신경망을 사용한 사상체질 진단검사 개발 연구

채 한* · 황상문 · 엄일규[1] · 김병철[2] · 김영인[2] · 김병주 · 권영규

부산대학교 한의학전문대학원 양생기능의학부, 1 : 부산대학교 전자전기공학부,
2 : 부산대학교 생명자원과학대학 바이오메디컬공학과

Development of Sasang Type Diagnostic Test with Neural Network

Han Chae*, Sang Moon Hwang, Il Kyu Eom[1], Byoung Chul Kim[2], Young In Kim[2],
Byung Joo Kim, Young Kyu Kwon

*Division of Longevity and Biofunctional Medicine, School of Korean Medicine, 1 : School of Electrical Engineering,
2 : Department of Biomedical Engineering, College of Natural Resource and Life Science, Pusan National University*

The medical informatics for clustering Sasang types with collected clinical data is important for the personalized medicine, but it has not been thoroughly studied yet. The purpose of this study was to examine the usefulness of neural network data mining algorithm for traditional Korean medicine. We used Kohonen neural network, the Self-Organizing Map (SOM), for the analysis of biomedical information following data pre-processing and calculated the validity index as percentage correctly predicted and type-specific sensitivity. We can extract 12 data fields from 30 after data pre-processing with correlation analysis and latent functional relationship analysis. The profile of Myers-Briggs Type Inidcator and Bio-Impedance Analysis data which are clustered with SOM was similar to that of original measurements. The percentage correctly predicted was 56%, and sensitivity for So-Yang, Tae-Eum and So-Eum type were 56%, 48%, and 61%, respectively. This study showed that the neural network algorithm for clustering Sasang types based on clinical data is useful for the sasang type diagnostic test itself. We discussed the importance of data pre-processing and clustering algorithm for the validity of medical devices in traditional Korean medicine.

Key words : neural network, sasang type diagnostic test, data pre-processing, clustering algorithm, medical informatics

서 론

이제마의 사상의학은 인간을 태양인, 소양인, 태음인, 소음인으로 분류하여, 각 유형이 지니고 있는 생리, 병리적 특성에 따라 질병을 예방, 진단, 치료하는 한국적 맞춤의학이다. 이에 각 체질을 진단한다는 것은 임상적으로 필수적인 과정으로, 진료과정에서 추출된 임상 생체 데이터를 활용하여 사상체질 진단검사(Sasang type diagnostic test, STDT)를 개발한다는 것은 임상적 측면에서 매우 중요한 의미를 지니게 된다[1-3].

이에 체질별로 고유한 임상 데이터들을 용모사기(容貌詞氣), 체형기상(體形氣像), 항심심욕(恒心心慾), 성질재간(性質才幹), 병증약리(病證藥理), 유전학 등의 측면에서 추출함으로써 다양한 사상체질 진단검사를 개발하고자 하는 노력이 다양한 측면에서 꾸준히 진행되어 왔으나, 이들 데이터를 활용하여 진단검사를 만들어내는 과정에 있어서 중요한 구성 요소가 되는 데이터의 획득 방법[4,5], 데이터의 전처리 방법[2,6] 등과 같은 세부 방법론적 요소들에 대한 체계적 연구들은 충분히 이루어지지 못하여 왔다.

기존의 사상의학 연구들은 각 체질 유형별로 유의한 차이를 보이는 단일 지점에서의 데이터 추출에만 집중함으로 인하여, 추출된 데이터들의 수학적 특징이나 추출된 데이터들을 모아서 분석하는 방법 등에 대한 연구들은 충분하지 못하여 왔다. 통계적 관점에서 볼 때, 기존 연구들은 특정 임상데이터 필드를 대상으로 '체질 유형 간 유의한 차이가 존재하느냐'라는 가설 검증(hypothesis test) 혹은 유의성 검증(significance test)에만 집중함으로써, 단일 필드에서의 차이가 유의한지에 대한 믿음의 크기를 측정[7]하는데 주력하였기 때문에, 결과적으로는 T-test나 ANOVA 등을 활용한 차이의 탐색(exploration)과 확인(confirm)에만 치중하는 결과를 보여 왔다[3,8,9].

* 교신저자 : 채 한, 경남 양산시 물금읍 부산대학교 한의학전문대학원
· E-mail : han@chaelab.org, · Tel : 051-510-8470
· 접수 : 2009/05/25 · 수정 : 2009/06/26 · 채택 : 2009/07/09

이에 사상의학 진단과 관련한 많은 연구들을 통해서 확보된 체질별 고유 임상 생체 데이터가 상당한 양에 이르고 있음에도 불구하고, 이러한 결과들을 구체적으로 종합, 가공, 분석하여 임상 검사에 활용하기위한 체계적 분석 방법론에 대한 연구는 부족하여 왔으며, 이는 현실적으로 시급한 문제가 되고 있다[8].

사상체질 분류검사를 개발함에 있어서는 특정 데이터 필드의 통계적 유의성만큼이나 데이터들을 통합하여 활용하는 데이터 분석 또한 중요하다. 그러나 기존의 연구에 있어서는 SPSS 등 상용 통계 패키지에 포함되어 있다는 피상적 이유와 함께 몇 번의 클릭만으로도 손쉽게 결과를 얻을 수 있다는 편의성으로 인하여, 대부분의 사상체질 진단검사의 개발에 있어 별다른 고민없이 판별분석(discriminant analysis)이 활용되어온 것이 사실이다. 본 연구에서는 데이터 분석도구의 기존 범주를 확장하여 생체데이터의 통계적 분석과 분석결과의 한의공학적 활용을 동시에 고려하는 새로운 연구 방법론이 필요하다는 인식[8]하에 Self-organizing feature map (SOM)[10]의 사상의학적 활용 가능성에 대하여 검토하였다.

