맥파를 이용한 사상체질의 진단에 있어서 분류방법에 따른 진단의 정확도 비교

Comparisons of the Accuracy of Classification Methods in Sasang Constitution Diagnosis with Pulse Waves

신상훈*, 김종열**
상지대학교 한방의료공학과*, 한국한의학연구원 체질의학연구본부**

Sang-Hoon Shin(shshin@sangji.ac.kr)*, Jong-Yeol Kim(ssmed@kiom.re.kr)**

요약

사상의학은 체질에 따라 치료하는 방법을 달리하므로, 체질진단의 객관화가 절실히 요구되고 있다. 본 연구는 맥파를 이용하여 사상체질을 객관적으로 진단함에 있어서, 정확도가 높으면서 실용적인 체질분류 방법을 탐색하는 것이 목적이다. 한방병원에 건강검진을 목적으로 내원한 2848명의 피험자를 대상으로 전문의가 진단한 체질, 체질량지수, 혈압, 맥파 자료를 입수하였다. 자료의 선별과정을 통하여 최종적으로 1635명의 자료를 분석에 사용하였다. 판별분석, 회귀분석, 의사결정나무, 신경망으로 체질을 예측하고 전문의가 진단한 결과와 비교하여 분류방법의 정확도를 비교하였다. 판별분석은 체질별로 공분산 행렬이 동일해야 한다는 가정을 만족시키기 어려웠으며, 체질량지수를 고려하지 않은 의사결정나무와 신경망 분석의 결과는 분석표본의 변동에 민감했다. 체질분류에 결정적인 영향을 미치는 변수인 체질량지수가 고려된 로지스틱 회귀분석 또는 의사결정나무 방법이 체질분류 방법으로 추천할 만하다.

■ 중심어 : | 맥파 | 사상체질 | 데이터마이닝 |

Abstract

The purpose of this study is to find a classification method with high accuarcy in regard with sasang constitutional diagnosis. The BMI, blood pressure, pulse wave, and Sasang constitution diagnosed by a specialist was collected from 2848 subjects who were apparently healthy. Through a selective procedure, the data of 1635 subjects was used in the analysis. The results with the classification methods such as the discriminant analysis, regression, decision tree and neural network were compared with the diagnosis of a Sasang constitutional specialist. In result, the discriminant analysis method was hard to qualify the assumption of the equality of covariance matrices within constitutional groups. Moreover, without BMI, the decision tree and neural network methods were very sensitive to the change of the analysis data. Therefore, the Logistic regression and the decision tree is recommended on condition that the decisive factors of constitution are well concerned.

■ keyword : | Pulse Wave | Sasang Constitution | Data Mining |

* 본 연구는 지식경제부 차세대기술개발사업 중 지능형 한방 컨텐츠 개발(10028438)에 의해 이루어졌습니다.

접수번호 : #090911-001
접수일자 : 2009년 09월 11일
심사완료일 : 2009년 09월 17일
교신저자 : 신상훈, e-mail : shshin@sangji.ac.kr

I. 서 론

사상의학은 체질을 태양인, 소양인, 소음인, 태음인의 네가지로 나누고, 체질에 따라 질병을 예방하고 치료하는 방법을 달리하고 있다. 그러므로 체질의 정확한 판단은 매우 중요하다. 전통적인 체질 진단방법은 체형기상(體刑氣像), 용모사기(容貌詞氣), 성질재간(性質材幹), 병증약리(病證藥理)의 4가지 결과를 종합하여 의사가 최종적으로 판단하는 것이다. 체형기상은 장부가 존재하는 체간부위의 형태를, 용모사기는 얼굴의 형태와 목소리를, 그리고 성질재간은 성격과 행동특성을 관찰하는 것이다 마지막으로 병증약리는 약물에 대한 인체의 반응을 관찰한다. 병증약리를 제외한 나머지 진단방법은 한의사의 주관이 개입될 수 있는 부분들이 많아 체질진단의 객관화가 절실히 요구되고 있다.

의사의 주관적인 감각과 판단으로 이루어졌던 사상체질 진단과정을 객관화하려는 연구들이 활발히 진행되고 있다. 체형기상의 진단 객관화를 위하여 3차원 스캐너를 이용한 체간형상 측정의 자동화가 시도되었으며[1], 용모사기 진단의 객관화를 위하여 정면과 측면사진을 이용하여 두면부의 형태학적 특징을 정량화하거나[2], 음성분석기를 이용하여 음성의 고저, 강약, 빠르기 등의 특성과 체질과의 상관성을 연구하기도 하였다[3]. 성질재간의 객관적 진단을 위하여 개발된 설문지는 표준화과정을 거쳐서 널리 사용되고 있다[4].

