Journal of Korean
Data & Information Science Society
2005, Vol. 16, No. 3, pp. 515~527

Data Mining Model Analysis for The Risk Factor of Hypertension
- By Medical Examination of Health Data -

Jea-Young Lee[1)] · Joon Sakong[2)] · Yong-Won Lee[3)]

Abstract

The data mining is a new approach to extract useful information through effective analysis of huge data in numerous fields. We utilized this data mining technique to analyze medical record of 39,900 people. Whole data were separated by gender first and divided into three groups, including normal, stage 1 hypertension, and stage 2 hypertension. The data from each group were analyzed with data mining technique. Based on the result that we have extracted with this data mining technique, major risk factors for the hypertension are age, BMI score, family history.

Keyword : Data Mining, HDL cholesterol, hypertension, Triglyceride

1. 서론

컴퓨터와 네트워크의 엄청난 발전으로 데이터베이스를 만들어 많은 양의 자료를 저장, 보관하게 되었다. 여러 분야에서 저장, 보관되는 대량의 데이터를 효과적으로 분석하여 유용한 정보를 획득하기 위해 새로이 등장한 것이 데이터 마이닝이다. 데이터 마이닝이란 대량의 데이터나 복잡한 구조의 데이터들을 정교한 통계분석과 모델링(Modeling) 테크닉을 이용하여 정확히 식별되지 않는 패턴이나 자료간의 상관관계를 밝혀내어 여러 가지 결과를 예측해 내는 새로운 통계적 기법이다. 이러한 데이터 마

1) First Author : Professor, Department of Mathematics and Statistics, Yeungnam University, 214-1 Daedong Kyungsan Kyungbuk, 712-749, South Korea
E-Mail : jlee@yu.ac.kr
2) Associate Professor. Department of Preventive medicine & Public health. Yeungnam University college of medicine. 317-1 Daemyeung-dong Namgu Daegu, 705-717, South Korea
3) Corresponding Author : Research fellow, Institute of Medical Science, Yeungnam University, 317-1 Daemyeung-dong Namgu Daegu, 705-717, South Korea
E-Mail : antonio@yumail.ac.kr

이닝 기법을 이용하여 제조업의 부도예측에 영향을 주는 주요변수를 선택하거나(최병권, 2004), 웹 페이지 분석을 통한 성별과 관심분야에 대한 웹 페이지 성향 파악이나(Baglion 등, 2003), 두 개의 서로 다른 분포에서 나온 데이터가 섞여있을 때 데이터 마이닝의 신경망 기법을 이용하여 올바른 판별을 할 수 있게 되었다(이성원, 2001). 기술, 경제적인 분야 등 다양한 분야에 활용되고 있는 데이터 마이닝을 의학 분야에 이용하여 고혈압 위험요인들의 정확한 평가를 위하여 성별에 따라 로지스틱 분석을 적용하고, 로지스틱 분석을 통하여 위험요인들의 상대 위험도(Relative Risk)를 구하였다(오희숙 등, 2000). 또한, 한국인에게 비만이 고혈압 발생에 독립적인 위험요인을 규명하였다 (이성희, 2001). 의학에서 로지스틱 모형을 사용하여 자료를 분석하였지만, 본 연구는 대규모 데이터를 바탕으로 데이터 마이닝 분석을 실시하기로 하였다.

2. 연구 내용 및 대상

순환계질환(심혈관 및 뇌혈관질환)은 우리나라 사람의 전체 사망원인 가운데 24.96%를 차지하고 있으며(통계청, 2002), 순환계질환 중 가장 유병률이 높은 고혈압은 뇌혈관질환과 관상동맥질환의 중요한 위험요인으로 알려져 있다.(윤석중 등, 2001, 황은희, 2003) 또한, 고혈압에 의한 사망은 전 세계적으로 연간 사망자 5천 6백만명 중 12%를 차지하며, 뇌혈관계, 심장 그리고 신장에까지 합병증을 초래하는 근본적인 만성질환으로 전 세계적으로 중요한 의학적 관심사이다.(W.H.O., 2002) 고혈압은 진단하기도 쉽고 치료도 어렵지 않지만 무증상으로 지내는 경우가 많아 환자 자신이 심각함을 깨닫지 못하게 되고 그런 상태로 방치되다가 치명적인 합병증을 일으키는 소위 현대 문명병으로서 최선단에 서 있는 질환이라고 할 수 있다.