SOM은 1982년 Teuvo Kohonen에 의해서 개발된 신경망의 하나로서, 주어진 데이터에 존재하는 상관도와 고유한 규칙성을 학습하여 복잡한 데이터들을 몇 개의 클러스터들로 조직화할 수 있는 도구로 알려져 있다(Fig. 1A).

SOM은 비교사 훈련방법 (unsupervised learning)인 경쟁적 학습방법(competitive learning)을 이용하여 데이터들을 자기조직화(self-organization)하는 능력을 가지고 있다. 이는 인간이 가진 학습능력 가운데 교사나 조언자 없이도 스스로의 판단에 의해 관찰한 데이터를 분류하고 이러한 학습 결과나 경험을 바탕으로 새로운 데이터가 주어졌을 때 그것이 속한 범주를 추정할 수 있는 능력을 모방하는 것으로, SVM 등이 임상적 경계점에 대한 인식이 사전에 선행된다는 점에서 차이를 보이고 있다[11-13]. 또한 '자기 조직화'와 '경쟁적 학습'을 특징으로 하는 SOM 신경망은 요소 간 구조나 통계적 특징이 전혀 알려져 있지 않은 데이터를 분석하는 데 유용한 도구이며, 다른 신경망에 비해 비교적 간단한 구조와 학습규칙을 사용하는데, 생물학적 학습 메커니즘보다는 인간의 사고 능력에 대한 기능적 측면을 모사하도록 학습규칙을 모델링하는 특징을 지니고 있다.

본 연구에서는 이와 같은 신경망을 활용한 사상체질 분류검사(STDT)의 전형적인 모델을 제시해보고, 입력된 데이터의 프로파일을 어떻게 재현해내는가에 대한 고찰을 통해 한의학 임상에서의 활용 가능성을 타당도 분석을 통해 검토하였다.

연구방법

1. 연구 데이터

신경망의 분류 능력을 평가함에 채 등[14]의 연구에서 사용하였던 기존 데이터를 사용하였다. 연구 참여자는 19세에서 43세 사이의 한의과대학 재학생 102명(남자 89명, 여자 13명)으로서, Questionnaire for the Sasang Constitution Classification II (QSCCII) 검사, Myers-Briggs Type Indicator (MBTI) 검사, Bio-Impedance Analysis (BIA) 검사를 수행하였다. 참가자 중 12명이 QSCCII설문을, 3명이 MBTI 검사를 12명이 BIA 검사를 완료하지 않았다. 이에 23명을 제외한 79명(남자 69명, 여자 10명)의 검사 결과를 활용하였다.

QSCCII는 사상의학에 기반을 둔 설문 형태의 체질 판별법으로서 경희대학교에서 1993년에 개발되고 1996년에 보완된 121문항의 설문지이다. 체질 감별 정확도(PCP)는 70%라고 보고되었으며[15], 내적일치도(Cronbachs a)는 태양인 0.57, 소양인 0.57, 태음인 0.57, 소음인 0.63이라고 보고되었다[14].

MBTI 검사는 95개의 문항을 사용하여 개인의 성격적 특징을 검사하는 도구로서, 성격을 EI, SN, TF, JP의 네 영역 즉, 외향(Extraversion)/내향(Introversion), 감각(Sensing)/직관(Intuition), 사고(Thinking)/감정(Feeling), 판단(Judging)/인지(Perceiving) 영역에 있어서 선호하는 정도를 측정한다. 본 연구에서는 영역별 선호 유형과 선호 정도를 100점을 기준으로 표준화하여 사용하였다. 예를 들면, 외향/내향에 있어서 100보다 낮은 점수는 외향적이며, 높은 경우에는 내향성이 높은 것을 의미한다.

BIA는 임상연구를 목적으로 인체의 체성분 분포를 측정하기 위한 전자적인 분석 장치이다. 간단하면서 비침습적인 방법을 사용하여 손과 발에 연결된 전극으로부터 임피던스의 변화를 관찰하는 것만으로도 체수분, 단백질, 무기질, 체지방 등의 분포에 대한 신뢰할만한 측정치를 추출할 수 있는 방법론적 장점을 지니고 있다[1].

이에 사용된 30개의 데이터 필드의 명칭과 의미는 다음과 같은데, 성별(Sex), 연령(Age), MBTI EI(M_EI), MBTI SN(M_SN), MBTI TF(M_TF), MBTI JP(M_JP), Intracellular water(ICF), Extracellular water(ECF), Protein mass(PM), Amino mass(AMM), Body fat mass(BFM), 키(Height), 몸무게(Weight), Percent body fat(PBF), Waist-hip ratio(WHR), Fluid of right arm(FRA), Fluid of left arm(FLA), Fluid of trunk(FT), Fluid of right leg(FRL), Fluid of left leg(FLL), 신체발달(Development of body), Total body water(TBW), Body mass index(BMI), arm leg ratio(ALR), Proper body weight(PBW), average body weight(ABW), average body fat(ABF), average muscle weight(AMW), right and left ratio of the upper limb(RLU), right and left ratio of the lower limb(RLL)이다.