맥진은 손가락 감각을 이용하여 경맥의 박동상태를 관찰함으로써 장부와 경락의 상태를 판단하는, 변증시치의 중요한 수단이다. 맥진은 설진과 함께 한의학적 진단에 가장 널리 사용되는 진단방법의 일부이다. 맥진기를 이용하여 체질을 판별하려는 기존의 연구들이 있었으나[5-7], 체질별 표본의 크기가 작고 연구자의 경험적 확신을 통하여 입력변수를 선정한 경우가 많았다. 체질진단의 객관화연구에서 가장 널리 사용되는 분류방법은 판별분석인데[8], 이는 판별분석에서 제공되는 분류함수가 체질과 입력변수들 사이의 명확한 관계를 제공하기 때문이라고 생각된다. 그러나 판별분석을 사용하기 위해서는 까다로운 기본가정을 만족해야 하며, 지금까지의 연구결과에 의하면 체질진단의 정확도도 매우 낮았다[9].

그러므로 본 연구에서는 체질진단의 정확도를 향상시키면서, 적용하기에 까다롭지 않고 실용적인 새로운 체질분류 방법을 모색하여 보고자 한다.

II. 연구방법

1. 맥파신호의 특성

인체의 손목부위인 요골동맥에서 측정되는 맥파의 대표적인 형상은 [그림 1]과 같다[10].

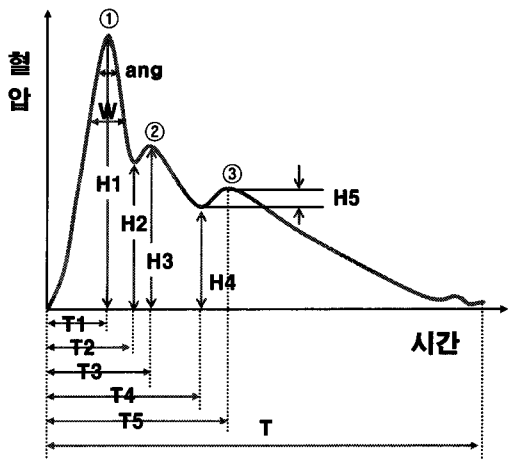


그림 1. 맥파의 특징을 나타내는 변수

일반적으로 맥파에서는 충격파(①), 조랑파(②), 중복파(③)라 부르는 3개의 봉우리를 관찰할 수 있다. 심장의 수축에 의하여 충격파(①)가 생성되며, 반사파의 중첩영향과 대동맥 밸브의 닫힘(T4)으로 인하여 조랑파(②)와 중복파(③)의 형상이 달라진다. H는 맥파의 상대적인 크기를 나타내며, T는 대응되는 시간을 나타낸다. 심장은 수축과 이완을 주기적(맥파주기:T)으로 반복한다. 수축기시간(T4) 동안 심장은 혈액을 동맥으로 내보내며, 이완기 시간(T-T4) 동안 이완된 심실에 혈액을 공급한다. [그림 1]에서 압력과 시간을 곱한 그래프의 면적을 맥파면적(A)이라고 하는데, 이는 맥동이 가지는 충격량을 나타낸다. 심장의 생리적 특성을 고려하여, 맥

파면적은 수축기면적(AS)과 이완기면적(AD)으로 나눈다. 충격파(①)는 심장의 수축특성을 나타내는 주요한 지표이므로 많은 변수정의를 필요로 한다. H1은 맥파의 최대크기를 나타내며, 주파의 형태적인 특성을 나타내기 위해 ANG를, 주파의 유효면적을 나타내기 위하여 W를 정의하였으며, 주파의 유효 충격량을 나타내기 위하여 AW를 정의하였다.

2. 자료수집

2.1 맥파신호의 획득

[그림 2]에 나타난 맥진기(3-D Mac, (주)대요메디)를 이용하여 피험자의 좌/우 관부(關部: 요골 경상돌기 위치에 있는 요골동맥)에서 맥파신호를 측정하였다.

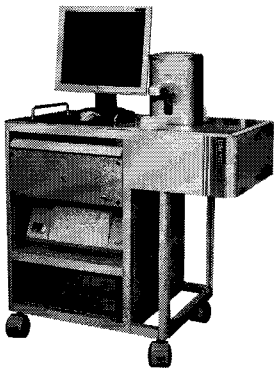


그림 2. 맥파 측정장비

2.2 체질의 진단

체질진단의 정확도를 비교하는데 있어서 기준이 되는 것은 의사의 진단소견이다. 본 연구에서는 사상체질 전문의가 사상체질 음성분석기[3]의 측정결과와 체형기상, 용모사기, 성질재간의 주관적인 평가를 통하여 피험자의 체질을 결정하였다.

