Journal of the Korean Data & Information Science Society
2014, **25**(2), 373–384

http://dx.doi.org/10.7465/jkdi.2014.25.2.373
한국데이터정보과학회지

데이터마이닝 모형을 활용한 호흡기질환의 주요인 선별

이제영[1] · 김현지[2]

[12]영남대학교 통계학과

접수 2014년 2월 12일, 수정 2014년 3월 5일, 게재확정 2014년 3월 10일

요 약

데이터 마이닝이란 대량의 데이터나 복잡한 구조의 데이터들을 정교한 통계분석과 모델링 테크닉을 이용하여 정확히 식별되지 않는 패턴이나 자료간의 상관관계를 밝혀내어 여러 가지 결과를 예측해 내는 통계적 기법이다. 이러한 데이터 마이닝 기법은 금융, 통신, 유통, 의학 등 다양한 분야에 활용되는데, 본 연구에서는 의학 분야에 적용하여 호흡기질환에 영향을 끼치는 요인을 선별하였다. 분석은 2012년도 경상북도 지역사회건강조사에 참여한 사람 중 의사에게서 폐결핵, 천식, 알레르기성 비염을 진단받은 경험이 있는 호흡기질환군과 건강군으로 정리한 자료를 대상으로 하였다. 호흡기질환이 영향을 끼치는 주요인을 선별하기 위해 인공신경망, 로지스틱 회귀모형, 베이지안 네트워크, C5.0, CART 기법을 이용하였다. 공정한 모형 평가를 위해 전체 데이터를 훈련용 데이터와 검증용 데이터로 나누었고, 훈련용 데이터에서 설정된 모형을 검증용 데이터에 적용하여 정확도를 비교하였다. 그 결과 CART가 최적 모형으로 선정되었으며 CART의 의사결정나무를 통하여 우울감 인지 여부, 현재 흡연여부, 스트레스 인지 여부 순으로 호흡기질환에 영향을 주는 것으로 나타났다. 그리고 호흡기질환의 주요인들에 대한 오즈비를 구하여 개별적인 영향력에 대해서도 밝혔다.

주요용어: 데이터마이닝, 의사결정나무, 호흡기질환.

1. 서론

폐결핵 (tuberculosis), 천식 (asthma), 알레르기성 비염 (allergic rhinitis) 등의 호흡기 질환은 만성적인 질환이다. 최근 우리나라는 산업이 발달하고 도시화 비율이 높아지면서 만성 호흡기 질환의 발생률뿐만 아니라 호흡기 질환에 의한 사망률도 계속 증가하고 있다 (Han 등, 2005). 통계청 자료 (2005∼2012)를 통해 우리나라 원인별 사망률을 살펴보면 폐결핵 및 호흡기계통의 질환 등 호흡기 질환에 의한 사망률이 인구 10만 명당 2005년도 37.1명, 2008년도 39.4명, 2010년도 45.2명, 2012년도 54명으로 증가하는 추세를 보이고 있다. 호흡기질환은 우리나라뿐만 아니라 전 세계적으로도 증가하고 있는 추세이며, 특히 선진국에서 유병률이 높다. 따라서 만성 호흡기질환, 특히 만성 폐쇄성 폐질환 (chronic obstructive pulmonary disease)과 기관지 천식, 폐암 (lung cancer) 등에 대한 관심이 증가하고 있다. 이러한 관심이 증가함에 따라 인터넷에서 호흡기질환에 관련된 정보들을 쉽게 수집할 수 있게 되었으나 이러한 정보들에는 과학적인 근거가 없는 정보들도 무분별하게 포함되어 있어서 피해 사례 또한 증가하고 있다.

[1] 교신저자: (712-749) 경북 경산시 대동 214-1, 영남대학교 통계학과, 교수. E-mail: jlee@yu.ac.kr
[2] (712-749) 경북 경산시 대동 214-1, 영남대학교 통계학과, 석사과정.

만성 호흡기질환은 악화와 완화가 반복되는 질환으로 지속적인 관리가 필요하며, 완치될 수 있는 질환이 아니므로 약물 사용과 자가 조절, 악화 인자 회피 등의 적절하고 종합적인 관리가 필요하다 (Ryu 등, 2007; Oraka 등, 2010; Kim 등, 2013). 그러나 경제협력개발기구 (OECD)의 한국 의료의 질 검토 보고서에 따르면 한국은 OECD 국가에 비해 천식 등 만성질환 관리가 취약하다고 하였다 (OECD, 2012; Kim 등, 2013). 따라서 본 연구에서는 만성적인 호흡기질환의 바람직한 관리를 위해서 호흡기질환에 영향을 미치는 여러 가지 요인들 중 가장 주요한 요인을 찾을 것이다.