2. 분석 방법

1) 특징 추출을 통한 잉여 데이터 필드의 제거

MBTI와 BIA 등을 이용하여 획득한 데이터는 나이와 성별을 포함하여 30개의 필드와 79개의 레코드를 가진다. 따라서 SOM 신경망의 입력층 노드의 개수는 30개가 되어야 하나, 적절한 특징추출 방법을 사용하여 데이터에 존재하는 잉여 정보(redundancy)를 제거함으로써 입력 차원을 줄일 수 있다.

일반적으로 실험에 사용되는 데이터 간에는 $d_r = f(d_1, d_2, \cdots)$와 같은 함수 관계가 존재할 수 있다. SOM 신경망은 앞서 언급한 바와 같이 복잡한 입력 정보들 간의 상호관계를 스스로 조직화하는 능력을 가지고 있으므로, 이러한 상호 함수 관계를 가진

데이터들은 관련 정보가 중복되는 형태로 파악될 수 있다. 위의 경우 데이터 d_r은 d_1, d_2,…이 가지고 있는 정보량 이상을 지니지 못하므로, 데이터 필드 상호간의 관계를 분석함으로써 잉여 정보인 d_r를 제거하였다.

2) 상관도 분석을 통한 잉여 데이터 필드의 제거

특징 추출을 통한 데이터 제거 이후에 남은 필드들을 대상으로 교차 상관도 (cross-correlation)를 구하고, 이들 중에서 상관도가 높은 항목 또한 잉여 정보(redundancy)라고 할 수 있으므로 상관도 분석을 통해 제거하였다. 데이터 d_m과 d_n(m=n=1, 2, …, N)을 랜덤 변수로 보았을 경우 교차 상관도는 다음과 같이 표현된다. 즉,

$$\rho_{mn} = \frac{E[(d_m - M_m)(d_n - M_n)]}{STD(d_m)STD(d_n)}$$

여기서 E[·]은 기대값을, STD[·]는 표준편차를 나타내고, M_m은 필드 d_m의 평균값이다. 이 때, 교차 상관행렬을 나타내면,

$$C = \begin{pmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1N} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{N1} & \rho_{N2} & \cdots & \rho_{NN} \end{pmatrix}$$

여기서 $p_{mn} = p_{nn}$ (대칭행렬)이고, p_{mm}=1이므로 행렬 C는 대각원소가 모두 1인 대칭행렬이 되며, $-1 \le p_{mn} \le 1$의 범위를 가진다. $|p_{mn}|$이 클수록 d_m과 d_n은 높은 상관도를 가지게 되는데, 상관도가 높다는 것은 d_m과 d_n의 선형 의존성이 높음을 의미한다. 따라서 교차 상관도를 구함으로써 상관도가 높은 데이터를 제거할 수 있다.

3) 신경망을 활용한 사상 체질의 분류

SOM은 입력층 (input layer)과 경쟁층 (competitive layer)의 두 개의 레이어를 가지는 형태로 구성된다(Fig. 1B). 입력층은 입력 패턴을 경쟁층에 있는 각 노드들로 분배하게 되며, 평면 형태로 배열된 노드들의 집합인 경쟁층은 입력층으로부터 전달된 신호와 학습 규칙에 따라 발화 상태가 결정된다. Fig. 1은 이러한 SOM 신경망의 구조를 보여주는데, $p = [p_1, p_2, \cdots, p_p]$ 는 입력 벡터이며, 입력 층의 i 노드와 경쟁 층의 j 노드를 연결하는 연결강도는 w_{ij}(j = 1, 2, …, MN)로 표현된다. 즉,

$$w_j = [w_{j1}, w_{j2}, \cdots, w_{jp}]$$

SOM 신경망은 대뇌 피질을 2차원 구조로 단순화한 형태를 가지게 되나, 생물학적 뉴런들의 학습 매커니즘은 아직 명확히 규명되어 있지 않다. 따라서 SOM 신경망은 생물학적 학습 메커니즘보다는 인간의 특정 능력에 대한 기능적 측면을 모사하도록 학습 규칙을 모델링하는데, SOM 신경망의 학습 규칙은 다음과 같다.

$$w_v(t+1) = w_v(t) + \alpha(t)(p - w_v(t))$$

여기에 있어서 $W_v(t)$ 는 t시점에서 연결강도 벡터들 중에서 승자 노드에 해당하는 연결강도 벡터를 의미하며, α(t)는 연결강도를 갱신할 경우 그 크기를 조절하는 학습 상수를 의미하며, 승자 노드 v는 다음과 같이 결정된다. 즉,

$$v = \underset{j=1,\cdots,MN}{\arg\min} \; p - w_j$$

여기에서 $\| p\text{-}w_j \|$는 입력벡터 p와 연결강도 벡터 w_j의 유클리드 거리 (Euclidean distance)를 의미한다. 식 (4)과 (5)에서 볼 수 있듯이 학습 과정은 입력 층에서 제시된 p와 가장 가까운 연결강도 벡터를 가지는 노드에 대해 p와 보다 더 가까운 거리를 가지는 방향으로 해당 연결강도를 갱신하게 된다. 이와 같은 경쟁학습 규칙은 승자 노드만의 학습을 허락하기 때문에 종종 승자독점 (winner-take-all) 법칙이라고 불린다. SOM 신경망은 위와 같은 승자 노드 결정 방식에 더해서, 층 내에서 경쟁하되 승자 노드와의 기하학적 거리에 반비례하도록 학습기회를 부여하는 '측면제어(lateral control)'를 이용하기도 한다. 일반적으로 이웃반경의 크기는 인접한 몇몇 뉴런으로 제한하며 이웃반경의 크기는 Fig. 1C 와 같이 학습이 진행됨에 따라 줄어들게 된다. SOM 신경망의 경쟁학습과 측면제어 효과는 학습이 진행됨에 따라 연결강도를 자기 조직화한다. 학습의 진행에 따른 연결강도의 변화를 연결강도 공간에서 좌표로 나타낼 경우, 이러한 자기조직화 과정을 시각화하여 볼 수 있는 데 따라서 이것을 자기조직화 지도(Self-Organizing Map, SOM)라고 부른다[16]. 이상의 SOM의 학습 과정은 다음과 같은 다섯 단계로 정리될 수 있다.

