사상체질 진단검사를 위한 데이터마이닝 알고리즘 연구

홍진우 · 김영인[1] · 박소정[2] · 김병철[1] · 엄일규[3] · 황민우 · 신상우[4] · 김병주[2] · 권영규[2] · 채 한[2]*

부산대학교 한의학전문대학원 임상의학부, 1 : 부산대학교 생명자원과학대학 바이오메디컬공학과,
2 : 부산대학교 한의학전문대학원 양생기능의학부, 3 : 부산대학교 전자전기공학부, 4 : 부산대학교 한의학전문대학원 응용의학부

Data mining Algorithms for the Development of *Sasang* Type Diagnosis

Jin Woo Hong, Young In Kim[1], So Jung Park[2], Byoung Chul Kim[1], Il Kyu Eom[3], Min Woo Hwang, Sang Woo Shin[4], Byung Joo Kim[2], Young Kyu Kwon[2], Han Chae[2]*

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

This study was to compare the effectiveness and validity of various data-mining algorithm for *Sasang* type diagnostic test. We compared the sensitivity and specificity index of nine attribute selection and eleven class classification algorithms with 31 data-set characterizing *Sasang* typology and 10-fold validation methods installed in Waikato Environment Knowledge Analysis (WEKA). The highest classification validity score can be acquired as follows; 69.9 as Percentage Correctly Predicted index with Naive Bayes Classifier, 80 as sensitivity index with LWL/Tae-Eum type, 93.5 as specificity index with Naive Bayes Classifier/So-Eum type. The classification algorithm with highest PCP index of 69.62 after attribute selection was Naive Bayes Classifier. In this study we can find that the best-fit algorithm for traditional medicine is case sensitive and that characteristics of clinical circumstances, and data-mining algorithms and study purpose should be considered to get the highest validity even with the well defined data sets. It is also confirmed that we can't find one-fits-all algorithm and there should be many studies with trials and errors. This study will serve as a pivotal foundation for the development of medical instruments for Pattern Identification and *Sasang* type diagnosis on the basis of traditional Korean Medicine.

Key words : waikato environment knowledge analysis, *sasang* type diagnosis, pattern identification, sensitivity and specificity, clustering algorithm, data field selection

서 론

데이터 마이닝(data mining)이란 1960년도 초 미국의 통계학자 제임스 마이어와 에드워드 포지가 제시한 것으로, 대량의 데이터에 내재되어 있는 중요한 잠재 패턴 등 유용하고 새로운 정보를 추출해냄으로써 의사결정을 위한 지식을 제공함에 사용되는 기법과 분석과정을 의미한다[1-3].

* 교신저자 : 채 한, 경남 양산시 물금읍 범어리 부산대학교 한의학전문대학원
· E-mail : han@chaelab.org, · Tel : 051-510-8470
· 접수 : 2009/09/23 · 수정 : 2009/11/09 · 채택 : 2009/11/23

최근 들어 데이터 마이닝이 다양한 생체정보의 분석에 응용[4]됨에 따라 임상연구자도 사용할 수 있을 정도로 사용법이 용이하여졌으며, 이에 한의학적 임상 데이터와 분석 모형에도 적용할 수 있는 여건이 급격히 성숙되어가고 있다. 그러나 다양한 데이터마이닝 기법들과 이들을 활용하기위한 SPSS Clementine, SAS Enterprise Miner, IBM DB2 Intelligent Miner for Data 등과 같은 상용 소프트웨어가 지속적으로 개발되고 있음에도 불구하고 자신의 목적과 환경에 맞는 소프트웨어를 선택하고 그 기능들을 충분히 활용하기위한 연구들은 아직 부족한 것이 현실이다[5].

더욱이 한의학적인 임상 특성을 분석하기 위한 데이터마이

닝 기법의 활용에 있어서는, 그 중요성에 비해 만족할 만한 연구나 기법의 개발이 이루어지지 못하여 왔다. 특히 한의 진료과정에 있어서 한의학적 변증이나 사상체질의 진단 또는 감별이라는 의료행위는 데이터마이닝 기법에 있어서는 유형들을 구분 또는 분류한다는 것을 의미하기에 분류 알고리즘에 대한 연구가 매우 중요한 의미를 지니게 되나, 이러한 알고리즘들의 특성을 체계적으로 검토할 수 있는 한의학적 데이터에 알맞은 타당도와 민감도와 같은 연구방법론은 최근에서야 제시되고 있다[6,7].

데이터마이닝 기법 중 분류 알고리즘은 적용 분야의 데이터 특성이나 활용 방식에 따라 확연한 성능 차이를 보이게 되는데, 사상의학 연구에 있어서 가장 적합한 분류 알고리즘을 찾기 위한 본격적인 비교연구는 아직까지 시도되지 못하였다. 기존의 연구에서는 판별분석[8]을 비롯한 neural network[9], decision tree[10-12], K-means[13], SVM[13] 등이 단편적으로 시도되어 왔으나, 이들의 분류 특성을 체계적으로 비교한 연구[6]는 이루어지지 못하여왔다.

이에 본 연구에서는 데이터마이닝 분류 알고리즘을 사용하여 사상의학 연구에 있어서 가장 높은 타당성을 보이는 분류 알고리즘을 찾아보고자 하였으며, 이와 함께 모든 속성을 적용한 경우와 속성 부분집합을 선택하여 적용한 결과를 비교함으로써 적은 수의 데이터만으로도 전체 데이터를 사용한 결과와 동일한 타당성을 보일 수 있을지에 대해서도 알아보고자 하였다.

