Journal of The Korea Society of Computer and Information
Vol. 19, No. 11, November 2014

www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2014.19.11.149

# Mixture of Expert 모형에 기반한 당뇨병 진단 분류

이 홍 기*, 명 성 민**

# Classification of the Diagnosis of Diabetes based on Mixture of Expert Model

Hong-Ki Lee*, Sung-Min Myoung**

## 요 약

당뇨병은 급성합병증을 예방하고 장기간의 합병증의 위험도를 감소하기 위하여 지속적인 치료와 환자 자가 관리 교육이 필요한 만성질환이다. 또한 전 세계적으로 당뇨병에 대한 유병률과 사망률이 대부분의 인구집단에서 역학적 비율에 도달하였다. 많은 연구에서 당뇨병에 대한 조기진단은 적절한 치료와 생활습관을 지키는 관리를 통하여 발병을 예방하는데 도움을 줄 수 있으며, 이를 통하여 당뇨병의 합병증을 감소시키고 생존률을 향상시킬 수 있다고 보고하고 있다. 본 연구는 PIMA Indians 당뇨 데이터에 대하여 mixture of expert 모형을 적용하여 당뇨유병환자의 여부를 분류하고, 이를 로지스틱 회귀분석, 신경망분석의 성능과 비교함으로서 그 유용성을 주장하고자 하였다. 연구결과 정확도 및 ROC 곡선, c-통계량에서 ME 모형이 다른 분류도구들에 비해서 높게 나타남을 확인할 수 있었다.

▸ Keywords : 자동 진단 시스템, 의사결정지원 시스템, 당뇨 진단, mixture of experts

## Abstract

Diabetes is a chronic disease that requires continuous medical care and patient-self management education to prevent acute complications and reduce the risk of long-term complications. The worldwide prevalence and incidence of diabetes mellitus are reached epidemic proportions in most populations. Early detection of diabetes could help to prevent its onset by taking appropriate preventive measures and managing lifestyle. The major objective of this research is to develop an automated decision support system for detection of diabetes using mixture of experts model. The performance of the classification

•제1저자: 이홍기 •교신저자: 명성민
•투고일 : 2014. 10. 13. 심사일 : 2014. 11. 3, 게재확정일: 2014. 11. 13.
* 중원대학교 경영학과(Department of Management, Jungwon University)
** 중원대학교 의료정보행정학과(Department of Medical Information and Administration, Jungwon University)
※ 이 논문은 중원대학교 교내학술연구비 지원에 의한 것임

algorithms was compared on the Pima Indians diabetes dataset. The result of this study demonstrated that the mixture of expert model achieved diagnostic accuracies were higher than the other automated diagnostic systems.

▸ Keywords : automated diagnostic system, decision support system, diabetes diagnosis, mixture of experts,

## I. 서 론

당뇨병(diabetes mellitus)은 고령화 및 생활습관의 서구화에 따라 전 세계적으로 유병률과 사망률이 계속적으로 증가하고 있는 만성 대사 장애 질환이다 [1]. 질병관리본부가 2013년 11월 발표한 '2012 국민건강영양조사' 결과에 의하면, 만 30세 이상 성인의 당뇨병 유병률은 남성 10.1%, 여성 8.0%였으며, 30~40대 당뇨병 유병자 절반이상이 미인지, 비치료 상태인 것으로 보고되었다 [2]. 또한 2013년 통계청에서 발표한 '사망원인통계'에 의하면 한국인의 사망원인 중 5위(23.0%)를 차지하며, 2011년에 비해 1.5% 증가하였으며, 남성의 순위(5위)보다는 여성이 더 높게(4위) 나타났다 [3].

당뇨병은 당뇨 자체의 문제 및 합병증으로 인한 건강문제가 나타날 수 있는데, 대표적인 문제점으로 망막합병증으로 인한 실명, 당뇨병성 신증으로 인한 말기신부전증으로의 진행, 신경합병증으로 인한 하지 절단 등이 있으며, 특히 심혈관 질환으로 인한 사망의 증가 및 의료비 상승으로 직결된다고 알려져 있다 [4]. 따라서 당뇨병에 대한 조기진단을 통한 적절한 치료와 생활습관을 지키는 관리가 이루어지는 것이 중요하며, 이를 통하여 당뇨병의 합병증을 감소시키고 생존률을 향상시킬 수 있다고 많은 연구에서 보고하고 있다 [5-7]. 이러한 의학적 처치의 수행 전 조기진단은 의과학 분야에서 매우 중요한 문제로 알려져 있다. 그러나 환자들에게 많은 검사를 수행하면 할수록 질병을 진단하는 것이 매우 복잡해지며, 또한 정확한 진단을 한다는 것 자체도 임상 전문가들에게 쉬운 문제는 아니다 [8]. 이러한 문제를 해결하기 위한 방법으로 최근 통계적/수학적 모형의 사용이 지속적으로 증가하고 있으며 [9], 이 모형을 이용한 의사결정지원시스템(clinical decision support system; CDSS)은 의료정보학, 의학통계학 및 컴퓨터 공학 연구자들에게 주요한 주제가 되고 있다. 기존 사례로서 Kim et al. [10]은 서포트 벡터 머신(support vector machine)을 이용하여 심전도 신호의 리듬 분류 기법를 제안하였으며, Lee and Kim [11]은 2012년 지역사회건강조사에서 호흡기질환군과 정상군을 대상으로 인공신경망(artificial neural network), 로지스틱 회귀분석(logistic regression), 베이지안 네트워크(Bayesian network), CART 등을 적용하여 위험요인을 규명하고자 하였다. Lee [12]는 209명의 여성을 대상으로 베이지안 네트워크를 이용한 유방암의 예측 성능을 비교하였으며, Kim et al. [4]은 2005년 국민건강영양조사 자료를 대상으로 로지스틱 회귀분석, CART, 신경망 모형등을 적용하여 당뇨환자의 관리요인을 규명하였다.