우리나라에서도 지난 20년 동안 순환계질환으로 인한 사망이 약 10배가 증가하였으며, 30세 이상의 순환계질환 사망자 중 8.3%가 고혈압에 의한 사망으로 나타났다 (통계청, 2002). 하지만, 고혈압 환자의 거의 반수는 본인이 고혈압인 것을 모르고 지내거나 고혈압이 발견되더라도 대수롭게 생각하지 않는 경우가 많아, 고혈압 환자로 진단받은 후 계속 치료받는 수는 20%정도에 불과하며, 65%이상의 환자들이 한두 번 의료기관을 방문하다가 치료를 중단하고 있다고 한다(오희숙 등, 2000; 황은희, 2003).

이러한 고혈압에 대한 주요인을 파악하기 위하여 2003년 1월 1일부터 12월 31일까지 1년간 서울 소재 종합건강검진센터에 건강검진을 받은 20세 이상의 39,900명을 대상으로 분석을 실시하였다. 대상자 전원에서 고혈압 가족력, 과거 고혈압 약 복용여부, 과거 당뇨약 복용여부, 흡연, 음주, 운동습관, 성별, 연령을 면담조사 하고, 신장, 체중, 신체질량지수(BMI, Body Mass Index), 총 콜레스테롤(Total cholesterol), HDL 콜레스테롤 (High-density Lipoprotein Cholesterol), 중성지방, 혈당, 혈압에 관한 측정 자료를 이용하였다.

본 논문에서 종속변수가 되는 고혈압 환자를 1차 고혈압 그룹과 2차 고혈압 그룹으로 분류하였다. 고혈압 그룹의 혈압 분류기준은 Joint National Committee (JNC) 7차 보고서를 기준으로 하였다 (JNC, 2003).

<표 1> 혈압 분류기준

분류	혈압(mmHg)	판정
수축기혈압	> 180	고혈압 제 3기
이완기혈압	> 110	
수축기혈압	160~179	고혈압 제 2기
이완기혈압	100~109	
수축기혈압	140~159	고혈압 제 1기
이완기혈압	90~99	
수축기혈압	130~139	높은 정상
이완기혈압	85~89	
수축기혈압	< 130	정상
이완기혈압	< 85	

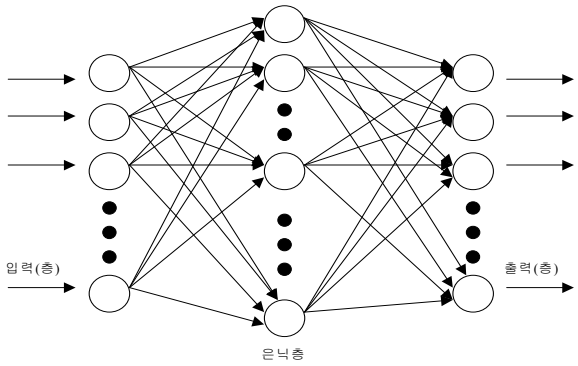
JNC에서 정한 혈압 분류기준과 과거 고혈압 약 복용여부를 바탕으로 1차 고혈압 그룹과 2차 고혈압 그룹으로 분류하였다. 과거 고혈압 약을 복용한 적이 있거나 수축기혈압 140이상, 이완기혈압 90 이상인 사람들을 1차 고혈압으로, 과거 고혈압 약을 복용한 적이 있거나 수축기혈압 160이상, 이완기혈압 100 이상인 사람들을 2차 고혈압으로 분류하였다.