이에 데이터 마이닝 알고리즘간의 체계적 비교를 위한 방법론적 편의를 제공함과 동시에 차후 연구결과의 응용이 용이한 지식분석을 위한 와이카토 환경(Waikato Environment for Knowledge Analysis, WEKA)을 사용하였다. 본 연구에서 활용한 WEKA[14]는 뉴질랜드 와이카토 대학교의 컴퓨터 과학과의 이안 위튼(Ian Witten) 교수가 이끄는 프로젝트의 결과물로서, 학습과 연구에 사용할 목적으로 누구나 쉽게 사용할 수 있는 JAVA기반 범용 기계 학습 라이브러리이다(Fig. 1).

WEKA는 오픈소스(open source) 소프트웨어이기 때문에 누구나 원하는 코드를 덧붙이거나 수정하여 사용할 수 있다는 장점도 지니고 있는데, 일반적으로 데이터마이닝 분야에서 필요로 하는 기본적인 알고리즘들이 대부분 구현되어 있다. 예를 들면 스팸메일 등을 자동으로 분류하는데 사용되는 베이지언 네트워크(bayesian network)나, 문서의 자동 군집화에 사용되는 k-means 알고리즘, 예측 규칙 도출을 위한 ID4, 상관규칙(association rule) 추출 알고리즘 등 관련 분야의 교과서에 나오는 기계 학습 알고리즘 등이 포함되어 있다. 사용자들은 간단한 인터페이스를 통해 이러한 알고리즘들을 불러 사용할 수 있으며, 그 결과를 그래프로 쉽게 표현할 수도 있다[4,14].

본 연구에서는 이러한 WEKA를 사용하여 사상의학에서의 체질 분류에 활용할 수 있는 다양한 데이터마이닝 알고리즘의 성능 특성을 비교하였으며, 분류알고리즘의 성능을 비교함에 있어서 최적의 속성을 선정하기위한 속성선정 알고리즘 사용을 동시에 진행하여 효율성에 대한 검토를 진행하였다. 아울러 정확예측율, 일반화 유형 민감도, 일반화 유형 특이도를 타당도 지표로 사용하여 장점과 단점에 대한 객관적인 비교를 시행하고 이를 보고하는 바이다.

연구방법

1. 연구 데이터

알고리즘의 분류 능력을 평가함에 있어서 채 등[9]의 연구 데이터를 사용하였다. 연구 참여자는 19세에서 43세 사이의 대학생 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[8].

MBTI 검사는 95개의 문항을 사용하여 개인의 성격적 특징을 검사하는 도구로서, 성격을 외향(Extraversion)/내향(Introversion), 감각(Sensing)/직관(Intuition), 사고(Thinking)/감정(Feeling), 판단(Judging)/인지(Perceiving)의 네 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, 체지방, 무기질, 단백질 등의 분포에 대한 신뢰할만한 측정치를 추출할 수 있다는 장점을 지니고 있다[16].

본 연구에 사용된 31개 속성(attribute)의 명칭과 의미는 다음과 같다. 1=성별 (Sex), 2=연령 (Age), 3=MBTI EI (M_EI), 4=MBTI SN (M_SN), 5=MBTI TF (M_TF), 6=MBTI JP (M_JP), 7=Intracellular water (ICF), 8=Extracellular water (ECF), 9=Protein mass (PM), 10=A mass (AMM), 11=Body fat mass (BFM), 12=키 (Height), 13=몸무게 (Weight), 14=Percent body fat (PBF), 15=Waist-hip ratio (WHR), 16=Fluid of right arm (FRA), 17=FLA, 18=Fluid of trunk (FT), 19=Fluid of right leg (FRL), 20=Fluid of left leg (FLL), 21=Total body water (TBW), 22=Body Mass Index (BMI), 23=Arm/Leg ratio (ALR), 24=신체발달 (BODY_DEV), 25=적정체중 (PROP_WGT), 26=체중조절 (WGT_CTL), 27=지방조절 (FAT_CTL), 28=근육조절(MUCL_CTL), 29=상체우/좌 (UPR_RL), 30=하체우/좌(DWN_RL), 31=사상체질 (Class)

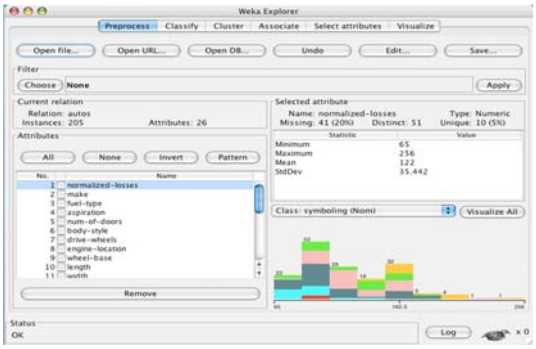
2. 연구 방법

1) 와이카토 지식분석 환경 (Waikato Environment for Knowledge Analysis, WEKA)

본 연구에서는 WEKA ver. 3.6.0에 포함되어 있는 분류 및 속성 부분집합 선정 알고리즘을 사용하였다. WEKA의 메뉴14)는 전처리(pre-processing), 분류(classify), 군집화(cluster), 상관(associate), 속성 부분집합의 선택(select attributes), 가시화(visualize)로 구성되어있다(Fig. 1). 본 소프트웨어는 인터넷 주소 http://www.cs.waikato.ac.nz/ml/weka/ 에서 무료로 다운로드 받을 수 있다.