국외 사례로서, Übeyli [13]는 Mixture of Experts(ME) 모형을 이용한 유방암 진단방법을 제안하였으며, Chen et al. [14]은 서포트 벡터 머신과 선형판별분석(linear discriminant analysis)에 기초한 LFDA_SVM 이라는 결합방법(hybrid method)을 제안하였고, 이를 간염 데이터에 적용하여 그 유용성을 확인하였다. Sakai et al. [15]는 급성 맹장염이라고 의심되는 169명의 환자를 대상으로 9개의 위험요인 변수들을 이용하여 베이지안 네트워크 모형과 신경망분석모형, 로지스틱 회귀모형 등을 적용하고, 이에 대한 정밀도를 확인하였다.

위의 선행연구들을 볼 때, 의사결정지원을 위한 통계적모형의 적용방법은 굉장히 다양하기 때문에 가장 효율적이고 효과적인 기법의 선택이 중요하다고 할 수 있다.

Jacobs et al. [16]에 의해 제안된 mixture of expert(ME) 모형은 복잡한 문제를 최대한 단순하게 분리를 하며, 이렇게 단순화된 해는 합쳐져서 최종적인 해를 산출한다는 분리와 해결의 원리(divide and conquer principle)를 이용하는 방법이다. 이 ME 모형은 목표 결과의 조건부 확률 밀도(conditional probability density)이며, 이는

mixture 추정문제의 학습 및 EM-알고리즘을 통한 mixture 모수들을 추정하는 것과 동일하다. 이 ME 모형은 예전에는 음성인식 및 영상처리기법 등에서 이용되었으며, 최근 마이크로어레이 자료 분석 등 많은 분야에서 응용되고 있다 [17-19].

본 연구는 ME 모형을 기존의 Smith et al. [20]의 연구에서 제시된 PIMA Indians 당뇨 데이터 셋을 기반으로 당뇨유병환자의 여부를 분류하고 기존 모형과의 비교를 통하여 정밀도를 확인하고 또한 임상의사결정에 도움이 되는 타당도를 확인함으로서, 그 유용성을 주장하고자 한다.

본 논문은 다음과 같이 구성된다. 분석에 사용되는 PIMA Indians 당뇨 자료와 ME 모형에 대한 분석방법을 2장에서 설명하고, 3장에서는 ME 모형 및 기존모형과의 분석결과를 제시하며, 4장에서는 분석결과에 대한 고찰과 더불어 5장에서는 결론 및 본 연구의 성과에 대하여 설명한다.

## II. 연구방법

### 1. 연구자료

본 연구에서는 UCI의 당뇨에 대한 PIMA Indian 데이터 셋에서 당뇨의 유병여부를 ME 분류모형을 이용하여 기존 모형과 비교하고자 한다. PIMA Indian 데이터 셋은 미국 국립 당뇨병, 소화기병, 신장병 연구소(national institute of diabetes and digestive and kidney diseases, NIDDK)에서 수집한 자료이며, 신경망기법인 ADAP(adapted neural network) 알고리즘 예측에 최초로 사용되었다 [20]. 최근에는 UCI Machine Learning Repository에 의해 제공되어 많은 연구자들에 의해 기계학습에 사용되고 있다 [21]. PIMA Indians 데이터셋은 768명에 의해 얻어졌고 Table 1과 같이 당뇨여부를 나타내는 변수와 8개의 위험인자(risk factor)들로 구성 되어있다.

8개의 위험인자들은 나이(age), 출산횟수(Number of times pregnant), 이완기 혈압(Diastolic blood pressure), 2시간 구강 혈당 부하검사(Plasma glucose concentration a 2h in an oral glucose tolerance test), 어깨 삼두근의 피부 주름 두께(Triceps skin fold thickness), 2시간 인슐린 양(2-h serum insulin), 체질량지수(Body mass index), 당뇨 직계 가족력(Diabetes pedigree function)으로 고려하였다. 본 연구에서는 자료분석을 위한 전 처리과정(pre-processing)으로, 768개의 관측값들 중 2