3. 자료의 처리

3.1 분석자료의 선별과정

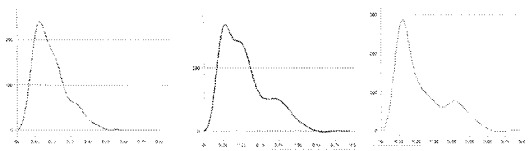
한방병원에 건강검진을 목적으로 내원한 2848명을 대상으로 체질량지수, 혈압, 체질과 맥파 자료를 수집하였다. 해석의 정확도를 높이기 위하여 [표 1]과 같이 분석에 사용될 자료를 선별하였다. 체질, 연령, 맥파 중 하나의 항목이라도 누락된 자료(540)를 제외하였으며, 집단의 크기가 매우 작은 태양인(2)도 제외하였다. 맥파신호의 측정상태가 불량한 경우(368)를 제외하였으며, SPSS의 Box Plot기능을 이용하여 이상치가 발생한 경우(87)를 제외하였다.

표 1. 자료의 선별과정

순서	내 용	최종 자료갯수
1	맥파 파일 입수	2848
2	분석자료 추출 및 정리	2308
3	태양인 제외	2306
4	측정불량 제외	1938
5	이상치 제거	1851

3.2 분석집단의 선정

[그림 3]은 태음인 여성에게서 측정한 맥파이다. 성별과 체질이 동일함에도 불구하고, 연령에 따라 맥파의 형상이 달라짐을 알 수 있다. 이는 노화에 따른 동맥의 경화현상이 맥파 전달속도를 증가시키고, 맥파 전달속도의 증가가 진행파와 반사파의 중첩시기를 단축시켜, 맥파형상이 변형된 결과이다[10]. 맥파를 이용한 체질분류에서는 맥파의 형상이 매우 중요하므로, 분석집단을 선정할 때는 연령대의 범위를 좁히는 것이 바람직하다. [표 1]에서 선정된 표본에서는 20대 이하(4%)와 60대 이상(7.8%)의 비율이 작았으므로, 본 연구에서는 분석집단을 30대, 40대, 50대로 한정하였다.



79세 45세 18세

그림 3. 연령에 따른 맥파의 모양 (태음인, 여성)

노화에 따른 맥파의 형상변화를 고려하여 20대 이하와 60대 이상의 데이터(216)를 제외하였으므로, 최종적으로

1635명의 자료가 분석에 사용되었다. [표 2]는 분석에 사용된 자료의 체질, 연령, 남녀의 분포를 나타낸다.

표 2. 분석표본의 체질, 연령, 남녀 분포

		나 이			남 자	여 자	합 계
		30-39	40-49	50-59			
소양인	N	67	177	116	151	209	360
	%	19	49	32	42	58	22
소음인	N	187	249	122	293	265	558
	%	34	45	22	53	47	34
태음인	N	202	323	192	501	216	717
	%	28	45	27	70	30	44
합 계	N	456	749	430	945	690	1635
	%	28	46	26	58	42	

3.3 입력변수의 선정 및 축소

맥진기를 이용하여 좌우의 관부(關部)의 맥파를 측정하였으나, 인체는 좌우대칭이므로 본 연구에서는 좌측의 맥파만 사용하였다. [표 3]은 체질분류에 사용될 입력변수를 나타낸다. BMI, PS, PD는 맥진기로부터 추출되는 자료는 아니나, 통상적으로 신체검사에서 제공되는 중요한 지표이므로 본 연구에 포함하였다,

표 3. 입력변수

변수	설명	변수	설명
BMI	체질량지수	T1	충격파 시간
PS	수축기 혈압	T2	조량파협 시간
PD	이완기 혈압	T3	조량파 시간
HR	맥박수	T4	절흔 시간
PRS	가압력	T5	중복파 시간
ENG	맥파 에너지	TD	이완기 시간
D_PRS	좌우 가압력 차이	T	맥파주기
D_ENG	좌우 맥파에너지 차이	W	주파너비
H1	충격파 높이	A	맥파면적
H2	조량파협 높이	AS	수축기면적
H3	조량파 높이	AD	이완기면적
H4	절흔 높이	AW	W 면적
H5	중복파 높이	ANG	주파각

H3와 T3는 맥진기에서 제공된 특징점 추출기능의 문제점으로 인하여 고려대상에서 제외하였다. 체질판별을 위하여 사용되는 입력변수는, 체질집단에 따라 평균값이 서로 달라야 한다. 그러므로 소양인, 소음인, 태음인으로 구별되는 세집단의 평균값을 비교할 수 있는 일원분산분석(one-way ANOVA)을 입력변수들에 적용하였다. 분석결과 체질집단의 변동을 감지할 수 없는, 즉 유의확률>0.05인, 입력변수 4개 (H4_N, TD, T, D_ENG)를 제외하여 최종적으로 20개의 입력변수로 압축하였다.