A. Weka GUI chooser

B. Window for exploration



C. Window for atrribute selection

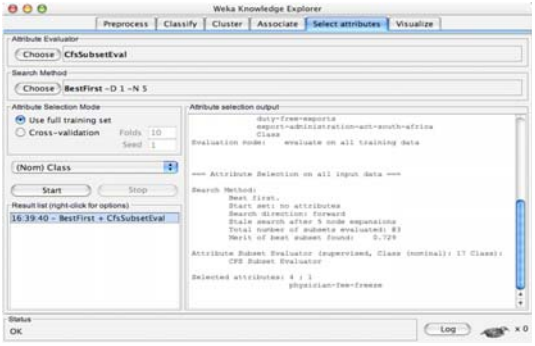


Fig. 1. Characteristic Features of Waikato Environment for Knowledge Analysis (WEKA)

2) 체질 분류 알고리즘

분류(classification) 학습 방법들로서는 decision tree, list, instance-based classifier, support vector machine, multi-layer perceptron, logistic regression, Naive bayes 등이 포함되어 있으며, 메타 분류 방법들로서는 bagging, boosting, stacking, error-correcting output codes, locally weighted learning 등이 포함되어 있다[14].

본 연구에서 우수한 성능을 보인 주요 분류 알고리즘들의 명칭과 약어는 다음과 같으며, 알고리즘별 parameter로는 대부분 WEKA의 default 값[14]을 사용하였으며 경우에 따라 일부를 수정하였다.

NB : Naive Bayes Classifier, Jrip : a propositional rule learner using Repeated Incremental Pruning to Produce Error Reduction (RIPPER), J48 : C4.5 decision tree, MLP : Multilayer Perceptron, SMO : Sequential Minimal Optimization algorithm of a support vector machine, LWL : Logistic Model Trees, DT : Decision Table, RF : Forest of Random Trees, SL : linear logistic regression model, LMT : logistic model tree, REP T : decision/regression Tree using Reduced-Error Pruning DT : Decision Table, RF : RandomF, B1 : Baysian1 (Bayesian Network using a hill climbing algorithm), B2 : Baysian2 (Bayesian Network using K2 algorithm)

3) 속성의 선정 (Attribute selection)

본 연구에 있어서는 분류 학습에 30개의 모든 속성 부분집합을 사용하여 분류 알고리즘의 타당도를 고찰 한 후, 속성 부분집합 선정을 통해 확인한 유의한 수개의 속성 부분집합을 선택하여 한 번 더 분류 학습을 시행하였는데, 이는 보다 적은 숫자의 데이터 필드를 사용하여 가장 높은 분류 효과를 보일 수 있는 데이터 필드를 찾기 위한 통상적인 방법이다.

WEKA에서 시행할 수 있는 주요 속성 부분집합 선택(select attributes) 방법으로서, 탐색법(search method)으로는 best-fit, forwqrd selection, random, exhaustive, genetic algorithm, ranking 등이 포함되어 있으며, 평가법(evaluation method)으로는 correlation-based, wrapper, information gain, chi-squared 등이 포함되어 있다. 아울러 WEKA는 이러한 두가지 방법들을 혼합하여 사용할 수도 있다[14].

본 연구에 사용된 속성 부분집합 선택 알고리즘으로는 모두 9개가 사용되었는데 그 명칭과 사용 기법은 다음과 같다. AS1 : CfsSubsetEval, AS2 : ChiSquaredAttributeEval, AS3 : ClassifierSubsetEval, AS4 : GainRatioAttributeEval, AS5 : InfoGainAttributeEval, AS6 : OneRAttributeEval, AS7 : PrincipalComponents, AS8 : SVMAttributeEval, AS9 : Symmetrical UncertAttributeEval, 또한 선정된 9개를 토대로 2개를 추가로 추출해내었는데, AS10은 중복 선정된 상위 속성 6개를 선택한 것이며, AS11은 중복 선정된 상위 속성 9개를 선택한 것이다(Table 1).

4) 성능 타당도 비교

성능 타당도 검증에 있어서는 분류의 정확도를 검증하기위하여 10겹 교차검증법(10 fold colss-validation)이 사용되었다. 또한 성능 타당도를 비교하는 지표[6]로서는 정확예측율, 체질별 일반화 유형 민감도(Qd)와 체질별 일반화 유형 특이도(Qm)를 사용하였다.

정확예측율은 유형이 옳게 분류된 총 도수를 사용된 모든 빈도수로 나눈 것이며, 체질별 일반화 유형 민감도(Qd)란 해당

유형의 실제 총수에 대한 해당 유형으로 진단한 경우의 비율이며, 체질별 일반화 유형 특이도(Qm)는 해당 유형으로 진단한 총수에 대한 실제 해당 유형인 경우의 비율을 의미한다[6]. 예를 들어 소양인 일반화 유형 민감도(SY-Qd)는 실제 소양인을 소양인으로 진단해 낸 비율을 의미하며, 소양인 일반화 유형 특이도(SY-Qm)는 소양인 진단한 경우에 있어서 실제 소양인의 비율을 의미한다.