표 1. PIMA 데이터 셋 정보
Table 1. PIMA data attribute information

| Variable | Mean±SD | type of variable |
|---|---|---|
| Number of times pregnant | 3.8±3.4 | Continuous |
| Plasma glucose concentration a 2h in an oral glucose tolerance test | 120.9±3.0 | Continuous |
| Diastolic blood pressure (mmHg) | 69.1±19.4 | Continuous |
| Triceps skin fold thickness (mm) | 20.5±16.0 | Continuous |
| 2-h serum insulin ($\mu$U/ml) | 79.8±115.2 | Continuous |
| Body Mass Index | 32.0±7.9 | Continuous |
| Diabetes pedigree function | 0.5±0.3 | Continuous |
| Age (years) | 33.2±11.8 | Continuous |
| Diabetes | - | Dichotomous |

시간 구강혈당 부하검사, 확장기혈압, 삼두근 피부 주름 두께, 체질량 지수 의 4개의 위험인자가 0으로 표기되어 있는 관찰치를 모두 제거하였다. 그 결과 768개의 데이터에서 236개가 제외된 532개의 데이터가 분석대상이 되었다.

### 2. Mixture of Experts 모형

Jacobs et al. [16]에 의해 제안된 ME 모형은 그림 1과 같은 구조이며, 크게 게이팅 네트워크(gating network)와 엑스퍼트 네트워크(expert network)로 구성된다.

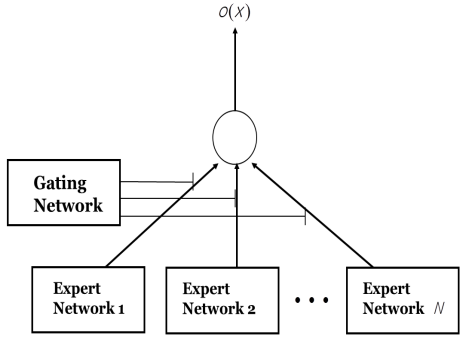


그림 1. ME 모형의 일반화된 구조
Figure 1. General architecture of the ME Model

게이팅 네트워크는 입력벡터 $x$를 기반으로 하여, 입력 공간(input space)에서 각 지점으로 분할하는 스칼라 결과를 생성하며, 각 엑스퍼트 네트워크는 입력 벡터에 대한 출력 벡터를 생성한다. 또한 입력 네트워크는 엑스퍼트 네트워크에 대한 확률로서의 선형결합(linear combination)을 제공하므로, ME 구조의 최종결과는 엑스퍼트 네트워크들에 의해 생성되는 모든 출력벡터들의 가중합(weighted sum)으로 나

타난다. ME구조에서 $N$개의 엑스퍼트 네트워크들이 존재한다고 가정하면, 모든 엑스퍼트 네트워크들은 '일반화 선형'으로 나타나는 비선형성을 가지는 선형모형이다. $i$ 번째 엑스퍼트 네트워크는 입력벡터 $x$에 대하여 다음과 같은 일반화 선형 함수 $o_i(x)$를 생성한다.

$$o_i(x) = f(W_i x) \tag{1}$$

여기에서 $W_i$는 가중치 행렬이며 $f(\bullet)$은 고정된 연속적인 비선형성을 가진다. 입력 네트워크는 또한 일반화 선형모형을 가지며, $i$ 번째 결과 $g(x,\nu_i)$는 매개변수 $\xi_i$ 의 다범주 로짓 모형(multinomial logit model) 혹은 소프맥스 함수(softmax function)로서 아래 식(2)와 같이 정의한다.

$$g(x,\nu_i) = \frac{e^{\xi_i}}{\sum_{k=1}^{N} e^{\xi_k}} \tag{2}$$

여기서 $\xi_i = v_i^T x$ 이며, $\nu_i$는 가중치 벡터이다. ME 모형의 전체 결과 $o(x)$는 다음과 같다.

$$o(x) = \sum_{k=1}^{N} g(x,\nu_k) o_k(x) \tag{3}$$

ME 모형은 확률적 해석을 제공한다. 데이터 쌍 $(x,y)$에 대한 게이팅네트워크 $g(x,\nu_i)$의 값은 $x$에서 $y$로 매핑하는 회귀과정(regressive process)을 종단하는 결정과 관련된 다항 확률로서 해석되어진다. 일단 그 결정이 만들어지면, $i$번째 회귀과정으로 선택지어지는 결과로 나타나게 되어, 출력변수 $y$는 확률밀도함수 $P(y|x, W_i)$에서 선택되어진다. 여기서 $W_i$는 모형에서 $i$번째 엑스퍼트 네트워크의 가중치 행렬 또는 모수들의 집합이라고 할 수 있다. 그러므로, $x$로부터 $y$를 생성하는 전확률(total probability)은 각 성분의 밀도함수로부터 $y$를 생성하는 확률을 가지는 mixture로 나타낼 수 있다.

$$P(y|x,\Phi) = \sum_{k=1}^{N} g(x,\nu_k) P(y|x, W_k) \tag{4}$$

위 식에서 $\Phi$는 엑스퍼트와 게이팅 네트워크 모수들을 모두 포함하는 모수들의 집합이다. 더불어, 모형의 확률적인 성분은 일반적으로 회귀모형의 경우에는 정규분포로 가정하며, 이항분류 문제인 경우에는 베르누이 분포, 다범주 분류인 경우 다항분포로 가정한다 [22]. 위의 식 (4)에서의 확률모형에 기초하여, ME 모형의 학습은 최대우도함수(maximum likelihood function)의 문제로 귀결된다. Jordan and Jacobs는 이 모형의 모수를 구하는 방법으로 EM 알고리즘을 제안하였다 [23]. 즉, 학습 집합(training set)이 $\aleph = \{(x_t, y_t)\}_{t=1}^{T}$ 로 주어진다고 가정한 경우, EM 알고리즘은 아래와 같이 2 단계로 구성될 수 있다.