그리고, 독립변수가 되는 고혈압 가족력, 과거 당뇨약 복용여부, 혈당, 흡연, 음주, 운동습관, 나이, BMI, 총 콜레스테롤, HDL 콜레스테롤에 관한 내용은 다음과 같다. 직계가족 중 고혈압 판정을 받은 사람을 고혈압 가족력이 있다고 분류하였고, 과거 당뇨약 복용여부와 혈당의 두 변수를 사용하여 과거 당뇨약을 복용한 적이 있으며 혈당 수치가 126이상인 집단과 과거 당뇨약을 복용한 적은 없지만 혈당 수치가 126이상인 집단을 당뇨그룹으로, 과거 당뇨약을 복용한 적이 없으며 혈당 수치가 126미만인 집단을 정상그룹으로 분류하였다. 흡연은 현재 흡연자, 과거 흡연자, 미 흡연자로 분류하였고, 음주는 즐겨 마시는 그룹, 거의 마시지 않지만 어쩌다 1~2잔정도 마시는 그룹, 전혀 음주하지 않는 그룹으로 분류하였다. 운동습관은 운동을 전혀 하지 않는 그룹과 운동을 하는 그룹으로 분류하였고, 연령은 10대 단위로 범주화 하여 20~29세를 20대, 30~39세를 30대, 40~49세를 40대, 50~59세를 50대, 60~69세를 60대 그리고 70대 이상으로 분류하였고, BMI는 18.5미만을 저체중, 18.5~23미만을 정상체중, 23~25미만을 과체중, 25~30미만을 중도비만, 30이상을 고도비만으로 분류하였다(WHO, 2000). 총 콜레스테롤은 200미만을 정상그룹, 200~239를 경계그룹, 240이상을 위험그룹으로 분류하였고, HDL 콜레스테롤은 46이상을 정상그룹, 35~45를 경계그룹, 35미만을 위험그룹으로 분류하였으며, 중성지방은 150미만을 정상그룹, 150~200을 경계그룹, 201이상을 위험그룹으로 분류하였다.

3. 데이터 마이닝 기법의 배경

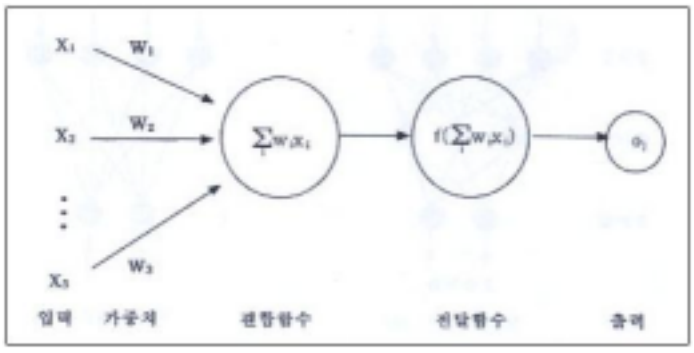
고혈압의 주요인을 분석하기 위하여 데이터 마이닝 기법을 사용한다. 데이터 마이닝은 보통 탐색, 변형, 모형화, 평가의 단계를 거쳐 수행하게 되는데, 이 중 모형화 단계에서 사용되는 기법들에 대한 연구가 활발히 진행되고 있다. 이러한 데이터 마이닝 기법은 기계학습 방법의 개념으로 대표되며, 이 중 가장 유명한 것은 Neural Network 기법과 의사결정 나무분석방법이다.

Neural Network 기법은 인간의 신경-두뇌 시스템을 흉내낸 것으로 <그림 1>과 같이 입력층(input layer), 은닉층(hidden layer), 출력층(output layer) 등 3개의 층으로 구성되며, 입력층에서 보내지는 값을 가중치에 따라 은닉층이 합산하고 이를 활성함수(activation function)에서 변환하여 출력층으로 보내는 구조를 가지고 있다.



<그림 1> Neural Network의 원리

일반적으로 로지스틱 함수와 쌍곡탄젠트 함수(hyperbolic tangent function)가 활성함수로 가장 보편적으로 쓰이며, 활성함수를 사용하여 입력층에서 이루어지는 자료처리 과정의 모습은 <그림 2>와 같다.



<그림 2> Neural Network에서의 자료처리 과정

입력변수 n개, 1개의 은닉층과 은닉노드 m개, 1개의 출력층과 출력노드 c개인 신경망 모형에서 j번째 은닉노드의 값을 H_j라고 한다면, 이것은 다음과 같이 표현된다.

$$H_j = f(b_0 + \sum_{i=1}^{n} w_{ji} \times x_i), \quad j = 1, 2, \cdots, m$$

여기서 x_i는 독립변수, w_{ji}는 j번째 은닉노드와 i번째 독립변수간의 가중치, b_0은 bias, f는 활성함수가 되며, 여기서 활성함수를 로지스틱 함수로 사용한다면 다음과 같다.