본 연구는 2012년 8월 16일부터 10월 31일까지 2012년 7월 기준으로 시·군·구에 거주하는 만 19세 이상의 성인을 대상으로 하는 2012년도 지역사회건강조사 중 경상북도 조사 참여자 22,304명의 자료를, 의사에게서 폐결핵, 천식, 알레르기성 비염을 진단받은 경험이 있는 참여자를 호흡기질환군으로, 만성질환을 진단받은 경험이 없는 참여자를 건강군으로 분류하여 이용하였다. 이 자료를 데이터마이닝 기법인 인공신경망 (neural network), 로지스틱 회귀모형 (logistic regression), 베이지안 네트워크 (Bayesian network), C5.0, CART (classification and regression tree)에 적용하여 호흡기질환의 주요인을 찾을 것이다. 여기서 데이터마이닝이란 대용량 자료로부터 의미있는 패턴과 규칙을 발견하기 위해서 자동화되거나 반자동화된 도구를 이용하여 데이터를 탐색하고 분석하는 과정이다 (Berry와 Linoff, 1997, 2011; Kim과 Kim, 2013).

본 연구는 다음과 같이 구성되었다. 2절에서는 데이터마이닝 기법을 소개하고 3절에서는 호흡기질환 자료에 소개한 기법들을 적용하여 정확도를 비교하여 최종모형을 선택한다. 4절에서는 선택된 모형을 이용해 호흡기질환에 영향을 미치는 주요인을 선별하고 그 요인의 오즈비 (odds ratio)를 통해 가장 높은 위험인자를 규명한다. 5절에서는 연구의 결과를 요약한다.

2. 데이터마이닝 기법 소개

2장에서는 호흡기질환에 적용할 다양한 데이터마이닝 기법들에 대해서 소개한다. 2.1절에서는 인공신경망 기법, 2.2절에서는 로지스틱 회귀모형, 2.3절에서는 베이지안 네트워크, 2.4절에서는 C5.0, 마지막으로 2.5절에서는 CART기법을 소개한다.

2.1. 인공신경망 기법 소개

인공신경망에 대한 연구는 생물의 신경 체계를 따라하려는 시도에 의해 시작되었다. 인간의 뇌는 뉴런 (neuron)이라 불리는 신경 세포와 뉴런을 연결하는 축색돌기와 수지상돌기로 구성되어 있으며, 그 연결점을 시냅스 (synapse)라고 부른다 (Sarle, 1994; Tan 등, 2006). 신경정신과 의사들은 인간의 두뇌가 같은 충동이 반복적으로 자극되면 뉴런사이의 시냅스 연결의 강도를 변경함으로써 학습하는 것을 발견하였다. 인공신경망은 이러한 인간의 신경-두뇌 시스템을 흉내 낸 것으로, 몇 개의 뉴런과 이것들이 배열된 층 (layer)으로 구성된다 (Heo와 Lee, 2008; Park 등, 2011). 각 뉴런은 특정의 작업을 수행하고 신경망은 이들 뉴런을 연결함으로써 자극과 반응간의 관계를 학습하고, 새로운 데이터에 대한 분류·추정 및 예측을 하게 된다. 여기서 분류는 목표변수가 이산형인 경우가 되며, 예측은 목표변수가 연속형인 경우를 일컫는다.

Figure 2.1은 신경망을 표현한 것인데, 입력 층 (input layer), 은닉 층 (hidden layer), 출력 층 (output layer) 등 3개 층으로 구성되어 있고 각 층에 몇 개씩의 뉴런이 들어있다. 원으로 표시된 것을 노드라고하며, 입력 층에서 입력노드가 자극을 접수하면 은닉 층의 은닉노드가 각 입력노드로부터 전달되는 신호들을 모아 선형결합을 한다. 그리고 이 신호를 최종적으로 출력 층의 출력노드가 전달받아 결합함으로써 최종 반응을 내보내게 된다. 신경망은 특히 복잡한 비선형에 적합 시 좋은 결과를 얻을 수 있다.