Table 1. Result of attribute selection.

Attribute Set	Selected attribute
AS1	M-EI, ICF, BFM, Weight, WHR, WGT-CTL, FAT-TCL
AS2	M-EI, ICF, BFM, Weight, WGT-CTL, FAT-CTL, TBW, PM, AMM
AS3	M-EI, M-JP, Weight, WGT-CTL, FAT-CTL
AS4	M-EI, ICF, BFM, WGT-CTL, FAT-CTL, TBW, PM, AMM
AS5	Same as AS2
AS6	M-EI, Weight, FLA, BMI
AS7	ICF, PM, AMM, FT, FLA, FRA, TBW, FRL, FLL, Weight, PROP-WGT
AS8	M-EI, M-TF, M-JP, ECF, BFM, WHR, DWN-RL, FLA, FLL, ALR, FAT-CTL
AS9	Same as AS2
AS10	M-EI, ICF, BFM, Weight, WGT-CTL, FAT-CTL
AS11	M-EI, FAT-CTL, Weight, GFM, ICF, WGT-CTL, AMM, PM, TBW

* Attribute Sets are as follows. AS1 : CfsSubsetEval, AS2 : ChiSquaredAttributeEval, AS3 : ClassifierSubsetEval, AS4 : GainRatioAttributeEval, AS5 : InfoGainAttributeEval, AS6 : OneRAttributeEval, AS7 : PrincipalComponents, AS8 : SVMAttributeEval, AS9 : Symmetrical UncertAttributeEval, AS10 : Six frequently selected attributes, AS11 : Nine frequently selected attributes. Selected attributes of AS2, AS5 and AS9 are the same.

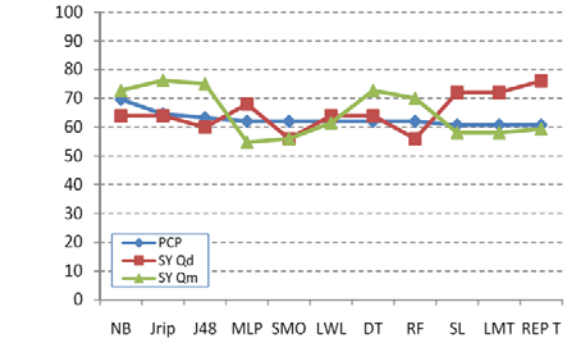
결 과

1. 모든 속성을 사용한 분류 알고리즘의 성능 타당도 비교

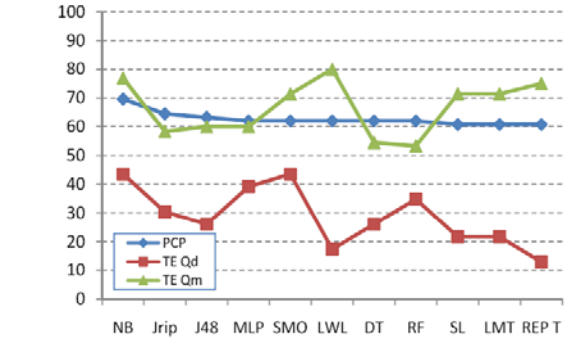
30개의 모든 속성을 사용한 분류 알고리즘의 성능 타당도 비교 결과 높은 효율을 보인 11개의 알고리즘은 다음과 같다(Fig. 2). 정확예측율(PCP)을 기준으로 비교하였을 때 가장 높은 효율을 보인 세 개의 알고리즘은 Naive Bayes Classifier (69.6), Jrip(64.6), J48(63.3)이었으며, 이들 중 가장 낮은 효율을 보인 것은 SimpleLogistics(60.8), LMT(60.8), REP Tree(60.8)이었다.

각 체질별 분류 알고리즘의 분류 성능을 일반화 유형 민감도(Qd)와 일반화 유형 특이도(Qm)를 기준으로 비교한 결과는 Fig. 2와 같으며, 각 체질별 민감도와 특이도의 고저를 정확예측율과 비교하기 위하여 함께 표시하였다. 소양인(Fig. 2A)에 있어서 가장 높은 일반화 유형 민감도를 보인 세 알고리즘은 REP Tree(76), Simple Logistics(72), LMT(72)이었으며, 가장 높은 일반화 유형 특이도를 보인 세 알고리즘은 Jrip(76.2), J48(75), Naive Bayes Classifier(72.7), DT(72.7)이었다. 태음인(Fig. 2B)에 있어서 가장 높은 일반화 유형 민감도를 보인 세 알고리즘은 Naive Bayes Classifier(43.5), SMO(43.5), MLP(39.1)이었으며, 가장 높은 일반화 유형 특이도를 보인 세 알고리즘은 LWL(80), Naive Bayes Classifier(76.9), REP Tree(75)이었다. 소음인(Fig 2C)에 있어서 가장 높은 일반화 유형 민감도를 보인 세 알고리즘은 Naive Bayes Classifier(93.5), J48(93.5), LWL(93.5)이었으며, 가장 높은 일반화 유형 특이도를 보인 세 알고리즘은 MLP(69.7), Naive Bayes Classifier(65.9), SMO(62.5)이었다.