$$f(s_j) = \frac{e^s}{1 + e^s} = \frac{1}{1 + e^{-s_j}}$$

따라서, 은닉노드의 값 H_j는 0과 1사이의 값을 취하게 되며, 계산된 H_j는 다시 출력층의 함수로 들어가서 Y_k값을 구하게 된다.

$$Y_k = f_k(b_1 + \sum_{j=1}^{m} w_{kj} H_j), \quad k = 1, 2, \cdots, c$$

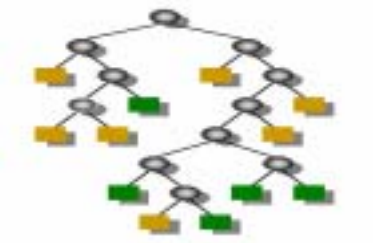
여기서 w_{kj}는 k번째 출력노드와 j번째 은닉노드의 가중치, b_1은 bias, f_k는 활성화/선택함수가 된다. 여기서도 활성함수를 로지스틱 함수로 사용하여 계산된 Y_k는 0~1 값이 되며, 임계치 θ를 기준으로 나뉘어진다.

$$Y_k = \begin{cases} 1, & y_k > \theta \\ 0, & otherwise \end{cases}$$

최적의 임계치를 구하기 위하여 은닉노드와 출력노드의 가중치를 조정하게 된다.

의사결정나무기법은 다양한 알고리즘에 의해 분리가 이루어지며 이런 과정은 나무구조로 표현된다. 이러한 나무구조는 여러 가지 마디(node)라고 불리는 구성요소들로 이루어져 있으며, 나무 구조가 시작되는 뿌리마디, 하나의 마디로부터 분리되어 나간 두 개 이상의 마디들인 자식마디, 자식마디의 상위마디인 부모마디, 각 나무줄기의 끝에 위치하는 끝마디 등 있다. 이러한 마디들의 분리기준은 어떤 입력변수를 사용하고 그 변수의 어떤 값을 기준으로 분리하는 것이 목표변수를 가장 잘 구별할 수 있는지에 초점을 두며, 몇 가지 알고리즘에 의해 분리기준이 정해지게 된다. (허명회 등, 2003)

CART(Classification and Regression Tree)는 설명변수들과 목표변수로 이루어진 자료들에서 설명변수들의 특성에 따라 자료들을 이진분류(binary split)하여, 2개의 하위노드를 생산하는 과정을 반복하여 자료들을 목표변수의 값이 유사한 부분집합으로 만드는 방법으로 <그림 3>과 같이 나타난다.



<그림 3> CART의 분류방식

CART의 알고리즘은 마디의 순수함을 나타내는 지니지수(Gini Index)에 의해 분리 여부를 결정하게 된다. 특정 변수에 의해 집단이 구분되면, 구분된 하나의 집단에서 나머지 집단의 개체가 선택될 확률을 계산하여 집단을 분리하며 집단이 순수할수록 지니지수의 값이 작아지며 확률 또한 작아지게 된다. 지니지수는 다음과 같다.

$$\sum_{i=1}^{r} P(i)(1-P(i))$$

여기서 r은 목표변수의 범주의 수이며, $P(i)$는 주어진 자료 중 i범주에 분류될 확률을 나타낸다.

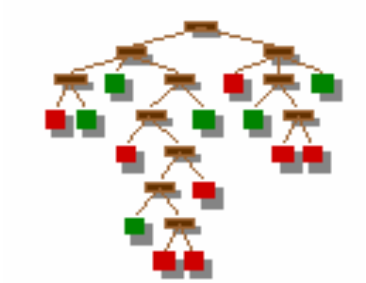
CART는 지니지수를 가장 감소시켜 주는 예측변수와 그 변수의 최적분리를 자식노드로 선택하는데, 지니지수의 감소량은 다음과 같이 계산한다.

$$\Delta G = G - \frac{n_L}{n} G_L - \frac{n_R}{n} G_R$$

여기서 n은 부모노드의 관측치 수, n_L과 n_R은 각각 자식노드의 수를 의미한다. 즉, 자식노드로 분리되었을 때 불순도가 가장 작도록 자식마디를 형성하는 것이며, 이는 다음과 같은 자식마디에서 불순도의 가중합을 최소화하는 것과 동일하다.

$$P(L) G_L + P(R) G_R = \frac{n_L}{n} G_L + \frac{n_R}{n} G_R$$

C5.0 알고리즘은 정보이론(Information Theory)에 따른 엔트로피(Entropy)개념을 이용하여 마디의 정보량에 따른 엔트로피 지수에 의해 분리가 되며 <그림 4>와 같이 나타나게 된다.