E-step에서는 $s$번째 반복(epoch)에서, $P(i|x_t,y_t)$의 확률로 해석되는 사후확률 $h_i^{(t)}(i=1,\cdots,N)$을 아래와 같이 계산한다.

$$h_i^{(t)} = \frac{g(x_i,\nu_i^{(s)}) P(y_t|x_t, W_i^{(s)})}{\sum_{k=1}^{N} g(x_t,\nu_k^{(s)}) P(y_t|x_t, W_k^{(s)})} \tag{5}$$

M-step은 다음의 최대화 문제를 해결한다.

$$W_i^{(s+1)} = \arg\max_{W_i} \sum_{t=1}^{T} h_i^{(t)} \log P(y_t|x_t, W_i) \tag{6}$$

$$V^{(s+1)} = \arg\max_{V} \sum_{t=1}^{T} \sum_{k=1}^{N} h_k^{(t)} \log g(x_t,\nu_k) \tag{7}$$

여기서 $V$는 게이팅 네트워크에서의 모든 모수들의 집합을 의미한다. 그러므로 EM 알고리즘은 다음과 같이 요약된다.

1. 각 데이터 쌍 $(x_t,y_t)$에 대하여 모수의 초기치를 이용하여 사후확률 $h_i^{(t)}$를 계산한다.
2. 각 엑스퍼트 네트워크 $i$에 대하여 관찰치 $\{(x_t, y_t)\}_{t=1}^{T}$와 관찰치에 대한 가중치 $\{h_i^{(t)}\}_{t=1}$를 가지고 식 (6)을 최대화 한다.
3. 게이팅 네트워크에 대하여, 관찰치 $\{(x_t, h_k^{(t)})\}_{t=1}^{T}$를 가지고 식 (7)을 최대화 한다.
4. 업데이트된 모수값을 이용하여 반복한다.

## 3. 성능평가 방법

본 절에서는 PIMA Indians 데이터 셋에 대한 기존의 분류방법인 로지스틱 회귀분석, 신경망분석과 본 논문에서 제안한 ME 모형에 의한 분류방법들의 성능평가를 성능평가 측도인 정확도(accuracy), c-통계량(c-statistics) , ROC 곡선(receiver operating characteristics curve)을 가지고 비

교하고자 한다.

만일 $n$개의 소속집단을 알고 있으며, 두 개의 집단 $G_1$과 $G_2$가 에 대한 데이터가 있다면 분류모형을 이용하여 각 데이터를 분류한 후, 모형에 의해 분류된 집단과 실제집단을 비교하여 그 결과를 표 2와 같이 요약할 수 있다.

표 2. 데이터의 실제집단과 분류된 집단의 결과 교차표
Table 2. Cross-classified table bwteen observed group and classified group

| | | Classified Group | |
|---|---|---|---|
| | | $G_1$ | $G_2$ |
| Observed Group | $G_1$ | $f_{11}$ | $f_{12}$ |
| | $G_2$ | $f_{21}$ | $f_{22}$ |

분류모형의 정확도(accuracy)는 다음과 같이 전체 데이터 $n$개 중 올바로 분류된 수 $f_{11}+f_{22}$의 비율이며, 오류율(error rate)은 전체 데이터 중 올바르게 분류되지 않은 수 $f_{12}+f_{21}$의 비율로서 정의하며, 정확도를 최대화 하거나 또는 오류율을 최소화하는 분류모형을 선택하는 것이 일반적이다 [24].

ROC 곡선은 분류모형의 1-특이도(specificity)를 $x$축으로 하고 민감도(sensitivity)를 $y$축으로 한 그래프이다. 민감도를 정분류율(true positive rate)이라 하고, 1-특이도를 오분류율(false positive rate)이라 한다. ROC 그래프의 한 점은 한 분류모형의 결과를 의미하는데 모형에서 기준값(cut-off value)을 변화시켜 가면서 각 기준값에 의해 나타나는 오분류율과 정분류율의 변화를 그래프로 나타낸 것이다. 즉, ROC 곡선은 분류모형의 최종결과인 사후확률 등의 기준값이 변화할 때 민감도와 특이도의 변화를 살펴보는 것이다. ROC 곡선이 좌측상단으로 더 위에 위치할수록 좋은 모형이라고 판정한다 [25]. ROC 곡선 아래의 면적을 c-통계량(c-statistics)이라고 하는데 어떤 모형의 ROC 곡선 아래의 면적이 다른 모형의 면적보다 크면 평균적으로 더 우수한 모형이라 할 수 있다.