단계 1] 연결강도의 초기화. 보통 임의의 작은 값으로 초기화 한다(난수발생).

단계 2] 입력층에 입력벡터를 제시한다.

단계 3] (2)에 의해 입력벡터와 모든 경쟁 층 노드와의 거리를 계산하여 승자노드를 결정한다.

단계 4] 승자 노드와 이웃 반경에 포함되는 노드들의 연결강도 벡터를 식 5에 의해 갱신한다. 이때 학습 상수는 0과 1사이의 값을 가지며 시간이 경과될수록 작아지도록 설계해야 한다. 이웃 노드 갱신 시에는 측면효과를 고려해야 한다.

단계 5] 연결강도의 변화가 별로 없거나 입력 벡터들과 연결강도간의 거리가 문턱치 이하이면 학습을 멈추고 그렇지 않으면 단계 2로 진행한다.

SOM 신경망은 주어진 데이터에 존재하는 상관관계를 학습하여 클러스터링을 수행하는데, 이때 각 노드의 연결강도 벡터는 각 클러스터를 대표하는 통계상의 표준 벡터(prototype)가 된다. 학습이 종료되면, 입력층에 임의의 입력이 주어질 경우 입력벡터와 가장 가까운 표준 벡터를 연결강도로 가지는 경쟁층의 노드가 발화하게 된다.

4) 결과의 타당도 분석

본 연구에서는 피 실험자들로 부터 수집하였던 MBTI 점수와 BIA 데이터 등을 대상으로 SOM 신경망을 활용한 클러스터링을 수행하여 데이터 내에 존재하는 통계적 특성과 분류 결과의 특징을 살펴보고자 하였다. 그러나 클러스터링에 있어서는 SOM의 특성이 더 명확히 발현될 수 있도록 하기 위하여, 기존 연구에서 확인되었던 태음인의 높은 BIA와 소음인의 높은 MBTI 외향성, 소양인의 높은 MBTI 내향성이라는 임상적 문맥(context)[14]은 본 연구에서 고려되지 않았다. 신경망을 활용한 사상체질 진단 결과와 실측 사상 체질과의 일치도를 확인하여 타당도를 분석함에 있어서는 유형(체질)이 옳게 분류된 총 도수를 사용된 모든 빈도수로 나눈 정확 예측율(PCP)과 유형(체질)별로

옳게 분류된 도수의 비율인 체질별 민감도(type specific sensitivity)를 사용하였다[2].

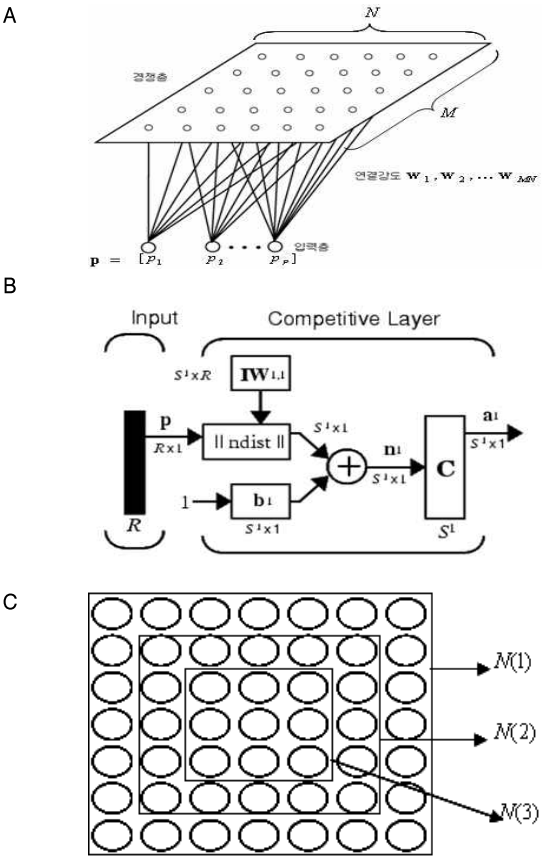


Fig. 1. Schematic explanation of Self-Organizing Map (SOM). (A) structure of Self Organizing Map, (B) network diagram of SOM, (C) Change of outer range during the SOM learning

결 과

1. 특징 추출을 통한 잉여 데이터 필드의 제거

30개 데이터 필드간의 상호 관계를 분석하여, 잉여 정보들을 제거하였다. Table 1에서 보는 바와 같이 9개의 데이터 필드는 기존 필드간의 함수로서 규정될 수 있었기에 본 연구에서는 이들을 제거하였다.