<그림 4> C5.0의 분류방식

분리된 마디에서의 집단의 정보량이 부족 할수록 엔트로피 지수는 커지게 되며, 많은 정보량을 가질수록 엔트로피 지수는 작아지게 되어 부족한 것이 없는 완전한 정보를 얻었음을 뜻하며, 엔트로피 지수는 다음과 같다.

$$-\sum_{i=1}^{r} p(i)(log\ (p(i))$$

여기서 r은 목표변수의 범주의 수이며, $p(i)$는 주어진 자료 중 i범주에 분류될 확률을 나타낸다.

만일, 자료 D가 변수 Y에 의하여 k개의 부그룹으로 분할되고, 각 부그룹의 상대적 빈도를 $p_1, p_2, \cdots, p_k$라고 하면, 자료 D의 엔트로피는 다음과 같이 계산된다.

$$info(D) = -\sum_{l=1}^{k} p_l log\ p_l$$

또한, 자료 D가 변수 X_i값을 바탕으로 자료를 분할하여 $D_1, D_2, \cdots, D_n$을 얻는다고 하자. $|D_j|$을 1개의 분할 자료 D의 크기라고 하면, 변수 X_i값을 바탕으로 분할된 $D_1, D_2, \cdots, D_n$의 엔트로피는 다음과 같이 계산된다.

$$infox_i(D) = \sum_{j=1}^{n} \frac{|D_j|}{|D|} info(D_j)$$

일반적으로, 변수 X_i값을 바탕으로 자료 D를 분할하여 나온 정보와 단순히 자료 D를 구별하는데 필요한 정보의 차이가 나게 되는데, 이러한 정보의 차이량을 정보의 이득(Gain)이라 하며,

$$Gain(X_i) = info(D) - infox_i(D)$$

로 정의한다.

$Gain(X_i)$이 큰 순서대로 어떤 변수 X_i가 더 많은 정보를 가지고 있는지를 판단

할 수 있으며, $Gain(X_i)$이 큰 값의 X_i부터 자료 D를 분할하게 된다.

이득(Gain)기준이 이론적으로는 명확하지만, 많은 수의 범주를 갖는 예측변수를 선호하는 편향이 내제되어 있어서 실제로 이득기준을 그대로 사용하지 않는다. 따라서, 이득 비율을 사용하여 실제적인 변수선택의 기준으로 한다.

$$GainRatio(X_i) = \frac{Gain(X_i)}{SplitInfox_i(D)}$$

여기서

$$SplitInfox_i(D) = -\sum_{j=1}^{n} \frac{|D_j|}{|D|} log(\frac{|D_j|}{|D|})$$

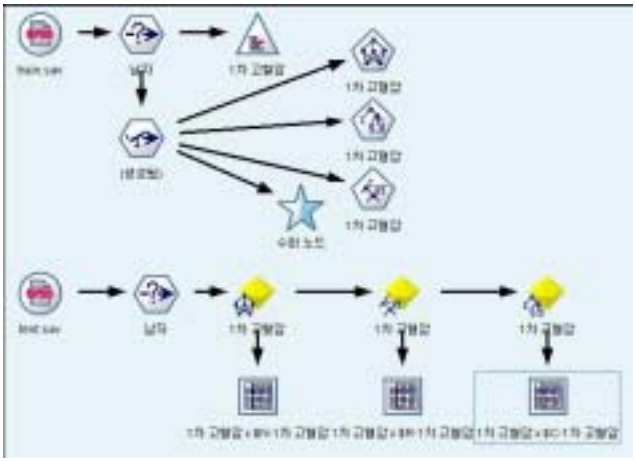
로 정의된다.