표 4. 입력변수의 일원분산분석 결과

변수	유의확률	변수	유의확률
BMI	0.000	T4	0.000
PS	0.000	T5	0.000
PD	0.000	TD	0.162
HR	0.000	T	0.061
PRS	0.000	W	0.000
ENG	0.000	A	0.031
H1	0.026	AS	0.000
H2_N	0.000	AD	0.000
H4_N	0.253	AW	0.000
H5_N	0.007	ANG	0.042
T1	0.000	D_PRS	0.002
T2	0.002	D_ENG	0.125

입력변수 중에서 H2_N, H4_N, H5_N은 각각 H2, H4, H5를 충격파의 높이(H1)로 나누어서 정규화한 것이다. 일반적으로 맥파높이(H)는 맥파의 기저선을 설정하는 알고리즘에 따라 상대적으로 변할 수 있으므로, 맥파높이에서 가장 큰 값을 가지는 H1을 이용하여 정규화하는 것이 바람직하다.

III. 분석결과

최종적으로 선정된 입력변수를 판별분석, 회귀분석, 의사결정나무분석, 신경망분석에 적용하여 체질분류의 정확도를 비교하였다. 판별분석에는 한글 SPSS 14.0을 사용하였으며, 회귀분석, 의사결정나무분석, 신경망분석에는 SPSS사에서 개발한 Clementine 12.0 을 사용하였다.

1. 판별분석

[표 5]는 각각의 체질집단에 대한 공분산행렬의 동질성을 검사한 결과이다. Box's M 검증에서 유의확률 <0.05가 의미하는 것은 집단별 분산분포가 동일하지 않음을 나타내며, 이는 분산이 큰 집단으로 잘못 분류될 가능성이 커진다는 것을 의미한다[11].

표 5. 공분산 동일성 검증

Box의 M		564.943
F	근사법	1.461
	자유도1	380
	자유도2	4102096.5
	유의확률	.000

세 개의 집단을 구별하므로 두 개의 판별함수가 도출되었다. [표 6]은 판별함수의 고유값에 관한 결과이며, 판별함수(1)이 체질구분을 대부분 설명(96.3%)하고 있다.

표 6. 고유값

함수	고유값	분산의 %	누적 %	정준상관
1	.981	96.3	96.3	.704
2	.037	3.7	100.0	.190

[표 7]은 판별함수를 이용하여 각 체질별 중심값을 구한 결과인데, 판별함수(1)은 태음인과 기타체질을 구분하고, 판별함수(2)는 소양인과 소음인을 구분한다.

표 7. 함수의 집단중심점

체질	함수	
	1	2
태음인	1.067	-.066
소양인	-.337	.358
소음인	-1.154	-.146

본 연구에서는 체질구분의 대부분을 설명하고, 태음인과 기타체질을 구분하는 판별함수(1)에 초점을 맞추었다. 판별함수(1)에 영향을 입력변수를 조사하기 위하여 구조행렬을 구하였다. [표 8]의 구조행렬에 의하면 판별함수(1)에 영향을 미치는 변수는 BMI, PS, PD, ENG이며, 나머지 변수는 판별함수(2)와 관련이 있다. 판별함수(1)에 가장 큰 기여도를 가지는 입력변수는 체질량지수(BMI)이다.

표 8. 구조행렬

	함수	
	1	2
BMI	.984*	-.032
PS	.296*	.005
PD	.281*	-.099
ENG	.138*	.042
AS	.111	.638*
W	.037	.571*
H2_N	.041	.558*
AW	.062	.426*
T4	.091	.426*
T1	.120	.355*
ANG	-.013	.315*
PRS	.200	.308*
H5_N	-.054	-.291*
T5	.111	.286*
A	.053	.199*
D_PRS	.082	-.154*
HR	-.101	-.135*
H1	.063	.125*
T2	.086	-.107*

판별함수(1)에 영향을 미치는 입력변수의 특성을 자세히 알아보기 위하여 체질별 입력변수의 평균값을 [표 9]에 정리하였다.

표 9. 입력변수의 체질별 평균

변수명	태음인	소양인	소음인
BMI	25.7	22.6	20.9
PS	125	119	115
PD	77	73	71
ENG	452	418	392

2. 데이터마이닝 분석 (전체표본)

판별분석은 종속변수와 독립변수들 사이의 관계를 명확하게 보여준다는 장점이 있는 반면 집단별 정규성

과 등분산성 가정을 만족해야 하므로 적용에 어려움이 많다. 본 연구에서는 SPSS사에서 개발한 Clementine 12.0을 사용하여 회귀분석, 의사결정나무분석, 신경망 분석을 실시하였다.