A. So-Yang type



B. Tae-Eum type



C. So-Eum type

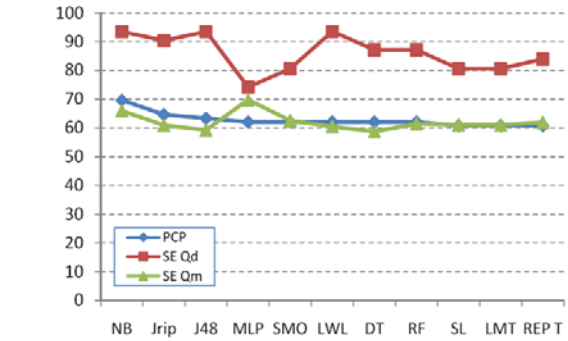


Fig. 2. Eleven algorithms with high sensitivity(Qd) and specificity(Qm) using all attributes (Diamond; PCP, Square; type specific sensitivity(Qd), Triangle; type specific specificity(Qm)). * NB: Naive Bayes Classifier, Jrip: a propositional rule learner using Repeated Incremental Pruning to Produce Error Reduction (RIPPER), J48: C4.5 decision tree, MLP: Multilayer Perceptron, SMO: Sequential Minimal Optimization algorithm of a support vector machine, LWL: Logistic Model Trees, DT: Decision Table, RF: Forest of Random Trees, SL: linear logistic regression model, LMT: logistic model tree, REP T: decision/regression Tree using Reduced-Error Pruning.

2. 속성 부분집합 선택 결과를 활용한 분류 알고리즘의 성능 타당도 비교

30개의 속성 중에서 속성 선정 알고리즘을 사용하여 11가지의 데이터 세트를 얻을 수 있었는데, 그중 2개는 9개의 선정을 토대로 추출해 낸 것이다(Table 1). 아울러 AS2, AS5, AS9로는 동일한 데이터 세트가 제시되었으므로, 연구 과정에 있어서 분류 성능을 비교함에 있어서는 AS2만 사용되었다.

선정된 9개의 데이터세트를 대상으로 다양한 분류 알고리즘별 성능 타당도를 정확예측율(PCP)을 사용하여 비교한 결과 상위 10개의 알고리즘은 Table 2에 제시한 바와 같다. 이러한 결과를 Fig. 2에서 제시된 알고리즘과 비교해 볼 때 NB, Jrip, J48,

MLP, SMO, LWL, DT, RF는 속성 전체를 사용했던 경우와 동일하게 비교적 높은 타당도를 보임을 알 수 있었다. 이에 데이터 세트 1, 2, 10, 11을 사용한 Naive Bayes Classifier, Baysian1, Baysian2의 세 알고리즘이 가장 높은 69.62라는 정확예측율(PCP)을 보였으며, MLP(데이터 세트 4, 11), SMO(데이터 세트 1, 3, 4), LWL(데이터 세트 3), Random F(데이터 세트 11)은 64.56으로 가장 낮은 정확예측율(PCP)을 보였다.

Table 2. Ten algorithms with high Percentage Correctly Predicted (PCP) using selected attributes.

Algorithm	NB	Jrip	J48	MLP	SMO
PCP	69.62	67.09	65.82	64.56	64.56
Attr. set No.	AS1, AS2, AS10, AS11	AS6	AS6	AS4, AS11	AS1, AS3, AS4
Algorithm	LWL	DT	RF	B1	B2
PCP	64.56	67.09	64.56	69.62	69.62
Attr. set No.	AS3	AS4	AS11	AS1, AS2, AS10, AS11	AS1, AS2, AS10, AS11

* NB : Naive Bayes Classifier, Jrip : a propositional rule learner using Repeated Incremental Pruning to Produce Error Reduction (RIPPER), J48 : C4.5 decision tree, MLP : Multilayer Perceptron, SMO : Sequential Minimal Optimization algorithm of a support vector machine, LWL : Logistic Model Trees, DT : Decision Table, RF : Forest of Random Trees, RF : RandomF, B1 : Baysian1 (Bayesian Network using a hill climbing algorithm), B2 : Baysian2 (Bayesian Network using K2 algorithm)

고찰 및 결론

WEKA와 같은 데이터마이닝 소프트웨어는 자동화된 단백질 분석, 유전발현 분석을 위한 probe 선택, 자동화된 암 진단, 식물 유전형 구별, 유전 발현 프로파일 구분, 흡연 상태 구별[17] 등과 같은 다양한 바이오인포메틱스(bioinformatics) 분야에 사용될 수 있는 여러 가지 알고리즘들을 포함하고 있는데[4,14], 이처럼 구별이나 구분 등에 사용한다는 점에서 볼 때 임상 과정에 있어서 사상체질의 진단 또는 변증(辨證)이 필수적인 한의학적 임상 데이터 분석에 매우 유용한 도구라 할 것이다.

그러나 현재까지 이러한 데이터마이닝을 한의학적 임상에 활용할 수 있는 방법들에 대한 연구는 활발히 진행되지 못하여 왔다. 기존 한의학 연구에 있어서는 생물학적 데이터의 분석방법으로서 단일 변수에 있어서 두(세) 그룹간 '차이가 있을 것인가'에 대한 가설에서 시작되어, 영가설(Ho)을 받아들일 것인가(그룹간 차이가 없다) 혹은 영가설을 기각할 것인가(그룹간 차이가 있다)라는 차이의 확인 또는 검증에 대한 논의가 주로 사용되어 왔다. 이러한 단일변수를 위주로 하여온 서양의학적 방법론의 한계는, 다원성 또는 복합적 분석과 치료를 특징으로 하는 한의학적 연구방법론의 개발에 있어 모델링 또는 그룹 예측을 위한 통계적 방법론이 개발되지 못하여 온 주요 원인이 되어 왔다.

