 김규곤 (2003b). 한방 통계분석방법에 관한 사례연구, *Journal of the Korean Data Analysis Society*, Vol. 5, No. 4, pp. 907-917.

[4] 김규곤, 강창완 (1999). 한의학에서의 변증점수개발에 대한 가중주성분분석의 응용, *응용통계연구*, 12(1), pp. 17-28.

[5] 김규곤, 조민형 (2004). 사상체질 판별함수의 개발에 관한 연구, *Journal of the Korean Data Analysis Society*, Vol. 6, No. 1, pp. 303-315.

[6] 송일병 (1993). *알기쉬운 사상의학*, 서울, 사상사, pp. 50-89.

[7] 李濟馬 (2000). *格致藁*, 서울, 청계출판, p. 10.

[8] 李濟馬 (1986). *東醫壽世保元*, 서울, 행림출판, pp. 137-142.

[9] 이태호 (1961). *실제적 동의사상진료의 비결*, 서울, 행림서원, pp. 31-47.

[10] 전란희, 이인선, 김규곤, 강창완 (1999). 한방 부인과 자료에서의 수량화분석, *Journal of The Korean Data Analysis Society*, Vol. 1, No. 1, pp. 53-63.