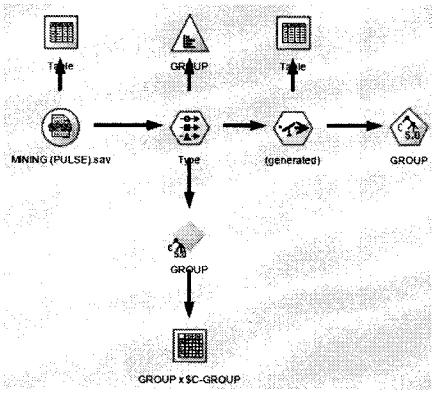


그림 4. 데이터마이닝 모델 (전체표본)

[그림 4]는 Clementine으로 구현한 데이터마이닝 모델이다. 본 연구에서는 동일한 표본으로 학습과 검증을 실시하는 경우를 전체표본이라고 하며, 학습과 검증에 사용된 표본이 서로 다른 경우를 학습-검증 표본이라고 부르기로 한다. 분류결과는 [표 10]과 같다.

표 10. 체질분류의 정확도 비교

분석방법	체 질		
	태음인	소양인	소음인
판별분석	75.7%	53.3%	69.4%
Logistic	78.0%	49.4%	70.3%
CART	78.9%	57.8%	66.1%
C5.0	86.8%	89.4%	81.9%
CHAID	80.1%	51.4%	81.7%
QUEST	69.6%	59.7%	63.3%
Neural (MLP)	74.2%	56.1%	65.8%
Neural (Quick)	75.6%	54.4%	66.7%

3. 데이터마이닝 분석 (학습-검증 표본)

체질분류의 정확도를 객관적으로 검증하기 위해서는 학습에 사용되는 표본과 검증에 사용되는 표본을 반드시 구별할 필요가 있다. 본 연구에서는 랜덤시드(random seed)를 이용하여 전체표본을 학습표본과 검증표본으로 50:50으로 랜덤하면서도 배타적으로 분리하였다. 그러므로 두 표본의 합은 전체표본이 된다. [그림 5]는 해석에 사용된 모델이다.

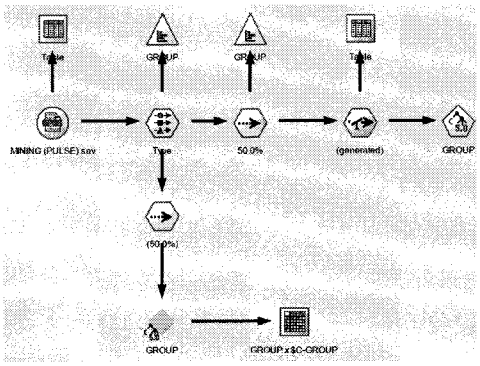


그림 5. 데이터마이닝 모델 (학습-검증 표본)

표 11. 학습-검증표본과 전체표본의 정확도비교

해 석 방 법			태음인	소양인	소음인
BMI 포함	학습 검증 표본	Logistic	75.1%	44.1%	66.1%
		C5.0	70.3%	38.7%	56.9%
		Neural (Quick)	78.2%	46.2%	66.4%
	전체 표본	Logistic	78.0%	49.4%	70.3%
		C5.0	86.8%	89.4%	81.9%
		Neural (Quick)	75.6%	54.4%	66.7%
BMI 제외	학습 검증 표본	Logistic	52.1%	37.6%	46.6%
		C5.0	49.3%	25.8%	51.6%
		Neural (Quick)	38.2%	57.5%	38.9%
	전체 표본	Logistic	56.1%	37.8%	53.9%
		C5.0	86.3%	83.9%	80.5%
		Neural (Quick)	65.4%	45.8%	37.6%

회귀분석, 의사결정나무분석, 신경망분석에서 높은 분류 정밀도를 보여준 방법들을 하나씩 선정하여 전체표본과 학습-검증 표본에서의 체질분류 정확도를 비교하였다. 본 연구에서는 판별분석의 문제점을 개선하기 위하여 다른 분류방법을 찾는 것이 목적이므로, 학습-검증 표본에 대해서는 판별분석을 제외하였다. 체질량지수는 맥진기로부터 측정될 수 있는 지표는 아니지만, [표 8]의 결과에 의하면 체질구분의 대부분을 설명하는 가장 기여도가 높은 변수이다. 그러므로 체질량지수가

체질분류의 정확도에 미치는 영향도 고려하였다.