Table 1. Data elimination by analyzing latent relationship

Eliminated data field	Latent relationship between data fields
Total Body Water (TBW)	TBW = ICF +ECF
Body Mass Index	BMI = weight / (height * height)
ALR	ALR = (FRA + FLA) / (FRL +FLL)
Adequate body weight(ABW)	ABW = f(height)
Body weight control (BWC)	BWC = adequate weight - weight
Right/Left arm	FRA/FLA = ratio of right/left upper limb
Right/Left leg	FRL/FLL = ratio of right/left lower limb
Body fat control Body muscle control	= f(BFM, height and weight)

2. 상관도 분석을 통한 잉여 데이터 필드의 제거

상관도 분석에 있어서 $|p_{mn}| > 0.9$인 경우를 높은 상관성을 지니고 있다고 보고, 이에 해당되는 데이터 필드를 제거하였다. 앞서 특징 추출을 통해 제거된 9개의 필드를 제외한 21개의 항목에 대하여 교차 상관도를 구하였으며, 0.9 이상의 높은 상관도를 가지는 9개의 데이터 필드는 Table 2와 같이 '●'로 표시하였다. Table 2의 7행(ICF)과의 상관성이 높은 9(PM), 10(AMM), 16(FRA), 17(FLA), 18(FT), 19(FRL), 20(FLL)을 제거할 수 있었으며, 11행(BFM)과 높은 상관성을 지닌 15(WHR) 또한 제거될 수 있었다. 아울러 연령(2) 또한 체질 분류에 영향을 주지 않을 것이라는 가정 하에 제거하였다. 최종적으로 두 단계에 걸쳐 18개의 데이터 필드를 제거하고 남은 12개의 데이터 필드를 SOM 신경망에 입력하였는데, 12개의 입력된 데이터 필드는 sex, M_EI, M_SN, M_TF, M_JP, ICF, ECF, BFM, height, weight, PBF, 신체발달이었다. 실험 결과 12개를 사용하였을 때와 30개 모두를 입력 데이터로 사용한 결과가 동일하게 나왔는데, 이는 실험에서 사용한 특징 추출법의 타당성을 확인하는 것으로 사료된다.

Table 2. Correlation matrix between data fields

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1																					
2																					
3																					
4																					
5																					
6																					
7									●	●						●	●	●	●	●	
8																					
9							●			●						●	●	●	●	●	
10							●		●							●	●	●	●	●	
11															●						
12																			●	●	
13																					
14																					
15											●										
16							●		●	●							●	●	●	●	
17							●		●	●						●		●	●	●	
18							●		●	●						●	●		●	●	
19							●		●	●		●				●	●	●		●	
20							●		●	●		●				●	●	●	●		
21																					

* 1=Sex, 2=Age, 3=M_EI, 4=M_SN, 5=M_TF, 6=M_JP, 7=ICF, 8=ECF, 9=PM, 10=AMM, 11=BFM, 12=Height, 13=Weight, 14=PBF, 15=WHR, 16=FRA, 17=FLA, 18=FT, 19=FRL, 20=FLL, 21=Development of body. ● means higher than 0.9.

3. 신경망을 활용한 사상 체질의 분류

신경망을 활용하여 분류한 사상체질별 MBTI, BIA 특성들은 기존 연구 결과와 유사한 패턴을 보이고 있음을 확인할 수 있었다. 신경망을 활용하여 사상체질별 MBTI를 분석함에 있어서, 신경망을 활용한 결과(Fig. 2B)가 각 체질별로 실측된 결과(Fig. 2A)와 매우 흡사한 패턴을 보였다. 소양인은 EI(93.8±18.2), SN(90.3±20.3), TF(80.4±19.6), JP(97.4±17.6)로 나타났으며, 태음인은 EI(122.5±20.4), SN(110.3±17.3), TF(106.0±19.4), JP(127.1±15.9), 소음인은 EI(137.4±10.8), SN(86.6±20.8), TF(88.0±22.6), JP(74.8±15.7)로 나타났다. 각 체질별 MBTI 프로파일을 실측치와 신경망 분석 결과에 있어서 비교해 본다면, 소양인과 소음인은 기존 실측치와 신경망 분석 결과가 매우 유사하게 나타났으나, 태음인은 이와 달리 기존의 프로파일보다 유의하게 높은 점수를

보이고 있었다. 실측치와 신경망 분석의 두 가지 경우에서의 각 체질 간 차이를 ANOVA (SPSS 14.0, SPSS Inc., Chicago)를 사용하여 분석한 결과, 기존 데이터에 있어서는 EI(F=16.700, p<0.001), SN(F=1.056, p=0.353), TF(F=0.606, p=0.548), JP(F=4.559, p=0.014)로 나타났으나, 신경망 분석에 있어서는 EI(F=42.663, p<0.001), SN(F=11.03, p<0.001), TF(F=10.127, p<0.001), JP(F=67.823, p<0.001)로 나타났다.