최근 들어 단일 변수에서의 차이보다 수개의 변수가 복합적으로 얽혀있는 거시적인 프로파일(profile)의 차이를 검증하는 연구방법론을 사용한 연구[8,18]가 시도되고 있는데, 이러한 측면에서 본다면 본 연구에서와 같이 기계학습법을 사용해 데이터에 내재되어있는 숨겨진 패턴을 찾아내서 모형화(modeling)한다는 것은 한의학 임상 및 한의학 연구 방법론에 있어서의 한 단계 진보라 할 수 있을 것이다.

한의학적 변증 과정을 위해서는 본 연구에서 사용된 데이터 마이닝 기법이 다양하게 응용되어야 할 것인데, 이는 데이터 마이닝이 많은 데이터 중에서 중요한 변수를 찾아내어 분석하고 이를 통해 높은 예측력을 지닌 모형을 신속하게 산출하는 능력을 지니기 때문이다[2]. 예를 들어 한의학적 변증(辨證) 시스템은 단순히 한두 개의 진단 정보에만 의존하는 것이 아니기 때문인데, 설진의 경우라면 혀의 형태, 색, 윤택함, 표면의 질감 등의 정보를 복합적으로 사용해야 하며 동시에 여타 맥진, 병의 증상, 치료 경과 등도 종합적으로 고려되어야 하는 복잡한 데이터 구조를 지니기 때문이다.

이러한 측면에서 본다면, WEKA는 여러 가지 장점을 지니고 있다. 전문적인 교육과 높은 비용을 전제로 하는 각종 상용 프로그램을 사용하기에는 연구비 또는 연구 방법론에 대한 전문지식 등이 매우 부족한 것이 현실이므로, WEKA의 사용은 이러한 한계에도 불구하고 다양한 알고리즘의 시도를 가능하게 한다.

또한 대용량 데이터를 사용하는 통상적인 데이터마이닝과는 달리 소용량 데이터에 기반한 다양한 알고리즘의 시도가 필요한 한의학 연구에 있어서는 본 연구에 사용한 WEKA가 더욱 큰 장점을 지니고 있는데, 최근 도입된 프로필 분석(profile analysis)[18] 과 같이 거시적인 경향성을 토대로 한 다양한 특성분석을 시도함으로써 한의학 연구와 임상에 적절한 새로운 통계적 기법을 발굴할 수 있는 여지를 제공하기 때문이다.

아울러 WEKA에서는 데이터의 전처리(pre-processing) 혹은 데이터 필드의 선택(data field selection) 등에 활용할 수 있는 다양한 소프트웨어 메뉴(Fig. 1)를 제공하고 있는데, 이는 데이터 분석과정에서의 불필요한 가중치를 적절히 제거할 수 있으며 동시에 최소한의 데이터 수집을 통해서도 높은 타당도를 유지할 수 있도록 한다. 한의학 임상 데이터의 분석에 있어서는 데이터에서 중요한 정보를 추출해내는 전체 과정, 즉 데이터 수집, 데이터 선택, 데이터 전처리, 데이터 분석, 한의학적 진단 등의 전 과정을 시스템적 측면에서 최적화할 필요가 있는데, WEKA에는 이를 위한 다양한 기법들이 포함되어 있다(Fig. 1).

본 연구에 있어서 정확예측율이나 민감도, 특이도에 있어 가장 높은 효율을 보인 것은 가설을 바탕으로 한 학습을 통해 사후확률을 구함으로써 최적의 상관성을 찾아나가는 베이지안 알고리즘이었다. 본 연구에 사용된 알고리즘에 있어서 NB는 supervised discretization을 사용한 Naive Bayes Classifier이며, B1(Baysian1)은 탐색 알고리즘으로서 hill climbing algorithm을 사용한 Bayesian Network이며, B2(Baysian2)는 탐색 알고리즘으로서 K2 algorithm을 사용한 Bayesian Network을 의미한다. 이에 B1/B2에 있어서 주요 파라미터로서 Markov Blanket correction (yes), maximum number of parents (1), score type (Bayes)을 사용하였다. 그러나 이러한 결과가 데이터의 특성에 기반한 상황의존적 결과이므로 데이터의 종류와 특성에 의해 크게 좌우될 수 있음도 확인할 수 있었는데, 정확예측율에 있어서 가장 낮은 타당도를 보였던 Simple Logistics, LMT, REP Tree가 소양인 일반화 유형 민감도에 있어서는 가장 높은 타당도를 보이고 있었다.

정확예측율에 있어서는 알고리즘에 따라 60.8에서 69.6까지

9.1(13%)의 차이가 있음을 확인할 수 있었다. 체질별 일반화 민감도에 있어서 알고리즘에 따른 최소-최대간의 차이가 소양인(Fig. 2A)에 있어서는 20(26.3%), 태음인(Fig. 2B)에 있어서 30.5(70.1%), 소음인(Fig. 2C)에 있어서는 19.3(20.5%)를 보였다. 또한 체질별 일반화 특이도에 있어서 알고리즘에 따른 최소-최대간 차이는 소양인(Fig. 2A)에 있어서는 21.4(28.1%), 태음인(Fig. 2B)에 있어서 26.7(33.4%), 소음인(Fig. 2C)에 있어서는 10.5(15.1%)를 보였다. 이에 태음인 일반화 민감도에서의 최대 차이가 70.1%로 가장 크게 나타난 반면 소음인 일반화 특이도에서는 최대 차이가 15.1%로 가장 작게 나타났음을 알 수 있었다.