IV. 고찰

한방병원에 건강검진을 목적으로 내원한 피험자를 대상으로 전문의가 진단한 체질과 의료기기를 사용하여 측정한 체질량지수, 혈압, 맥파 자료를 입수하였다. 자료의 선별과정을 통하여 최종적으로 1635명의 자료가 분석에 사용되었다. 분석에 사용된 표본에서 남녀의 비율은 남자(58%) 여자(42%), 연령별 비율은 30대(28%), 40대(46%), 50대(26%)이며, 체질별 비율은 소양인(22%), 소음인(34%), 태음인(44%)이었다. 판별분석, 회귀분석, 의사결정나무분석, 신경망분석을 이용하여 체질을 예측하고 전문의가 진단한 결과와 비교하여 각 분석방법들에 대하여 체질분류의 정확도를 계산하였다.

1. 판별분석

체질분류에서 가장 널리 사용되었던 방법은 판별분석인데, 이는 판별분석에서 제공되는 분류함수가 체질과 입력변수들 사이의 명확한 관계를 제공하기 때문이라고 생각된다. [표 6]과 [표 7]의 결과에 의하면, 함수 1은 태음인을 다른 체질과 구별하며 체질분류에 대하여 설명력이 96.3%인 매우 중요한 함수이다. 반면 소양인을 소음인으로부터 구분하는 함수 2가 체질분류에서 차지하는 비중(3.7%)은 매우 작은 편이다. 즉 태음인의 분류정확도가 가장 높을 것이라는 추측을 가능하게 해준다. [표 10]에 의하면 체질분류의 정확도는 태음인>소음인>소양인이었다. 이는 [표 2]의 체질별 표본크기 순서와 동일하였다. 기존의 연구[9]에서도 체질별 표본크기의 순서가 태음인(42.6%)>소음인(30.2%)>소양인(27.2%)일 때, 체질분류의 정확도의 순서가 표본크기의 순서와 동일하였다. 주어진 자료로부터 판별함수를 도출하기 위한 가정은 ① 입력변수들이 다변량 정규분포를 이루고, ② 종속변수에 의해 범주화 되는 집단들의 분산-공분산 행렬이 동일해야 한다는 것이다. 본 연구에서는 표본의 개수가 충분하므로 가정①을 위배할 염려는 없다. 문제는 가정②를 위배할 경우인데, 가정②의 위배여부를 검사하는 것이 SPSS에서 제공하는 Box's M 검정이다. [표 5]에서 유의확률<0.05가 의미하는 것은 가정②를 위배한다는 것이며, 집단별 분산분포가 동일하지 않음을 나타낸다. 가정②가 위배되면 보다 큰 분산-공분산 행렬을 갖는 그룹에 많은 관측치가 분류되는 문제가 발생된다. 즉 표본의 크기가 작을수록 체질분류의 정확도가 낮아지게 되며, [표 10]과 기존의 연구[9]에서 나타난 결과를 설명하는 이유가 된다.

구조행렬은 판별함수에 사용되는 입력변수들의 상대적인 기여도를 나타낸다. [표 8]에 의하면 태음인의 구분에 영향을 미치는 변수는 체질량지수(BMI), 수축기혈압(PS), 이완기혈압(PD), 맥파에너지(ENG) 인데 체질량지수의 기여도가 가장 크다. [표 9]에 의하면 4가지 변수(BMI, PS, PD, ENG) 모두 태음인이 다른 체질에 비하여 높은 값을 가진다. 태음인의 외형상 특징은 다른 체질에 비하여 체격이 좋고 비만하므로 체질량지수가 높다. 체질량지수가 높을수록 혈관의 중막(arterial intima-media)이 두꺼워지므로[12] 혈류저항이 높아져서 혈압(PS, PD)이 상승하게 된다. 맥파에너지에 가장 큰 영향을 미치는 요인은 최대 맥파진폭(H1)이며, 다른 연구결과[6]에 의하면 태음인의 H1값이 다른 체질에 비하여 크게 나타나고 있다. 그러므로 태음인을 구별하는 위의 특징들은 생리학적으로 매우 타당한 결과라고 할 수 있다.

2. 데이터마이닝 분석

체질분류를 위한 판별분석의 적용에서, 체질집단의 분산을 동일하게 하는 것이 필수적이다. 그러므로 본 연구의 초기과정에 있어서, 체질별 분산분포를 동일하게 유지하기 위하여 많은 시도를 하였다. 그러나 체질별로 등분산분포를 만들기 위한 수학적 변환과정이 너무나 인위적이었으며, 나아가 체질별로 분산이 다르다는 것도 체질특성의 일부일 수가 있다는 결론에 도달하게 되었다. 그러므로 체질별 분산이 동일하지 않아도 되는 분류방법을 탐색하게 되었다. 체질분류에 사용될 수 있는 방법들은 크게 회귀분석(Regression), 의사결정나무(Decision Tree), 그리고 신경망(Neural Network)이 있다. 본 연구에서는 회귀분석방법으로 로지스틱회귀

분석(Logistic)을, 의사결정나무 분석방법으로 CART, C5.0, CHAID, QUEST을 사용하였다. 신경망 분석방법으로 MLP와 Quick을 사용하였다[13].