## III. 연구결과

제안된 ME 모형의 성능평가를 위해 로지스틱 회귀분석, 인공신경망 모형을 고려하였고 R-package 3.1.1을 이용하였다. 로지스틱 회귀분석은 R-package의 glm() 함수를 이용하였으며, 인공신경망 모형은 nnet library를 [26] 사용하여 구현하였다.

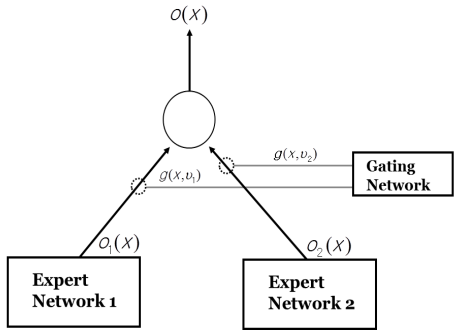


그림 2. PIMA 당뇨 데이터에 대한 ME 모형의 설정
Figure 2. Architecture of the ME model in PIMA dataset

당뇨병 진단을 위해 사용된 ME 아키텍처는 그림 2에 제시하였는데, 게이팅/엑스퍼트 네트워크가 2개로 설정하였다. 각 분류모형들의 분류결과는 표 3과 같은 오분류표(confusion matrix)로서 나타낼 수 있다. 오분류표는 목표변수의 범주별로 이를 제대로 분류한 빈도와 제대로 분류하지 못한 빈도를 함께 제시한 표이다.

표 3. 분류모형별 오분류표
Table 3. Confusion Matrix of the classifiers

| Classifiers | Desired Result | Output Result | |
|---|---|---|---|
| Logistic Regression | Non-diabetics | 270 | 85 |
| | Diabetics | 36 | 141 |
| Neural Network | Non-diabetics | 271 | 84 |
| | Diabetics | 35 | 142 |
| ME Model | Non-diabetics | 291 | 64 |
| | Diabetics | 25 | 152 |

표 4에서는 각 분류방법들의 성능결과를 민감도(sensitivity), 특이도(specificity), 정확도(accuracy)로 제시하였다.

여기서 민감도는 실제 당뇨를 가진 환자수가 분모이며, 분자는 당뇨로 예측한 환자수로 정의한다. 특이도는 당뇨가 아니라고 환자수를 실제 당뇨가 아닌 환자수로 나눈 값이며, 정확도는 제대로 분류한 환자수를 전체 환자수로 나눈 값을 의미한다. 분류방법에 의한 정확도 측면에서 보면 로지스틱 회귀분석은 77.26%이며, 신경망분석은 77.63%로 거의 비슷하게 나타난다. ME 모형의 정확도는 83.27%로서 제시한 3

가지 분류방법들 중 가장 높게 나타났다.

표 4. 분류모형별 성능비교결과
Table 4. Performance comparision of the classifiers

| Classifiers | Classification Accuracies(%) | | |
|---|---|---|---|
| | Sensitivity | Specificity | Accuracy |
| Logistic Regression | 79.66 | 76.06 | 77.26 |
| Neural Network | 80.23 | 76.34 | 77.63 |
| ME Model | 85.88 | 81.97 | 83.27 |

민감도의 경우 ME 모형이 85.88%로서 가장 높게 나타났고, 다음으로 신경망분석 80.23% 순이고 로지스틱 회귀분석이 79.66%로 가장 낮은 값을 보이고 있다. 특이도의 경우에도 민감도에서의 결과와 같이 ME 모형이 가장 높은 값으로 나타났다.

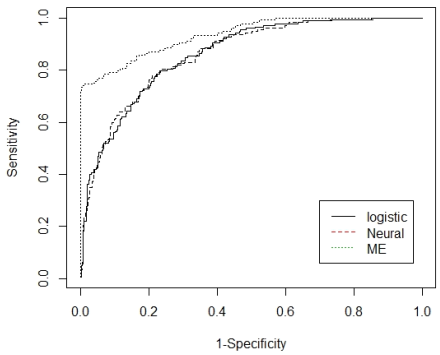


그림 3. 분류모형별 ROC 곡선
Figure 3. ROC Curves of the classifiers

그림 3은 각 분류방법들을 PIMA 데이터 셋에 적용하여 얻은 ROC 곡선들을 나타내고 있다. 로지스틱 회귀분석과 신경망분석의 ROC 곡선들은 거의 중첩되어 있어서 가시적으로 어느 방법이 더 좋다고 판단할 수 없었다. ME 모형은 나머지 곡선들보다 왼쪽 위에 위치하고 있어서 제시한 모형들 중에 가장 좋은 성능을 보인다고 판단할 수 있다. 표 5는 여러 가지 분류방법에 대한 c-통계량이다.

이를 확인해 보면 ME 모형이 가장 높게 나타났고(0.936) 그 다음 로지스틱회귀분석 0.861, 신경망분석 0.860 순으로 나타났다.