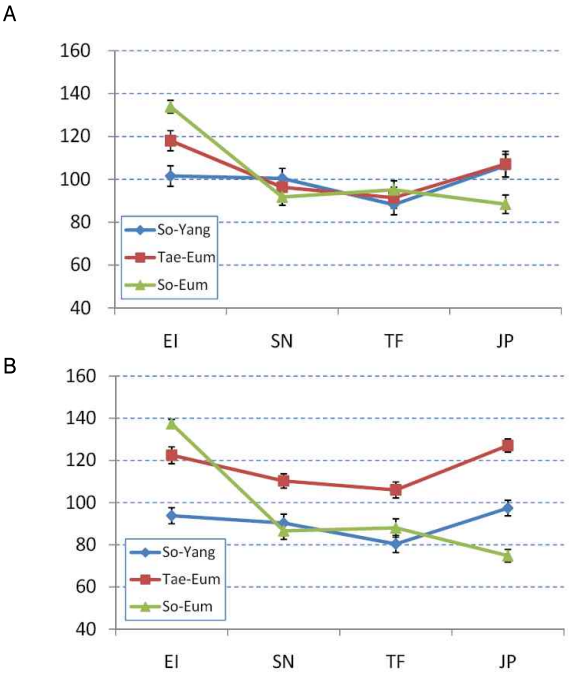
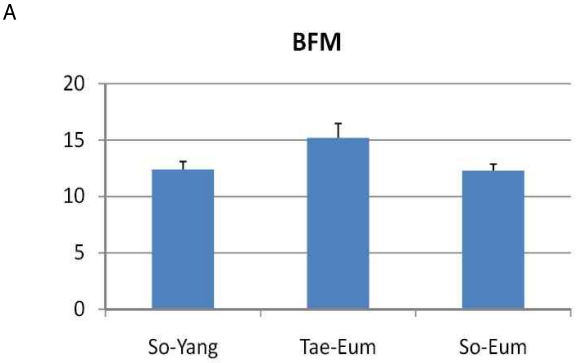


Fig. 2. Myers-Briggs Type Indicator profile of each Sasang types from (A) original data and (B) current study classified with SOM. Data shown as mean and standard error.

신경망을 활용하여 사상체질별 BIA를 분석함에 있어서, 신경망을 활용한 결과(Fig. 3B)가 소양인(12.5±3.9), 태음인(14.3±5.6), 소음인(12.7±3.1)로 나타났는데, 이는 각 체질별로 실측된 결과(Fig. 3A)인 소양인(12.4±3.4), 태음인(15.2±6.1), 소음인(12.3±3.2)과 매우 흡사함을 알 수 있었다. 다만 두 가지 경우에서의 각 체질 간 차이를 ANOVA를 사용하여 분석한 결과에 있어서는, 기존 데이터에 있어서는 F=3.61, p=0.032로 나타났으나, 신경망 분석에 있어서는 F=1.277, p=0.285로 나타났다.



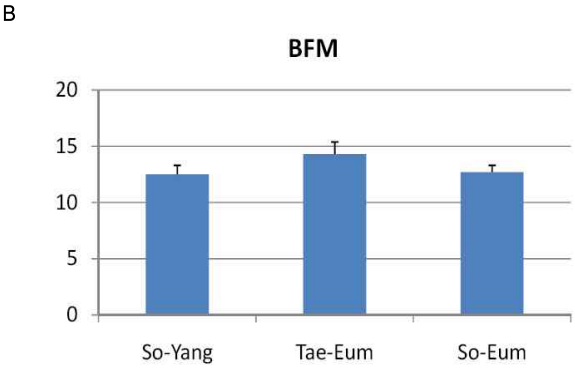


Fig. 3. Body fat mass profile of each Sasang types from (A) original data and (B) current study classified with SOM. Data shown as mean and standard error.

실측치에 대한 신경망 분석 결과의 비교에 있어서는 정확예측율(PCP)이 56%로 나타났는데, 체질별 민감도에 있어서는 소양인 민감도는 56%, 태음인 민감도는 48%, 소음인 민감도는 61%이었다.

고찰 및 결론

본 연구에서는 사상체질 진단 알고리듬으로 사용하여 왔던 기존의 판별분석(discriminant analysis)보다 유용할 것으로 사료되는 한의공학적 알고리듬으로서 Self-organizing feature map (SOM)을 제시하고, SOM의 유용성을 확인하기 위하여 임상 생체 데이터를 활용한 체질 분류를 시행한 결과를 기존의 연구 결과와 비교분석하였다. 이러한 과정에서 SOM을 활용한 사상체질 진단검사가 56%의 정확예측율과 함께 소양인 민감도 56%, 태음인 민감도 48%, 소음인 민감도 61%를 지니고 있음을 알 수 있었으며, MBTI와 PBF에 있어서 기존 연구에서의 실측치와 유사한 프로파일을 보이고 있음(Fig. 2, 3)도 확인할 수 있었다. 이는 본 연구에 활용된 SOM이 사상체질 분류 알고리듬으로서 유용성을 지니고 있음을 반증하는 것이라 사료된다.

데이터의 전처리 과정을 통해 데이터 필드간의 함수 관계 특징 (Table 1) 및 상관도 분석(Table 2)을 통해 잉여 정보를 제거한 결과, 처음의 30개 데이터 필드가 12개의 데이터 필드로 축소될 수 있었으며, 이들을 활용한 체질 진단 결과 또한 제거 이전과 이후가 동일하게 나타나는 것을 확인할 수 있었다. 이러한 분석 결과는, 기존 연구들에 있어서 사초(四焦)간의 형태학적 차이를 극대화하기 위하여 측정치(raw data)를 그대로 사용하기보다 각종 함수관계를 통해 새로운 데이터 필드를 생성하고자 하였던 시도들이 실제 분석에 있어서는 큰 도움이 되지 못한다는 기존의 연구결과[1]를 다시 재확인하는 것이라 사료된다.

아울러 본 연구에서 사용된 데이터 전처리 과정들은, 분석의 편의를 위해 데이터 필드의 개수를 줄임에 있어서 논리적이면서도 유용하게 활용될 수 있는 방법으로, 추후 한의학 임상 데이터를 분석하는 다른 연구에 있어서도 적절히 응용될 필요가 있을 것이다. 기존의 연구들에 있어서는 이러한 데이터 전처리 과정에 크론바하 알파[4]나 도수분석[5] 등이 사용되어야 한다고

제시하고 있다.