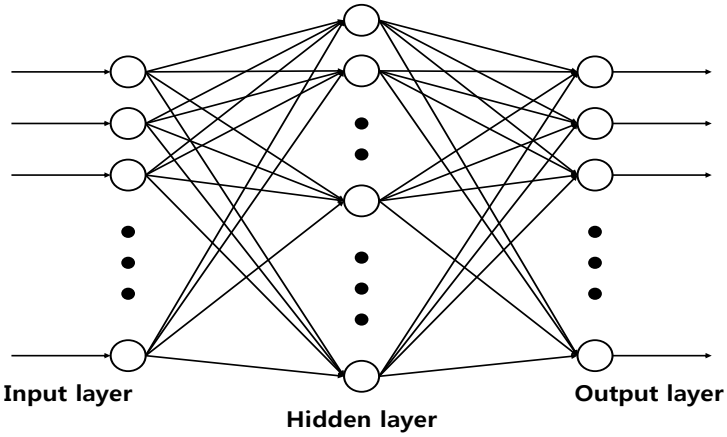

Figure 2.1 Structure of neural network

2.2. 로지스틱 회귀모형 기법 소개

로지스틱 회귀모형은 병리학 연구에서 질병과 연관 있는 위험요인들을 식별하거나, 임상 연구 자료에서 중요한 요인들을 식별하는 탐색적 분석에 많이 적용된다 (Berson 등, 2000; Lee 등, 2005; Heo와 Lee, 2008). 로지스틱 회귀모형에서 종속변수는 주로 이항반응이지만 여러 개의 수준을 갖는 다항 반응인 경우도 있다. 이항자료가 범주형 자료의 가장 일반적인 형태이며, 본 연구도 이항자료를 이용하므로 이항반응에 대해서만 살펴보기로 한다. 이항반응변수를 종종 베르누이 변수라고도 한다. 이 변수에 대한 분포는 성공에 대한 확률 $P(Y=0|x)=1-p_x$와 실패에 대한 확률 p_x로 명시된다. 여기서 성공확률 p_x에 대해 다음과 같은 선형 확률모형을 생각할 수 있다.

$$p_x = \alpha + \beta x \tag{2.1}$$

하지만 위 모형은 구조적인 결함을 가진다. p_x가 0과 1사이의 값을 가지고 $\alpha+\beta x$는 실수 전체의 값을 가지므로 성공확률 p_x가 범위 외의 값을 가질 수 있게 된다. 따라서 이런 식 대신에 성공확률 p_x에 대해 k개의 설명변수 $X_1, X_2, \cdots, X_k$와 비선형인 아래의 식과 같은 함수를 생각할 수 있다.

$$p_x = \frac{exp(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k)}{1+exp(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k)} \tag{2.2}$$

위 식에서 성공확률에 대한 오즈(odds)는 아래 식으로 나타낼 수 있다.

$$\frac{p_x}{1-p_x} = exp(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k) \tag{2.3}$$

성공확률에 대한 오즈에 로그를 씌운 로그 오즈는 아래의 선형 식으로 타나낼 수 있다.

$$\log\frac{p_x}{1-p_x} = \alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k \tag{2.4}$$

위 식을 로지스틱 회귀모형이라고 한다.

2.3. 베이지안 네트워크 기법 소개

베이지안 네트워크는 확률 변수 집합사이의 확률적 관계를 네트워크로 표현하는 방법이다 (Tan 등, 2006). 베이지안 네트워크는 두 가지 요소가 있다. 첫째는 변수 집합 사이의 종속성 관계를 표현하는 방향성 비순환 그래프 (directed acyclic graph; DAG)로 표현하는 것이고 둘째는 각 노드를 그 부모 노드들과 연관시키는 확률표로 표현하는 것이다.