$SplitInfox_i(D)$는 자료 D를 단순히 $|D_1|,\cdot\cdot,|D_n|$에 비례하게 임의 분할하였을 때의 엔트로피를 나타낸다. 따라서, $GainRatio(X_i)$는 자료 D를 X_i로 분할함으로써 발생한 이득의 상대량을 의미한다.

간단히 요약하면 $Gain(X_i)$는 절대 비교, $GainRatio(X_i)$는 상대비교를 나타낸다. C5.0은 $GainRatio$를 이용하여, $GainRatio(X_i)$가 최대화되는 점에서 데이터의 분할을 선택한다. (허명회 등, 2003)

3. 고혈압의 주요인에 대한 데이터 마이닝 분석

2003년 1월 1일부터 12월 31일까지 1년간 서울 소재 종합건강검진센터에 건강검진을 받은 20대 이상의 35,671명의 건강검진자료를 전체 데이터로 이용하였다. 종속변수를 고혈압 유무로 하고, 독립변수로 고혈압 가족력, 당뇨병, 흡연, 음주, 운동습관, 나이, BMI, 총 콜레스테롤, HDL 콜레스테롤을 이용하였다.



<그림 5> Data Mining Stream

<그림 5>와 같이 공정한 모형 평가를 위하여 전체 데이터를 각각 반으로 나누어 만든 Train Data와 Test Data를 이용하여 클레멘타인 프로그램을 사용하여 데이터 마이닝을 실시하였다. Train Data를 이용하여 각각의 분석방법으로 만들어진 모형을 Test Data에 적용하여 공정한 모형 평가를 실시하였으며, 고혈압 집단의 비율차이에 대하여 균형을 맞추어 데이터 마이닝 기법 중 Neural Network 기법과 의사결정 나무 분석을 이용하였다 (허명회 등, 2003).

먼저 남성 1, 2차 고혈압에 대한 총 정확도를 살펴본 결과 다음 <표 2>와 같다.

<표 2> 데이터 마이닝 기법들의 총 정확도

	1차		2차		비고
	Train	Test	Train	Test	
Neural Network	**67.604**	**66.190**	**74.221**	**69.67**	← 오분류 측면
CART	66.852	65.079	70.338	70.086	
C5.0	74.798	65.113	84.624	66.602	

<표 2>에 따르면, 남자 1차 고혈압에서 Train Data에서 C5.0의 총 정확도가 가장 높게 나타났지만, Test Data에서는 Neural Network의 총 정확도가 높다는 것을 알 수 있다. 남자 2차 고혈압에서 Train Data에서 C5.0의 총 정확도가 가장 높게 나타났으며, Test Data에서는 CART의 총 정확도가 높다는 것을 알 수 있다. 그러나 오분류 측면에서 살펴보면 1, 2차 고혈압에서 Neural Network의 오분류 비율이 가장 낮게 나타나, Neural Network을 이용하여 남성 1, 2차 고혈압의 주요인에 대하여 살펴보았다.

<그림 6> 남자 1차 고혈압 Neural Network의 중요도

<그림 7> 남자 2차 고혈압 Neural Network의 중요도

Neural Network 기법을 통하여 남성 1, 2차 고혈압의 주요인을 판별 해 본 결과, 1, 2차 고혈압 모두 "**나이 > BMI > 고혈압 가족력**"이 다른 요인들보다 높은 중요도를

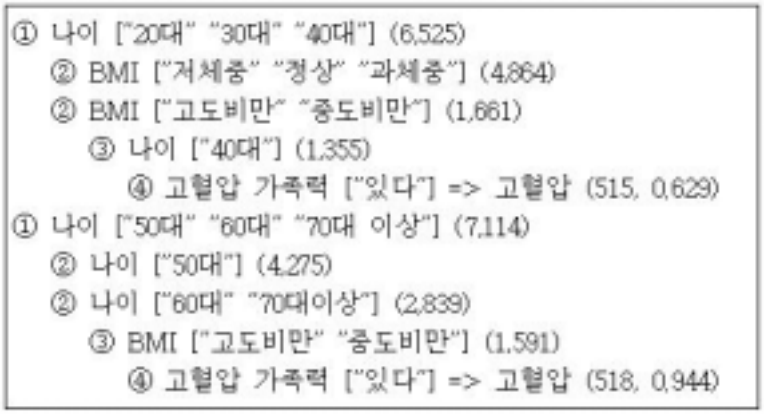
차지하고 있는 것을 알 수 있다.