표 5. 분류모형별 성능비교결과(c-통계량)
Table 5. Performance comparision of the classifiers(c-statisti cs)

| Classifiers | c-statistics |
|---|---|
| Logistic Regression | 0.861 |
| Neural Network | 0.860 |
| ME Model | 0.936 |

위의 결과들을 종합적으로 보았을 때, 정확도가 가장 높게 나타나는 ME 모형에 대한 모수추정결과는 표 6에 제시하였다. 2개의 엑스퍼트 네트워크를 가지는 모수 추정치들은 다음과 같다. 게이팅 네트워크는 (0.5730, 0.4270)으로 나타났는데, 1번째 엑스퍼트로 할당되는 확률이 57.30%이며, 2번째 엑스퍼트로 할당되는 확률은 42.70%임을 의미한다.

각 엑스퍼트 네트워크들에 하여 추정된 모수들은 각 변수별로 엑스퍼트 네트워크들의 부호 또는 크기가 다르게 나타남을 확인할 수 있었다. 출산횟수의 경우 엑스퍼트 1의 추정치는 2.0266으로서 양수이나, 엑스퍼트 2에서는 -4.7969로 음수로 나타났다. 구강 혈당부하 검사(2시간)의 경우에는 엑스퍼트 1보다 2에서 더 큰 추정치로 나타났으며, 확장기 혈압, 2시간 혈청주사한 인슐린의 양의 경우 엑스퍼트 2는 음수 추정치로, 엑스퍼트 1은 양수 추정치로 나타났다. 반대로 어깨 삼두근의 피부주름두께 및 연령의 경우에는 엑스퍼트 1이 음수추정치, 엑스퍼트 2가 양수추정치로 나타났다. 또한 체질량 지수와 당뇨 직계가족력의 경우에는 두 엑스퍼트 모두 비슷하게 추정되었다.

표 6. ME 모형에서 게이팅/엑스퍼트 네트워크 추정결과
Table 6. Parameter estimation of each gating/expert networks in ME model

| Variable | Expert 1 ($g_1 = 0.57$) | Expert 2 ($g_2 = 0.43$) |
|---|---|---|
| Number of times pregnant | 2.0266 | -4.7969 |
| Plasma glucose concentration a 2h in an oral glucose tolerance test | 0.9214 | 4.8413 |
| Diastolic blood pressure (mmHg) | 0.0368 | -0.9955 |
| Triceps skin fold thickness (mm) | -0.5829 | 0.7695 |
| 2-h serum insulin ($\mu$U/ml) | 1.0410 | -4.8087 |
| Body Mass Index | 1.1910 | 1.4190 |
| Diabetes pedigree function | 0.8357 | 0.4433 |
| Age (years) | -1.0449 | 4.5265 |

## Ⅳ. 고찰

최근 의료정보학 분야에서는 개발된 통계적 모형들을 분석

하여 컴퓨터를 이용한 의학적 의사결정(decision making)을 수행하며, 실제 임상을 통하여 새롭게 수집된 자료로 개발된 방법을 평가하는 방법이 활발하게 진행되고 있다 [27]. 이러한 자동화된 질병의 진단기법은 의과학연구자들과 컴퓨터공학 및 의학통계학자들과의 다학제간 연구로서 수행되어져 오고 있다 [28]. 또한, Brause는 사람이 진단하는 능력이 신경망 진단 시스템보다 더 나쁠 수도 있다고 주장하기도 하였다 [29].

의과학 분야에서 이러한 자동화된 의사결정 지원 시스템으로 많이 제안된 방법은 신경망 분석 및 로지스틱 회귀분석방법으로서, 이를 이용해 제안된 결과를 의사들의 진단결과와 비교하였으며, 또한 여러 분류기법들과도 동시에 비교하였다 [30-32]. 그러나 신경망 분석의 경우 분석진행 속도가 로지스틱 회귀분석에 비해 느리며, 결과 해석과정에서 어려움이 존재하며, 또한 모형 선택과정에서 적절한 연결함수와 결합함수의 선택, hidden unit의 숫자 결정 등이 쉽지 않다는 점이 단점으로 제시되고 있다 [33]. 또한 로지스틱 회귀분석은 모든 데이터 상에서 분류함수를 구함으로서 과적합 하는 경향이 있다고 주장되어지는 상황이다 [34].

본 연구는 당뇨의 진단 및 당뇨의 위험인자를 가지는 개체들에 대한 자동화된 의사결정 지원 시스템을 ME 모형을 이용하여 적용하고, 이에 대한 성능을 평가하기 위해 로지스틱 회귀분석, 신경망분석과 비교하였다. 본 연구에서 적용된 ME 모형은 기존에 제시된 민감도, 특이도, 정분류율 등에서 가장 높은 값(85.88%, 81.97%, 83.27%)을 갖는 것으로 나타났다. 또한 ROC 곡선 및 c-통계량에서도 ME 모형의 성능이 가장 우수한 것으로 나타났다.