아울러 결과(Fig. 2)에 있어서 한 가지 흥미로웠던 점은, 특이도(Qm)는 정확예측율(PCP)과 거의 유사하게 나타났던 반면, 민감도(Qd)는 정확예측율과 비교할 때 비교적 높거나(소음인, Fig. 2C), 유사하거나(소양인, Fig. 2A), 낮게(태음인, Fig. 2B) 나타났다는 점인데, 이러한 결과가 나타난 원인에 대해서는 심도 있는 추가연구가 필요할 것이라 보인다.

이상의 분류 타당성 비교 결과에서 각 지표별 타당도가 높고 낮음이 발생하는 상세한 원인에 대한 심도 있는 분석을 시행할 수는 없었다. 이는 각 알고리즘의 상세한 데이터 처리 방식이나 기계학습 과정들을 한 번에 모두 비교한다는 것은 현실적으로 불가능하기 때문이다. 또한 통상적인 데이터 프로세싱 과정에 수반되는 최적화 과정을 생략하였기 때문에 단정적으로 한 알고리즘이 다른 것보다 확연히 우월하다 언급할 수는 없는 것이다.

본 연구에서는 이처럼 알고리즘의 장단점들을 세세하게 분석할 수 없다는 한계점을 또 한 번 확인할 수 있었는데, 이는 알고리즘간의 상세한 타당성 비교를 위해서는 다양한 평가지표 또는 평가방법의 개발이 시급히 요구됨[6]을 또 한 번 반증하는 것이라 사료된다. 알고리즘들의 다면적 타당성을 객관적으로 비교[6]하기 위해서는 한의학적 증상 데이터와 진단 로직(logic)에 내재되어 있는 문턱값(threshold), 민감도와 특이도간의 상관성, ROC 곡선의 조정 등과 같은 한의학적 연구방법론의 개발[7] 또한 시급히 요구된다 하겠다.

데이터를 선택한 경우(Table 2)와 모든 데이터를 함께 사용하였던 경우의 최대 타당도는 모두 69.6으로 동일하게 나타났는데, 이는 데이터에서의 중복된 정보를 제거하는 적절한 방법으로서 기존에 제시되었던 특징추출과 상관도 분석[9] 외에 WEKA에 포함된 데이터 마이닝 방법을 활용함이 적절하다는 것을 의미한다 사료된다. 가장 높은 타당도를 보였던 선택 데이터 세트(Table 1, 2)는 AS1, AS2, AS10, AS11이었는데, M_EI, ICF, Weight, WGT_CTL, FAT_CTL 이 모든 데이터세트에 공통적으로 포함되어 있었으며, BFM은 AS1, AS2, AS10에, TBW, AMM과 PM은 AS2, AS11에, GFM은 AS11, WHR은 AS1에 포함되어 있었다. 이러한 결과는 심리 지표로서의 외향-내향성과 함께 신체 지표로서의 체내수분량, 몸무게, 그리고 연령/성별 표준 체중/지방량이 체질 구별에 있어서 가장 중요한 의미를 지닐 것임을 의미한다고 보는데, 이는 기존 연구 결과들[8,19,20]을 재확인하는 것이라 사료된다.

본 연구는 데이터 선택 알고리즘 및 분류 알고리즘[9], 그리고 사용된 데이터 자체의 특성에 따라 지표별 타당도가 전혀 다르게 나타날 수 있다는 점을 확인하였는데, 이러한 점이 본 연구의 가장 중요한 가치로서 decision tree[11,12]나 SVM[13] 등을 사용하였던 기존 연구에 사용되었던 알고리즘들의 타당도와 데이터의 활용성에 대하여 민감도나 타당도를 바탕으로 한 체계적인 재검토[6,7]가 필요함 것이라 보인다.

본 연구의 결과에서 볼 때 베이지안 알고리즘들이 높은 타당성을 보이고 있는 것을 확인할 수 있었는데, 한의학적 데이터 분석에 있어서의 연구방법론에 대한 기존 연구[21]에서도 한의학적 개념을 구현함에 가장 높은 타당성을 지니고 있는 알고리즘으로서 본 연구에서 살펴본 베이지안이 제시되고 있는 것으로 볼 때, 기존 연구 결과에 대한 전면적인 검토가 필요함이 한 번 더 역설되고 있는 것이라 사료된다.

예를 들어 SVM(SMO)의 경우 본 연구에 있어서는 polynomial kernel을 사용하고 주요 파라미터로서 complexity (1.0), epsilon (1.0E-12), exponent (1.0), tolerance (0.001)을 사용하였으나, 기존의 연구 보고에 있어서는 어떠한 kernel/parameter를 사용하였는지 확인할 수 없었으므로 본 연구에서 이들 특성에 따른 비교는 할 수 없었다.