다음으로 여성의 1, 2차 고혈압에 대한 정확도은 다음 <표 3>과 같다.

<표 3> 데이터 마이닝 기법들의 총 정확도

	1차		2차		비고
	Train	Test	Train	Test	
Neural Network	76.526	71.033	79.932	72.468	
CART	**74.12**	**73.625**	**76.007**	**75.188**	← 오분류 측면
C5.0	80.241	71.533	83.745	69.883	

<표 3>에 따르면, 여자 1차 고혈압에서 Train Data에서 C5.0의 총 정확도가 가장 높게 나타났으며, Test Data에서는 CART의 총 정확도가 높다는 것을 알 수 있다. 여자 2차 고혈압에서 Train Data에서 C5.0의 총 정확도가 가장 높게 나타났으며, Test Data에서는 CART의 총 정확도가 높다는 것을 알 수 있다. 그러나 오분류 측면에서 살펴보면 1, 2차 고혈압에서 CART의 오분류 비율이 가장 낮게 나타나, CART을 이용하여 여성 1, 2차 고혈압의 주요인에 대하여 살펴보았다.



<그림 8> 여성 1차 고혈압에 대한 의사결정 나무

<그림 8>에서 여성 1차 고혈압의 주요인을 판별 해 본 결과, "**나이 > BMI > 고혈압 가족력**"이 다른 요인들보다 높은 중요도를 차지하고 있는 것을 알 수 있다. 먼저 나이가 50대 미만과 50대 이상으로 나누어진다. 50대 미만에서 BMI가 중도비만이상일 경우 나이가 40대이면서 고혈압 가족력이 있으면 1차 고혈압으로 판정이 되며, 50대 이상에서 다시 나이가 60대 이상이며 BMI가 중도비만 이상인 경우 고혈압 가족력이 있다면 1차 고혈압으로 판정이 된다.

<그림 9> 여성 2차 고혈압에 대한 의사결정 나무

<그림 9>에서 여성 2차 고혈압의 주요인을 판별 해 본 결과, "**나이 > BMI > 고혈압 가족력**"이 다른 요인들보다 높은 중요도를 차지하고 있는 것을 알 수 있다. 먼저 나이가 50대 미만과 50대 이상으로 나누어진다. 50대 미만에서 BMI가 중도비만이상일 경우 고혈압 가족력이 있으면 2차 고혈압으로 판정이 되며, 50대 이상에서 다시 나이가 60대 이상이며 BMI가 중도비만 이상인 경우 고혈압 가족력이 있다면 2차 고혈압으로 판정이 된다.

4. 결론

2003년 1월 1일부터 12월 31일까지 1년간 서울 소재 건강검진센터에 건강검진을 받은 20세 이상 39,900명의 건강검진자료를 바탕으로 데이터 마이닝기법을 이용하여 성별에 따른 고혈압 발생 위험요인에 대하여 분석하였다. 공정한 모형 평가를 위하여 전체 데이터를 Train Data와 Test Data로 나누고, Train Data에 데이터 마이닝기법을 이용하여 만들어진 모형을 Test Data에 적용하여 평가하였다.

남성인 경우 1, 2차 고혈압 모두 Neural Network기법이 다른 기법들보다 정확도가 더 높게 나타났다. Neural Network기법을 이용하여 선별된 1차 고혈압 위험요인 중 나이, BMI, 고혈압 가족력이 가장 중요한 위험요인으로 나타났으며, 기타 위험요인 중에서 당뇨, 중성지방이 다른 요인들보다 더 영향을 준다는 것을 알 수 있다. 2차 고혈압 위험요인 역시 1차 고혈압과 동일하게 나이, BMI, 고혈압 가족력이 가장 중요한 위험요인으로 선별되었으며, 기타 위험요인 중에서 HDL 콜레스테롤, 중성지방이 다른 요인들보다 조금 더 영향을 준다는 것을 알 수 있다. 여성인 경우 1, 2차 고혈압에서 CART기법이 다른 기법들 보다 정확도가 더 높게 나타났다. CART기법을 이용하여 선별된 1차 고혈압 위험요인 중 나이, BMI, 고혈압 가족력이 가장 중요한 위험요인으로 선별 되었으며, 기타 위험요인 중에서 당뇨, 중성지방이 다른 요인들보다 조금 더 영향을 준다는 것을 알 수 있다. 2차 고혈압 위험요인에서도 동일하게 나이, BMI, 고혈압 가족력이 가장 중요한 위험요인으로 선별 되었으며, 기타 위험요인 중에서 당뇨병, 중성지방이 다른 요인들보다 조금 더 영향을 준다는 것을 알 수 있다.