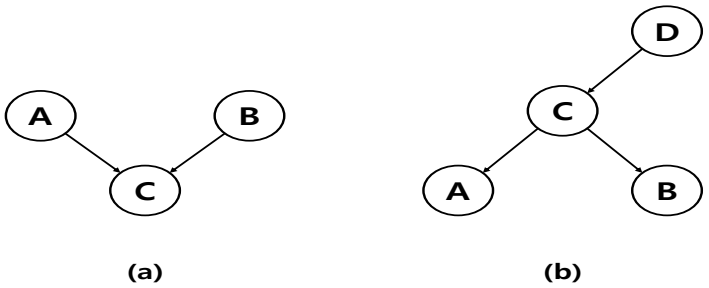

Figure 2.2 A directed acyclic graph (Tan *et al.*, 2006)

Figure 2.2의 (a)는 확률변수 A, B, C의 관계를 네트워크로 표현한 것이다. A와 B는 독립변수이고 각각은 변수 C에게 직접적인 영향을 가지고 있다. 그래프에서 노드는 변수, 화살표는 변수 쌍 사이의 종속관계를 보여준다. Figure 2.2의 (b)를 살펴보면 C는 A와 B의 부모노드이고 A와 B는 C의 자식노드가 된다. D는 A와 B의 조상노드가 된다. 베이지안 네트워크의 한 노드가 부모노드가 알려져 있다면 그 노드는 자손이 아닌 노드들과 조건부 독립적이다. 그러므로 노드 B와 D가 노드 A의 자손이 아니기 때문에 A는 B와 C에 조건부 독립적이라고 할 수 있다.

일반적으로 베이지안 네트워크는 다른 노드들의 확률 값들을 기초로 특정 노드가 가질 값에 대한 조건부확률을 계산하는데 이용할 수 있다. 따라서 하나의 베이지안 네트워크는 한 개체의 다른 속성들의 값이 주어졌을 때 분류 클래스 노드의 사후 확률 분포를 구해줌으로써 개체들에 대한 하나의 자동 분류기로 이용될 수 있다. 즉, 베이지안 네트워크를 기초로 분류 클래스를 확률적으로 예측할 수 있다. 베이지안 네트워크는 조건부 확률 계산에 베이즈 정리를 이용하여 다음의 수식을 얻을 수 있다.

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{2.5}$$

베이즈 정리는 사후 확률을 구할 때 즉, 예측문제를 해결할 때 사용될 수 있다. 베이지안 네트워크는 화살표로 연결된 노드들과는 서로 의존적이고 자손이 아닌 노드들과는 조건부 독립적이라는 속성을 가지고 있다. 이것을 아래의 식으로 표현할 수 있다. 먼저 조건부 확률의 성질에 의해서 다음이 성립한다.

$$P(A|B) = \frac{P(A,B)}{P(B)} \tag{2.6}$$

곱셈규칙에 의해서 다음 식이 성립한다.

$$P(A,B) = P(A|B)P(B) = P(B|A)P(A) \tag{2.7}$$

결과적으로 체인룰을 사용하면, 아래의 식이 만들어진다.

$$P(X_1, X_2, \cdots, X_n) = P(X_1)\prod_{i=2}^{n} P(X_i|X_1, X_2, \cdots, X_{i-1}) \tag{2.8}$$

베이지안 네트워크는 가정의 단순함에도 불구하고 많은 연구를 통해 비교적 높은 분류 성능을 보여주는 것으로 알려져 있다.

2.4. C5.0 기법소개

C5.0 알고리즘은 정보이론 (information theory)에 따른 엔트로피 (entropy)개념을 이용하여 마디의 정보량에 따른 엔트로피 지수에 의해 분리가 된다. 분리된 마디에서의 집단의 정보량이 부족 할수록 엔트로피 지수는 커지게 되며, 많은 정보량을 가질수록 엔트로피 지수는 작아지게 되어 부족한 것이 없는 완전한 정보를 얻었음을 뜻한다 (Freund와 Mason, 1999; Berson 등, 2000).

만일 자료 D가 목표변수에 의하여 범주가 k개로 분할되고, i번째 범주에 분류될 확률이 $p_1, p_2, \cdots, p_k$일 때, 자료 D의 엔트로피 지수는 아래의 식과 같이 정의된다 (Quinlan, 1993; Heo와 Lee, 2008).

$$info(D) = -\sum_{i=1}^{k} p_i(\log p_i) \tag{2.9}$$

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.

$$info\ x_i(D) = \sum_{j=1}^{n} \frac{|D_j|}{|D|} info(D_j) \tag{2.10}$$

일반적으로, 변수 값을 바탕으로 자료 를 분할하여 나온 정보와 단순히 자료 를 구별하는데 필요한 정보의 차이가 나게 되는데, 이러한 정보의 차이량을 정보의 이득 (gain)이라고 한다. 이득의 기준이 이론적으로는 명확하나 많은 수의 범주를 갖는 예측변수를 선호하므로 실제로 이득기준을 그대로 사용하지는 않는다. 따라서 이득비율을 사용하여 실제적인 변수선택의 기준으로 한다.