판별분석 결과와 동일한 비교를 위하여 학습에 사용된 표본집단에 대하여 분류의 정확도를 검증하였다. [표 10]에 의하면 판별분석과 회귀분석은 매우 유사한 결과를 보이고 있다. 태음인의 분류정확도는 의사결정나무 방법이 가장 높았으며, 나머지는 서로 비슷했다. 소양인의 분류정확도는 의사결정나무 분석의 C5.0을 제외하고는 비슷했다.

체질분류의 정확도를 객관적으로 검증하기 위하여, 랜덤시드(random seed)를 이용하여 전체표본을 학습표본과 검증표본으로 50:50으로 랜덤하면서도 배타적으로 분리하였다. 세 종류의 분석방법에서 높은 분류 정확도를 보여준 방법들을 하나씩 선정한 다음, 동일한 랜덤시드값을 적용시켜 동일한 학습표본과 검증표본을 만든 다음, 각 방법들의 분류정확도를 비교하였다. 본 연구에서는 판별분석 결과와 비교하기 위하여 [표 10]과 같은 전체표본에 대한 비교를 실시하였으나, 실질적인 의료환경에서는 진단 알고리즘을 개발하는데 사용되었던 표본과 진단 알고리즘이 적용되는 표본은 분명히 다를 것이다. 그러므로 실질적인 체질진단 정확도는 학습-진단 표본을 통해서 검증되어야 한다. 체질량지수를 포함시킨 경우에는 모든 방법에서 전체표본과 학습-검증 표본의 정확도 차이가 미소하다. 그러나 체질량지수를 제외시킨 경우에는 회귀분석을 제외한 두 방법에서는 전체표본과 학습-검증 표본사이의 차이가 심하게 나타난다. 이는 분석표본의 변동이 결과의 정확도에 큰 영향을 미치는 것으로서, 진단 정확도의 신뢰성이 부족하므로 의료환경에 적용되어서는 안된다. 로지스틱 회귀분석의 경우 표본변동이 정확도에 미치는 영향은 적으나, 체질량지수를 고려하면 정확도가 향상되었다. 분석방법이 제공하는 결과를 살펴보면, 로지스틱 회귀분석과 의사결정나무는 체질분류에 사용되는 입력변수가 체질결정에 미치는 영향도나 결정조건에 관한 정보를 제공하고 있다. 그러나 신경망은 단지 결과만을 제공할 뿐, 의사결정의 과정을 추측할 수 있는 어떠한 정보도 보여주지 않는다. 진단 알고리즘은 임상경험의 축적을 통하여 개선되어 나가야 하며, 진단결과 뿐만 아니라 진단과정도 매우 유용한 정보이다. 이러한 관점에서 신경망은 치명적인 단점을 안고 있다. 그러므로 본 연구를 통하여 가장 추천되는 체질분류 방법은 ① 체질량지수를 고려한 의사결정나무 방법 또는 ② 체질량지수를 고려한 로지스틱 회귀분석 방법이다.

V. 결 론

본 연구에서는 정확도가 높으면서도 의료환경에 적합한 사상체질 분류방법을 탐색하였다. 최종적으로 1635명의 자료가 분석에 사용되었다. 전문의가 확진한 사상체질, 건강검진에서 측정한 혈압과 맥파자료를 판별분석, 회귀분석, 의사결정나무, 그리고 신경망에 적용시켜 , 각 방법의 체질진단 정확도를 비교하였다. 체질분류에서 가장 기여도가 높은 체질량지수와 분석표본의 변동이 정확도에 미치는 영향을 조사하였으며, 다음과 같은 결론을 얻을 수 있었다

체질분류를 위하여 입력변수를 선정하는 경우에 있어서, 체질량지수는 체질분류의 정확도를 향상시켰다. 또한 체질량지수는 분석표본의 변동에 따른 체질분류 정확도의 변동을 최소화하였다.

판별분석은 체질과 입력변수들 사이의 관계를 명확하게 파악할 수 있는 장점이 있는 반면, 체질별 등분산가정을 만족하기 어려워 체질별 표본의 크기가 체질분류의 정확도에 영향을 미쳤다. 의사결정나무와 신경망은 분석방법을 적용하는데 있어서 제한성은 없었지만, 체질량지수를 고려하지 않는 경우 분석표본의 변동이 정확도에 미치는 영향이 매우 컸다. 신경망분석의 경우는 생성된 모형이 블랙박스로 되어 있어 사용자가 내부의 알고리즘을 알 수 없으므로, 경험의 축적을 통한 점진적인 알고리즘의 개선을 기대할 수 없다. 의사결정나무는 각각의 입력변수가 체질분류의 과정에 작용하는 규칙을 제공하므로 현실적으로 유용한 도구라고 생각된다. 그러나 체질량지수를 반드시 고려하여, 분석표본의 변동에 따른 체질진단 정확도의 악화를 방지해야 한다. 로지스틱 회귀분석 방법은 방법의 적용에 제한성이 없었으

며, 분석표본의 변동이 체질진단 정확도에 미치는 영향도 거의 없었다. 그러나 회귀분석 방법을 사용하기 위해서는 체질의 분류에 결정적인 영향을 미치는 체질량지수를 고려해야만 분류의 정확도를 높일 수 있다.