그러나 이와 같은 설문 응답의 도수분석을 통한 데이터 내부의 맥락(context)을 고려하거나 판별분석 과정에서의 중요도에 따라 변수를 임의적으로 제거하는 방법은, 연구 결과의 일반화를 지향하기 보다는 최종 결과에 있어서의 정확 예측율(PCP)을 높이기 위한 측정 데이터 자체의 인위적 왜곡 가능성을 배제할 수 없다는 측면에서 치명적 단점을 지니므로 본 연구에서 제시된 것과 같이 특징 추출과 상관도 분석을 활용한 데이터 필드의 제거가 사용되어야 할 것이다.

MBTI에 있어서 신경망을 활용한 결과는 실측치와 매우 흡사한 패턴을 보이고 있었으며(Fig. 2), 동시에 몇 가지 측면에 있어서는 기존 데이터와 다른 특성도 보이고 있었다. 소양인과 소음인의 MBTI 프로파일과 달리, 태음인의 SOM 분류 결과는 SN, TF, JP 값에 있어서 실측 데이터보다 높게 나타났음을 확인할 수 있었는데, 이는 신경망의 분류 학습 과정에 있어서 세 유형간의 차이가 과장되어 나타난 것으로 추정되며, 사상체질 유형을 3가지로 구분하는 것이 아니라 두 유형을 구분하도록 학습시켰다면 이와는 다른 양상을 확인할 수 있었을 것이라 사료된다.

BIA에 있어서는 신경망을 활용한 결과가 원 데이터와 거의 유사한 프로파일을 보이고 있음(Fig. 3)과 동시에 실측치와 SOM 분류 결과에 있어서 사상 체질 간 차이가 다르게 나타나는 것도 확인할 수 있었다. 체질별 차이를 통계적으로 확인하기 위하여 ANOVA를 실시하였을 때의 F값이 실측치의 3.61에서 SOM 분류치의 1.277로 줄어드는 것을 확인할 수 있었는데, 이러한 결과는 전반적인 경향성에 있어서는 그 프로파일이 그대로 재현되나 세부적인 수치들에 있어서는 상황에 따라 왜곡이 발생될 수 있음을 의미하는 것이라 사료된다. 이에 실질적인 사상체질 분류검사를 개발하는 과정에 있어서는 민감도와 특이도 같은 타당도 지표[2]를 중심으로 한 지속적 검토-수정-재검토의 과정이 필요할 것이다.

이상에 있어서 SOM라는 신경망 학습을 활용한 데이터 분석 과정을 살펴본다면, 임상가의 조언이나 방향 설정이 전혀 없는 상태에서도 스스로의 판단에 의해 관찰한 데이터를 자동으로 분류할 수 있다는 장점은 매우 높은 활용 가능성을 지니는 것으로 사료된다. 그러므로 만약 이러한 장점이 단일 데이터필드에 있어서 사상 체질 간 뚜렷한 차이를 확인하기 어려웠던 데이터 특성을 보였던 경우, 예를 들어 체질별 음성분석[17] 등에 활용되었다면 기존의 음성 특성의 분석에 판별분석(discriminant analysis)을 사용한 것[4,18]보다 더 높은 타당도를 얻을 수 있었을 것이라 보인다.

데이터 마이닝 알고리듬의 측면에 대해 고찰한다면, 한의학 임상 현장을 가장 잘 모사하는 것으로 주장되고 있는 의사결정나무(decision tree)[6,8,9]나 SVM(Support Vector Machine)[12] 등은 사전에 경계점에 대한 선행 인식을 요구한다. 만약 이러한 경계점을 결정하는 중요 변곡점에 주관적 가치평가 요소가 개입[6]되거나, 객관적 지표와 설문 응답의 결과가 불일치[6]할 경우에는 본연의 맥락에 심각한 왜곡이 초래될 가능성[8]이 내포되어 있다. 이에 사상체질진단검사(STDT)를 개발함에 있어서는 개개의 데이터 필드(문항)에 치중하기보다는 문항들의 화학적 총계로서 높은 안정성(stability)을 지니고 있는 요인(factor)을 추출하여 사용하는 것이 더 바람직할 것이며[19], 이들 요인 측정의 효율성을 높이는 연구가 중요하다 할 것이다.

본 연구에서 활용하였던 SOM은 데이터 내부에 존재하는 요소들 간의 구조나 통계적 특징들이 전혀 알려져 있지 않은 데이터를 분석함에 매우 유용하며, 여타 다른 신경망에 비하여 비교적 간단한 구조화 학습 규칙을 사용한다. 이에 사상체질 진단검사 뿐만 아니라, 차후 한의학적 변증(辨證) 시스템과 같은 한의공학적 연구개발에 있어서도 활용할 여지가 매우 높기 때문에 다양한 데이터마이닝 알고리즘에 대한 연구가 시급하다 할 것이다[20].