동일한 데이터와 모형을 사용하더라도 상황에 따라 알고리즘의 타당성이 현격한 차이를 보이는 것은 이미 알려져 있는 사실로서, 최신의 데이터마이닝 기법들을 안다고 해서 모든 것이 해결되는 것이 아니라 실제 분석에 들어가기에 앞서 데이터의 특성을 파악하고 정교한 분석 모형을 제시하여 분석의 목적과 환경에 적절한 알고리즘을 운용하는 것이 무엇보다 중요하다[5].

이에 본 연구 데이터와 다른 특성을 지닌 측정치, 예를 들어 신체 계측치, 성음분석 결과, 안면 인식 결과 등을 분석함에 있어서 WEKA에서 제공되는 다양한 알고리즘들을 적용하였을 때 다면적 타당도 지표가 어떻게 나타날 것인지 확인하기위한 시스템적 접근[6,7]이 시급히 요구된다 하겠다.

본 연구에서 한의학적 데이터 분석의 전체과정에 있어서 WEKA를 사용하여 체계적으로 접근할 수 있음을 확인하였다는 점은, 한의학 연구를 위한 새로운 연구방법론의 개발이라는 측면에서 큰 의의를 가진다고 하겠다. 앞으로 한의학적 특성을 지닌 데이터들을 분석하기위한 최적화된 알고리즘의 구현과 응용[9], 그리고 이들에 대한 효율적 타당성 분석법의 개발[6,7] 등에 대한 연구방법론 측면에 대한 추가적인 연구가 지속적으로 이루어져야 할 것이다.

감사의 글

이 논문은 2009년도 한국한의학연구원의 지원을 받아 기관고유사업의 일환으로 수행된 연구임(K09011).

참고문헌

1. 허 준, 최병주. (클레멘타인을 이용한)데이터마이닝 : 입문편. SPSS 아카데미, 서울, 2001.

2. 이창희. Feature selection을 이용한 중요변수 선택방법에 관한 연구. 석사학위논문, 중앙대학교, 2007.
3. Michael, J.A. Berry and Gordon Linoff, Data Mining Techniques: For Marketing, Sales, and Customer Support, John Wiley & Sons, Inc., 1997.
4. Frank, E., Hall, M., Trigg, L., Holmes, G., Witten, I.H. Data mining in bioinformatics using Weka. Bioinformatics. 20(15):2479-2481, 2004.
5. 이덕기. 데이터마이닝 소프트웨어의 효율성에 관한 비교 연구. 석사학위논문. 호서대학교, 2001.
6. 이수진, 김명근, 채 한. 사상체질 진단검사 타당성분석에 대한 연구. 대한한의학회지 29(1):7-14, 2008.
7. 황상문, 박소정, 강기림, 권영규, 채 한. 사상체질 진단검사 타당성 분석지표의 일반화 연구. 동의생리병리학회지 23(5):950-957, 2009.
8. Chae, H., Lyoo, I.K., Lee, S.J., Cho, S., Bae, H., Hong, M., Shin, M. An alternative way to individualized medicine: psychological and physical traits of Sasang typology. Journal of Alternative and Complementary Medicine. 9(4):519-528, 2003.
9. 채 한, 황상문, 엄일규, 김병철, 김영인, 김병주, 권영규. 신경망을 사용한 사상체질 진단검사 개발 연구. 동의생리병리학회지 23(4):765-771, 2009.
10. 박은경, 이영섭, 박성식. 의사결정나무법을 이용한 체질진단에 관한 연구. 사상체질의학회지 13(2):144-155, 2001.
11. 진희정, 문진석, 고성호, 구임회, 이시우, 이도현, 송미영, 김종열. 사상체질 의사결정시스템 구축을 위한 체질 진단 자료를 이용한 예비연구. 한국한의학연구원논문집, 13(2):75-81, 2007.
12. 박성식, 최재영. 의사결정나무법을 이용한 설문지의 응답특성에 대한 임상적 검토. 사상체질의학회지 15(3):177-186, 2003.
13. Zhang, Q., Lee, K.J., Whangbo, T.K. 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 pp 18-21, 2007, Jinan, China, pp 426-430, 2007.
14. Witten, I.H., Frank, E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco, C.A, 2000.
15. 김선호, 고병희, 송일병. 사상체질분류검사지(QSCCⅡ)의 표준화 연구. 사상의학회지 7(1):187-246, 1995.
16. 이수진, 박수현, 고유선, 박수진, 엄일규, 김병철, 김영인, 백진웅, 김명근, 권영규, 채한. 임피던스 분석을 활용한 사상인의 신체계측 연구. 동의생리병리학회지 23(2):433-437, 2009.
17. 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.
18. Park, S.H., Kim, M.G., Lee, S.J., Kim, J.Y., Chae, H. Temperament and Character Profiles of Sasang Typology in an Adult Clinical Sample. Evid Based Complement Alternat Med. doi:10.1093/ecam/nep034, Advance Access published April 20, 2009.
19. Chae, H., Park, S.H., Lee, S.J., Kim, M.g., Wedding, D., Kwon, Y.K. Psychological profile of Sasang typology: A systematic Review. eCAM, 6: 21-29, 2009.
20. 채 한, 박수잔, 이수진, 고광찬. 사상 유형학의 성격심리학적 고찰. 대한한의학회지 25(2):151-164, 2004.
21. Lukman, S., He, Y., Hui, S.C. Computational medthos for Traditional Chinese Medicine: A survey. Computer methods and programs in biomedicine. 88: 283-294, 2007.