참 고 문 헌

[1] 허만회, 고병희, 송일병, "체간 측정법에 의한 체질판별", 사상체질의학회, 제14권, 제1호, pp.51-66, 2002.

[2] 석재화, 윤종현, 이준희, 황민우, 조용진, 고병희, 이의주, 송일병, "사상체질진단 두면부 분석프로그램의 Upgrade 연구: 성별, 연령별 특징", 사상체질의학회, 제19권, 제3호, pp.30-50, 2007.

[3] 허재범, 정운기, 최민기, 유준상, 전종원, 김달래, "사상체질음성분석기를 통한 한국인 소아 청소년의 체질별 음성특성 연구 -단문을 중심으로", 사상체질의학회지, 제19권, 제2호, pp.40-52, 2007.

[4] 김선호, 고병희, 송일병, "사상체질분류검사지(QSCC II)의 표준화 연구", 사상체질의학회지, 제8권, 제1호, pp.187-246, 1996.

[5] 나경찬, "희수식 맥진기를 이용한 사상체질감별법", 대한한의학회지, 제14권, 제2호, pp.139-153, 1993.

[6] 박승창, 김대진, "사상 체질 판별 알고리즘과 자동맥진 시스템의 구현", 전자공학회논문지, 제41권, 제2호, pp.53-60, 2004.

[7] 이시우, 주종천, 김경요, 김종열, "어레이 압저항 센서를 활용한 체질맥 임상연구", 사상체질의학회지, 제18권, 제1호, pp.118-131, 2006.

[8] 김종원, 이의주, 김규곤, "성별 나이 비만도를 고려한 의사용 사상체질판별함수의 진단정확률 비교", Journal of the Korean Data Analysis Society, 제9권, 제3호, pp.1077-1088, 2007.

[9] 신상훈, 김종열, "맥상파를 이용한 체질 판별방법에 관한 연구", 동의생리병리학회지, 제22권, 제6호, pp.1403-1409, 2008.

[10] 신상훈, 임혜원, 박영재, 박영배, "심혈관 노화가 맥상에 미치는 영향", 대한한의진단학회, 제9권, 제1호, pp.59-68, 2005.

[11] 양병화, *다변량데이터 분석법의 이해*, 커뮤니케이션북스, 2006.

[12] W. Zhu, X. Huang, J. He, M. Li, and H. Neubauer, "Arterial intima-media thickening and endothelial dysfunction in obese chinese children," European J. of Pediatrics, Vol.164, No.6, pp.337-344, 2005.

[13] SPSS Korea, *Clementine Manual*, SPSS, 2007.

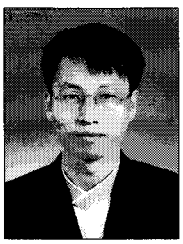
저 자 소 개

신 상 훈(Sang-Hoon Shin) 정회원

- 1987년 2월 : 부산대학교 기계설계학과(공학사)
- 1989년 2월 : 부산대학교 기계공학과(공학석사)
- 1995년 2월 : 부산대학교 기계공학과(공학박사)
- 2006년 2월 : 경희대학교 한의학과(한의학박사)
- 1995년 3월 ~ 1998년 2월 : LG산전 선임연구원
- 1998년 3월 ~ 2006년 2월 : 삼성종합기술원 수석연구원
- 2006년 2월 ~ 현재 : 상지대학교 한방의료공학과 교수

<관심분야> : 한방의료시스템, 생체역학

김 종 열(Jong-Yeol Kim) 정회원



- 1983년 2월 : 서울대학교 건축학과(공학사)
- 1985년 2월 : KAIST 토목공학과(공학석사)
- 1996년 2월 : 경희대학교 한의학과(한의학박사)
- 1998년 2월 : 원광대학교 한의학과(한의학석사)
- 2001년 2월 : 원광대학교 한의학과(한의학박사)
- 1996년 ~ 2004년 : 재단법인 익산원광한의원 원장
- 2004년 ~ 현재 : 한국한의학연구원 책임연구원

<관심분야> : 한방의료기기, 데이터 마이닝, u-헬스 의료기기