간단히 요약하면 $Gain(x_i)$는 절대 비교, $Gain\ Ratio(x_i)$는 상대비교를 나타낸다.

$$Gain(x_i) = info(D) - info\ x_i(D) \tag{2.11}$$

$$Gain\ Ratio(x_i) = \frac{Gain(x_i)}{Split\ info_i(D)} \tag{2.12}$$

여기서,

$$Split\ info_i(D) = -\sum_{j=1}^{n} \frac{|D_j|}{|D|} \log\left(\frac{|D_j|}{|D|}\right) \tag{2.13}$$

로 정의된다.

$Gain(x_i)$이 큰 순서대로 어떤 변수 x_i가 더 많은 정보를 가지고 있는지를 판단할 수 있으며, $Gain(x_i)$이 큰 값의 x_i부터 자료 D를 분할하게 된다. $Split\ info_i(D)$는 자료 D를 단순히 $|D_1|, |D_2|, \cdots, |D_n|$에 비례하게 임의 분할하였을 때의 엔트로피를 나타낸다. 따라서 $Gain\ Ratio(x_i)$는 자료 D를 x_i로 분할함으로써 발생한 이득의 상대량을 의미한다. C5.0은 $Gain\ Ratio$를 이용하여, $Gain\ Ratio(x_i)$가 최대화되는 점에서 데이터의 분할을 선택한다.

2.5. CART 기법 소개

CART (classification and regression tree)는 설명변수들과 목표변수로 이루어진 자료들에서 설명변수들의 특성에 따라 자료들을 이지분류 (binary split)하여, 2개의 하위노드를 생산하는 과정을 반복하여 자료들을 목표변수의 값이 유사한 부분집합으로 만드는 방법이다. CART의 알고리즘은 마디의 순수함을 나타내는 지니계수 (gini index)에 의해 분리여부를 결정한다. 특정 변수에 의해 집단이 구분되면,

구분된 하나의 집단에서 나머지 집단의 개체가 선택될 확률을 계산하여 집단을 분리하며 집단이 순수할수록 지니계수의 값이 작아지며 확률 또한 작아지게 된다 (Berson 등, 2000).

만일 목표변수의 범주가 k개로 분할되고, i번째 범주에 분류될 확률이 $p_1, p_2, \cdots, p_k$일 때, 지니계수는 아래의 식과 같이 정의된다 (Breiman 등, 1984; Heo와 Lee, 2008).

$$G = 1 - \sum_{i=1}^{k} p_i^2 = 1 - \sum_{i=1}^{k} \left(\frac{n_i}{n}\right)^2 \tag{2.14}$$

여기서 n은 그 마디에 포함된 관찰치 수, n_i는 목표변수의 i번째 범주에 속하는 관찰치의 수를 의미한다. CART는 지니계수를 가장 감소시켜 주는 설명변수와 그 변수의 최적분리를 자식노드로 선택하는데, 지니 계수의 감소량은 다음과 같이 계산한다.

$$\Delta G = G - \frac{n_L}{n} G_L - \frac{n_R}{n} G_R \tag{2.15}$$

여기서 n은 부모노드의 관측치 수, n_L과 n_R은 각각 자식노드의 수를 의미한다. 즉, 자식노드로 분리되었을 때 불순도가 가장 감소시켜주는 예측변수와 그 때의 최적분리에 의해 자식마디를 형성하는 것이며, 이는 다음과 같은 자식마디에서 불순도의 가중 합을 최소화하는 것과 동일하다.

$$P(L)G_L + P(R)G_R = \frac{n_L}{n} G_L + \frac{n_R}{n} G_R \tag{2.16}$$

3. 호흡기질환군의 주요인에 대한 데이터마이닝 분석

3.1절에서는 호흡기질환 자료에 대해서 살펴보고, 3.2절에서는 5가지 데이터마이닝 기법인 인공신경망, 로지스틱 회귀모형, 베이지안 네트워크, C5.0, CART를 호흡기질환 자료에 적용시켜 정확도가 가장 높은 모형을 선별한다. 이러한 데이터 마이닝 기법을 적용하기 위하여 IBM SPSS Modeler 버전 14.1을 활용하였다.