이러한 결과에 비추어 보았을 때, 자동화된 진단 시스템을 위한 ME 모형의 성능은 만족스럽다는 결론을 내릴 수 있으며, 이러한 모형을 개발 시 임상연구에서 충분히 사용할 수 있다고 판단된다. 이러한 ME 모형을 이용한 유용성은 많은 선행연구에서 제안되었는데, 인도인을 대상으로 2형 당뇨와 전당뇨의 예측을 위한 컴퓨터 기반 진단도구의 개발 [28], EEG 데이터를 통한 간질환자의 진단 [35], 비침습적 글루코즈 모니터링 시스템에서의 신호처리에서의 적용 [36], 마이크로 어레이 자료에 대한 백혈병환자의 진단예측 [37] 등에서 확인할 수 있었다.

## V. 결론

최근 질병에 대한 진단과 치료중심보다는 예방과 예측이 중요시되고 있다. 따라서 임상자료에 대한 분류모형의 적용 및 이에 대한 예측모형의 제시는 중요한 문제라고 할 수 있다. 본 연구의 목적은 자동화된 진단 시스템 중 ME 모형을 이용하여 당뇨병 여부를 분류하였다. 성능 평가를 위해 UCI의 Pima Indians 데이터에 대해 각 분류모형들에 대한 정확도, ROC 곡선, c-통계량을 가지고 비교분석하였다. 실제 자료에 적용결과 정확도면에서 ME 모형이 가장 높고 다음으로 신경망 분석, 로지스틱 회귀분석으로 나타났고 ROC 곡선에서 ME 모형이 높은 성능을 보였으며 신경망, 로지스틱 회귀분석은 비슷한 성능으로 나타났으며, c-통계량의 경우에는 ME 모형이 가장 높게 나타났다.

이러한 결과는 학습 알고리즘, 네트워크 모수들의 추정 및 특성들이 분류되는 특징 등 여러 요인들 때문이라고 판단되며, 제시된 결과는 임상의사결정에 도움을 줄 수 있는 분류도구의 타당성을 확인할 수 있었다. 향후 연구에는 국민건강영양조사 등 다양한 자료들로부터 양질의 변수를 확보하여 응용가능한 전략적 지식으로 전환할 수 있다면 당뇨관리 사업에 중요한 자료가 될 수 있을 것이라 판단된다.

## 참고문헌


[1] Georg, P., Ludvik, B., "Lipids and diabetes", Journal of Clinical and Basic Cardiology, Vol. 3, No. 3, pp. 159-162, 2000.

[2] Korean National Health and Nutrition Survey, http://knhanes.cdc.go.kr

[3] Statistics Korea, "Annual report on the cause of death statistics: 2012", Statistics Korea, 2013

[4] Kim, Y., Chang, D., Kim, S., Park, I., Kang, S., "A study on factors management of diabetes mellitus using data mining", Journal of academia-industrial technology, Vol. 10, No. 5, pp. 1100-1108, 2009.

[5] Diabetes Control and Complications Trial Research Group, "The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus", New England Journal of Medicine, Vol. 329, pp. 977-986, 1995.

[6] Holman, R., Paul, S., Bethel, M., Matthews, D., Neil, H., "10-year follow-up of intensive glucose control in type 2 diabetes", New England Journal

of Medicine, Vol. 359, No. 15, pp.1577-1589, 2008.

〔7〕 Sokol, M., McGuigan, K., Verbrugge, R., Epstein, R., "Impact of medication adherence on hospitalization risk and healthcare cost", Medical Care, Vol. 43, No. 6, pp. 521-530, 2005.

〔8〕 Hwang, S., Kim, D., Kang, T., Park, G., "Medical diagnosis system of breast cancer using FCM based parallel neural networks", In Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence, pp. 712-719, 2007.

〔9〕 Shortliffe, E., Cimino J., "Biomedical Informatics: computer applications in health care and biomedicine", Springer-Verlag: New York, pp. 76-79, 2013.

〔10〕 Kim, S., Kim, D., "Rhythm classification of ECG signal by rule and SVM based algorithm", Journal of the Korea Society of Computer and Information, Vol. 18, No. 9, 2013.

〔11〕 Lee, J., Kim, H., "Identification of major risk factors association with respiratory diseases by data mining", Journal of the Korean data and information science society, Vol. 25, No. 2, pp. 373-384, 2014.

〔12〕 Lee, S., "Comparisons of predictive modeling techniques for breast cancer in Korean women", Journal of Korean Society of Medical Informatics, Vol. 14, No. 1, pp. 37-44, 2008.

〔13〕 Übeyli, E,. "A mixture of experts network structure for breast cancer diagnosis", Journal of medical systems, Vol. 29, pp. 5569-5579, 2005.

〔14〕 Chen, H., Liu, D., Yang, B., Liu, J., Wang, G., "A new hybrid method based on local fisher discriminant analysis and support vector machines for hepatitis disease diagnosis", Expert systems with applications, Vol. 38, No. 9, pp. 11796-11803, 2011.

〔15〕 Sakai, S., Kobayashi, K., Nakamura, J., Toyabe, S., Akazawa, K., "Accuracy in the diagnostic prediction of acute appendicitis based on the bayesian network model", Methods of information in medicine, Vol. 46, No. 6, pp. 723-726, 2007.

〔16〕 Jacobs, R., Jordan, M., Nowlan, S., Hinton, G., "Adaptive mixture of local experts", Neural compuation, Vol. 3, pp. 179-87, 1991.