종합적으로 볼 때, 성별에 관계없이 **나이, BMI, 고혈압 가족력**이 고혈압에서 가장 중요한 위험요인으로 선별됨을 알 수 있다. 특히 나이와 BMI는 남성보다 여성인 경우에 그 위험도가 훨씬 크며, 고혈압 가족력은 여성보다 남성인 경우에 그 위험도가 크다고 할 수 있다. 또한, 잠재적 위험요인으로 볼 수 있는 당뇨병, 중성지방, HDL 콜레스테롤 역시 그 위험도가 상당히 크다는 것을 알 수 있다. 그러므로 고혈압에 대한 조기 예방 및 치료를 위해서는 성별에 관계없이 우선적으로 BMI를 감소시키며, 더불어 당뇨병을 예방하며 중성지방의 수치 또한 감소시키면서 HDL 콜레스테롤의 수치를 높일 수 있도록 해야 할 것이다. 본 연구에서는 대규모 데이터를 바탕으로 데이터 마이닝 기법을 활용하였는데, 데이터 마이닝 기법은 확률적 접근이기 때문에 충분치 못한 데이터인 경우 오류가 점차 커질 수 있다. 또한, 일반적으로 종합건강검진 설문지에서 흡연, 음주, 운동습관에 관하여 응답자의 대부분이 현재 상태를 응답하게 된다. 하지만, 정확한 분석을 위해서는 설문내용을 자신이 고혈압 판정을 받기 전의 흡연, 음주, 운동습관에 대한 응답이 이루어질 수 있도록 수정하여야 할 것이다. 식생활에 관련된 문항 역시 고혈압 판정을 받기 전의 상태에 대한 응답이 이루어질 수 있도록 하며, 사회심리학적 요인에 관한 응답이 이루어질 수 있도록 수정하여야 할 것이다.

참고 문헌

1. JNC. (2003). The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure : *The JNC 7 Report*. JAMA.
2. M. Baglion, U. Ferrara, A. Romei, S. Ruggier, F. Turini. (2003). Preprocessing and Mining Web log Data for Web Personalization *Proc. of 8th Natl' Conf. of the Italian Association for Artificial Intelligence (AI*IA 2003)*, Paris (to be held September 2003), Italy.
3. WHO. (2000). West Pacific Region. The Asia-Pacific Perspective: *Refining Obesity and its Treatment*. IOTF. Feb.
4. WHO. (2002). Reducing Risks and Promoting Life. *World Health Report*
5. 오희숙, 천병렬, 감신, 예민혜, 강윤식, 김건엽, 이영숙, 박기수, 손재희, 이상원, 안문영 (2000). 농촌지역 주민들의 고혈압 발생 위험요인:1년간 전향적 추적 조사. *예방의학회지* **33,** (2), 231-238
6. 윤석준, 하범만, 김창엽 (2001). 장애보정생존년수(DALY)를 활용한 우리나라 고혈압의 질병부담 측정. *보건행정학회지* **11,** (1), 89-101
7. 이성원. (2001). Logistic modelling for receiver operation characteristic curves with neural networks, Ph.D, 영남대학교
8. 이성희. (2001). 비만이 고혈압 발생에 미치는 영향에 관한 후향적 코호트 연구. Ph.D 서울대학교.
9. 최병권. (2004). 데이터 마이닝 기법을 이용한 제조업 부도예측 주요 변수 선택, 서울대학교
10. 통계청, http://www.nso.go.kr/

11. 허명회, 이용구. (2003). *데이터 마이닝 모델링과 사례*, SPSS 아카데미 p.29, 144-178
12. 황은희. (2003). 고혈압으로 인한 질병악화의 위험요인 및 관리양상에 관한 연구. 강원대학교.

[2005년 4월 접수, 2005년 6월 채택]