3.1. 실험자료

본 연구는 2012년도 지역사회건강조사의 경상북도 자료를 이용하였다. 대상자는 2012년 8월 16일부터 10월 31일까지 2012년 7월 기준으로 시·군·구에 거주하는 만 19세 이상의 성인이며 조사 참여자 수는 22,304명이다. 조사 문항 중 의사에게서 폐결핵, 천식, 알레르기성 비염을 진단받은 경험이 있는 2,496명을 호흡기질환군으로, 기타 만성질환을 진단받은 경험이 있는 11,823명을 기타질환군으로, 만성질환을 진단받은 경험이 없는 7,985명을 건강군으로 분류하여 기타질환군을 제외한 10,481명을 최종 분석 대상자로 결정하였다.

Table 3.1은 분석에 사용된 7가지 독립변수를 나타낸 것이다.

1) 현재흡연여부 (current smoking status)
   : 현재흡연자/ 비흡연자/ 과거흡연자
2) 현재음주여부 (current drinking status)
   : 최근 1년 동안 한 달에 1번 이상 음주를 한 자/ 1번미만 음주를 한 자
3) 고위험음주여부 (high risk drinking)
   : 한 번의 술자리에서 남자 7잔 이상, 여자 5잔 이상을 일주일에 1번 이상 한 자/ 1번 미만 한 자
4) 주관적인 체형 인지 여부 (perception of body type)
   : 마름/ 정상/ 비만
5) 체중 조절 시도 여부 (weight control experience)

: 최근 1년 동안 체중을 조절하려고 노력한 경험 여부

6) 스트레스 인지 여부 (perception of stress)

: 평소 일상생활 중에 스트레스를 인지하는지 여부

7) 우울감 인지 여부 (perception of depression)

: 최근 1년 동안 연속적으로 2주 이상 일상생활에 지장이 있을 정도의 우울감 (슬픔이나 절망)을 인지한 경험 여부

Table 3.1 Seven kinds of risk factors for respiratory disease

Independent variable	Description for the independent variable
Current smoking status	1 : Smoking 2 : Non-smoking 3 : Smoking in the past
Current drinking status	1 : Drinking $\geq$ Once a month during the last one year 0 : Drinking $<$ Once a month during the last one year
High risk drinking	1 : Drinking $\geq$ Once a week 0 : Drinking $<$ Once a week
Perception of bady type	1 : Thin 2 : Normal 3 : Fat
Weight control experience	1 : Weight loss 2 : Weight maintenance 3 : Weight gain 4 : Non control
Perception of stress	1 : Yes 0 : No
Perception of depression	1 : Yes 0 : No

Table 3.2 Result for respiratory disease by neural network

Group			N-Group		Total
			RD	Control	
Train data	RD	Count	661	552	1213
		Column(%)	30.931	18.170	23.440
	Control	Count	1476	2486	3962
		Column(%)	69.069	81.830	76.560
	Total	Count	2137	3038	5175
		Column(%)	100	100	100
Test data	RD	Count	725	553	1278
		Column(%)	31.743	18.427	24.182
	Control	Count	1559	2448	4007
		Column(%)	68.257	81.573	75.818
	Total	Count	2284	3001	5285
		Column(%)	100	100	100

RD; respiratory disease, N-Group; group classified by neural network

Table 3.3 Result for respiratory disease by logistic regression

Group			L-Group		Total
			RD	Control	
Train data	RD	Count	603	610	1213
		Column(%)	32.358	18.412	23.440
	Control	Count	1259	2703	3962
		Column(%)	67.615	81.588	76.560
	Total	Count	1862	3313	5175
		Column(%)	100	100	100
Test data	RD	Count	644	634	1278
		Column(%)	32.281	19.271	24.182
	Control	Count	1351	2656	4007
		Column(%)	67.719	80.729	75.818
	Total	Count	1995	3290	5285
		Column(%)	100	100	100

RD; respiratory disease, L-Group; group classified by logistic regression

Table 3.4 Result for respiratory disease by Bayesian network

Group			B-Group		Total
			RD	Control	
Train data	RD	Count	657	556	1213
		Column(%)	30.332	18.478	23.440
	Control	Count	1509	2453	3962
		Column(%)	69.668	81.522	76.560
	Total	Count	2166	3009	5175
		Column(%)	100	100	100
Test data	RD	Count	730	548	1278
		Column(%)	31.479	18.476	24.182
	Control	Count	1589	2418	4007
		Column(%)	68.521	81.524	75.818
	Total	Count	2319	2966	5285
		Column(%)	100	100	100