〔17〕 Gutta, S., Huang, J., Jonathon, P., Wechsler, H., "Mixture of experts for classification of gender, ethnic origin, and pose of human faces", Neural Networks, IEEE Transactions on, Vol. 11, No. 4, pp. 948-960, 2000.

〔18〕 Lê Cao, K., Meugnier, E., McLachlan, G., "Integrative mixture of experts to combine clinical factors and gene markers", Bioinformatics, Vol. 26, No. 9, pp. 1192-1198, 2010.

〔19〕 Hu, Y., Palreddy, S., Tompkins, W., "A patient-adaptable ECG beat classifier using a mixture of experts approach", Biomedical Engineering, IEEE Transactions on, Vol. 44, No. 9, pp. 891-900, 1997.

〔20〕 Smith, J., Everhart, J., Dickson, W., Knowler, W., Johannes, R., "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus", Proc Annu Symp Comput Appl Med Care, Vol. 9, pp. 261-265, 1988.

〔21〕 UCI Machine Learning Repository. University of California, Center for Machine Learning and Intelligent Systems, http://archive.ics.uci.edu/ml/datasets.html

〔22〕 Hong, X., Harris, C., "A mixture of experts network structure construction algorithm for modelling and control", Applied Intelligence, Vol. 16, pp. 159-169, 2002.

〔23〕 Jordan, M., Jacobs, R., "Hierarchical mixture of experts and the EM algorithm", Neural Computation, Vol. 6, pp. 2181-2214, 1994.

〔24〕 Lee, J., "Data mining using R, SAS, MS-SQL", Free Academy, pp. 172-173, 2011.

〔25〕 Bradley, A., "The use of the area under the ROC curve in the evaluation of machine learning algorithms", Pattern recognition, Vol.

30, No. 7, pp. 1145-1159, 1997.
〔26〕 Venables, W., Ripley, B., "Modern applied statistics with S", Springer: New York, pp. 245-250, 2002.
〔27〕 Übeyli, E., "Comparisons of different classification algorithms in clinical decision-making", Expert System, Vol. 24, pp. 117-131, 2007.
〔28〕 Shankaracharya, Odedra, D., Samanta, S., Vidyarthi, A., "Computational Intelligence-based diagnosis tool for the detection of prediabetes and type 2 diabetes in India", The review of diabetic studies, Vol. 9, No. 1, pp. 55-62, 2012.
〔28〕 Brause, R., "Medical analysis and diagnosis by neural networks", Lecture note in computer science, Vol. 99, pp. 1-13, 2001.
〔29〕 Shanker, M,. "Using neural networks to predict the onset of diabetes mellitus", Journal of chemical information and computer sciences, Vol. 36, pp. 35-41, 1996.
〔30〕 Lim, C., Harrison, R., Kennedy, R., "Application of autonomous neural network system to medical pattern classification tasks", Artificial intelligence in medicine, Vol. 11, No.3, pp. 215-239, 1997.
〔31〕 Übeyli, E., "Combining neural network models for automated diagnostic systems", Journal of medical systems, Vol. 30, pp. 6483-6488, 2006.
〔32〕 Lee, H., Park J., "Probabilistic filtering for a biological knowledge discovery system with text mining and automatic inference", Journal of the Korean data and information science society, Vol. 17, No. 2, pp. 139-147, 2012.
〔33〕 Min, D., "Comparison of neural network modeling and logistic regression based on pattern analysis of customer using SAS E-Miner", Journal of Korean Data Analysis Society, Vol. 9, No. 4, pp. 1861-1873, 2007.
〔34〕 Lim, J., Sohn, J., Sohn, J., Lim, D., "Breast cancer classification using optimal support vector machine", Journal of the Korea Society of Health Informatics and Statistics, Vol. 38, No. 1, pp. 108-121, 2013.
〔35〕 Subasi, A., "EEG signal classification using wavelet feature extraction and mixture of expert model", Expert system and applications, Vol. 32, pp. 1084-1093, 2007.
〔36〕 Kurnik, R., Oliver, J., Waterhouse, S., Dunn, T., Jaylakshmi, Y., Lesho, M., Lopatin, M., Tamada, J., Wei, C., Potts, R., "Application of the mixtures of experts algorithm for signal processing in a noninvasive glucose monitoring system", Sensors and actuators B, Vol. 60, pp. 19-26, 1999.
〔37〕 Corchado, JM., Paz, J., Rodriguez, S., Bajo, J., "Model of experts for decision support in the diagnosis of leukemia patients", Artificial intelligence in medicine, Vol. 46, pp. 179-200, 2009.

## 저 자 소 개

이 홍 기
2002: 경기대학교 경영학 박사
현재: 중원대학교 경영학과 조교수
관심분야: 병원경영학, 리더쉽, 조직관리
Email: it2020@jwu.ac.kr

명 성 민
2006: 연세대학교 의학전산통계학과 의학통계학 박사
현재: 중원대학교 의료정보행정학과 조교수
관심분야: 의학통계학, 데이터마이닝, 통계계산
Email: smmyoung@jwu.ac.kr