본 연구에서의 정확 예측율(PCP)은 56%로 그리 높은 편은 아니다. 그러나 앞에서도 기술한 바와 같이, 본 연구에서는 기존 연구에서 얻었던 임상적 문맥(clinical context)을 의도적으로 제외함으로써 알고리즘의 특성을 최대로 발현시키고자하는 의도를 지니고 있었기에, 그리 낮은 결과라고는 할 수 없을 것이다. 이론적 추론과 임상 연구 과정에서의 문맥에 대한 연구 결과가 추가된다면 그 효율은 더욱 높아질 것이다. 각 측정치별 중요성과 함께 기존 연구 결과를 반영한 분석 과정에서의 가중치, 예를 들어 심리특성[3]과 신체특성을 구별하여 처리하고, MBTI[3] EI와 JP, 그리고 BMI[8] 등에 있어서 기존 임상연구에서 확인된 임상적 문맥[14]들을 반영한다면 한층 높은 효율성(타당도)을 확보할 수 있을 것이다. 또한 양적(quantitative) 분석보다 질적(qualitative) 분석을 우선시하는 한의학적 임상 특징을 반영하기 위해서는, 임상 현장에서의 이산분포 데이터의 측정과 분석9)에 대한 연구가 시급히 진행되어야 할 것이며, 데이터 필드에 MBTI 점수나 BIA 측정치 등과 같이 표준화(standardized)된 수치가 사용되어야 할 것이다.

본 연구에서는 사상체질 진단검사(STDT)와 같은 한의공학에 활용될 높은 가능성을 지니고 있는 신경망분석의 하나인 SOM을 사용하여 사상체질별 임상 생체 데이터를 분석하여 보고, 실측 데이터와의 비교를 통해 그 특성을 살펴보았다. 이를 통해 상관도와 규칙성을 활용하는 신경망 학습이 임상적 문맥 없이도 사상체질 유형들을 클러스터링할 수 있음을 확인하였으며, 이를 토대로 사상체질 진단검사의 개발과정에서의 데이터 마이닝 알고리즘과 데이터 전처리의 중요성을 논하였다.

감사의 글

이 논문은 부산대학교 자유과제 학술연구비(2년)에 의하여 연구되었음.

참고문헌

1. 이수진, 박수현, 고유선, 박수진, 엄일규, 김병철, 김영인, 백진웅, 김명근, 권영규, 채한. 임피던스 분석을 활용한 사상인의 신체계측 연구. 동의생리병리학회지 23(2):433-437, 2009.
2. 이수진, 김명근, 채한. 사상체질 진단검사 타당성분석에 대한

연구. 대한한의학회지 29(1):7-14, 2008.
3. 채한, 박수잔, 이수진, 고광찬. 사상 유형학의 성격심리학적 고찰. 대한한의학회지 25(2):151-164, 2004.
4. 김규곤, 최승배. 한의학에서의 사상체질판별함수 개발에 관한 연구 (I) -크론박 알파 계수에 의한 변수선택- Proceedings of Joint Conference of Korean Data And Information Science Society and The Korean Data Analysis Society, April 30-May1, pp 61-68, 2004.
5. 김규곤, 조민형. 한의학에서의 사상체질판별함수 개발에 관한 연구(II) -도수분석에 의한 변수선택- Proceedings of Joint Conference of Korean Data And Information Science Society and The Korean Data Analysis Society, April 30-May1, pp 69-77, 2004.
6. 박성식, 최재영. 의사결정나무법을 이용한 설문지의 응답특성에 대한 임상적 검토. 사상체질의학회지 15(3):177-186, 2003.
7. Aviva, Petrie. Caroline Sabin. Medical Statistics at a Glance. 2nd Ed. Wiley-Blackwell. 2005.
8. 박은경, 이영섭. 박성식. 의사결정나무법을 이용한 체질진단에 관한 연구. 13(2):144-155, 2001.
9. 진희정, 문진석, 고성호, 구임회, 이시우, 이도현, 송미영, 김종열. 사상체질 의사결정시스템 구축을 위한 체질 진단 자료를 이용한 예비연구. 한국한의학연구원논문집 13(2):75-81, 2007.
10. Teuvo, Kohonen. Self-Organizing Maps. Springer-Verlag. 1997.
11. Savova, G.K., Ogren, P.V., Duffy, P.H., Buntrock, D.J., Chute, C.G. Mayo clinic NLP system for patient smoking status identification. Journal of the American Medical Informatics Association. 15(1):25-28, 2008.
12. Qian Zhang, Ki Jung Lee, Taeg Keun, Whang bo. K-mean and double cross-validation algorithm for LS-SVM in Sasang typology classification. Proceedings of the IEEE International Conference on Automation and Logistics. August 18-21, Jinan, China. pp 426-430, 2007.
13. Petrova, N.V., Wu, C.H. Prediction of catalytic residues using Support Vector Machine with selected protein sequence and sturctural properties. BMC Bioinformatics 7: 312, 2006.
14. Chae, H., et al. An alternative way to individualized medicine: psychological and physical traits of Sasang typology. Journal of Alternative and Complementary Medicine. 9(4):519-528, 2003.
15. 김선호, 고병희, 송일병. 사상체질분류검사지(QSCCⅡ)의 표준화 연구. 사상의학회지 7(1):187-246, 1995.
16. 이용구, 이윤수. 데이터마이닝에서 코호넨네트워크와 전통적 군집분석 방법의 비교 연구. 수학·통계논문집, 7: 17-29, 2000.
17. 조동욱. 음성 신호 분석에 의한 사상 체질 분류. 한국통신학회논문지, 31(5C):548-555, 2006.
18. 이의주, 송광빈, 최환수, 유정희, 곽창규, 손은혜, 고병희. 음성분석에 의한 체질진단에 관한 연구. 대한한의학회지 26(1):93-102, 2005.
19. 최정락, 최재영, 이영섭, 박성식. 태음인 수면의 임상적 특징(로지스틱 회귀분석을 이용하여). 사상체질의학회지 16(3):18-24, 2004.
20. 김대윤, 이재원. 사상의학 체질진단 객관화에 대한 통계적 연구. Proceedings of the spring conference, Korean Statistical Society. pp 228-233, 1999.