RD; respiratory disease, B-Group; group classified by Bayesian network

Table 3.5 Result for respiratory disease by C5.0

Group			C-Group		Total
			RD	Control	
Train data	RD	Count	652	564	1216
		Column(%)	32.963	17.576	23.443
	Control	Count	1326	2645	3971
		Column(%)	67.037	82.424	76.557
	Total	Count	1978	3209	5187
		Column(%)	100	100	100
Test data	RD	Count	660	620	1280
		Column(%)	32.417	19.030	24.178
	Control	Count	1376	2638	4014
		Column(%)	67.583	80.970	75.822
	Total	Count	2036	3258	5294
		Column(%)	100	100	100

RD; respiratory disease, C-Group; group classified by C5.0

Table 3.6 Result for respiratory disease by CART

Group			R-Group		Total
			RD	Control	
Train data	RD	Count	333	883	1216
		Column(%)	34.330	20.939	23.443
	Control	Count	637	3334	3971
		Column(%)	65.670	79.061	76.557
	Total	Count	970	4217	5187
		Column(%)	100	100	100
Test data	RD	Count	361	919	1280
		Column(%)	34.678	21.608	24.178
	Control	Count	680	3334	4014
		Column(%)	65.322	78.392	75.822
	Total	Count	1041	4253	5294
		Column(%)	100	100	100

RD; respiratory disease, R-Group; group classified by CART

Table 3.7 Comparison of accuracy for each data mining method

Model	Train data accuracy(%)	Test data accuracy(%)
Neural network	60.8116	60.0378
Logistic regression	63.8841	62.4409
Bayesian network	60.0966	59.5648
C5.0	63.5628	62.2969
CART	70.6960	69.7960

3.2. 데이터마이닝 기법 적용 결과 및 최종모형 선택

데이터는 훈련용 데이터와 검증용 데이터를 각각 50%, 50%로 분할하고 훈련용 데이터에 부스팅을 통한 균형화 작업을 실시하여 호흡기질환군과 건강군의 비율을 50:50으로 맞추었다. 그리고 인공신경망, 로지스틱 회귀모형, 베이지안 네트워크, C5.0, CART 순으로 데이터마이닝 기법을 적용하여 모형을 구

축하였다.

Table 3.7은 각 기법에서 훈련용 데이터와 검증용 데이터에서의 정확도를 계산하여 나타낸 것이다. 훈련용 데이터에서의 정확도를 살펴보면 인공신경망은 60.81%, 로지스틱 회귀모형은 63.88%, 베이지안 네트워크는 60.10%, C5.0은 63.56%, CART는 70.70%로 CART 기법에서 가장 높은 정확도가 나타났다. 훈련용 데이터에서 만들어진 모형을 검증용 데이터에 적용한 결과를 살펴보면 인공신경망은 60.04%, 로지스틱 회귀모형은 62.44%, 베이지안 네트워크는 59.56%, C5.0은 62.30%, CART는 69.80%로 CART 기법이 다른 기법보다 정확도가 매우 높은 것을 볼 수 있다. 그러므로 훈련용 데이터와 검증용 데이터를 종합적으로 봤을 때 정확도가 가장 높게 나타난 CART 기법을 최종모형으로 선택하여 호흡기질환의 주요인을 알아보았다.

4. 호흡기질환의 주요인 판별

4.1절에서는 CART 기법을 통해서 호흡기질환에 영향을 미치는 요인을 판별하고, 4.2절에서는 이 요인에 대한 오즈비를 계산하여 개별적인 영향력에 대해서 알아본다.

4.1. 호흡기질환의 주요인 판별을 위한 CART 기법 적용

CART 기법을 통하여 호흡기질환에 영향을 미치는 요인을 판별해 본 결과, “우울감 인지 여부 > 현재흡연여부 > 스트레스 인지 여부” 순서로 이 세 가지 요인이 다른 요인들 보다 높은 중요도를 가지고 있는 것으로 나타났다. CART 기법에서의 호흡기질환 주요인에 대한 의사결정나무를 살펴보면 Figure 4.1과 같다.

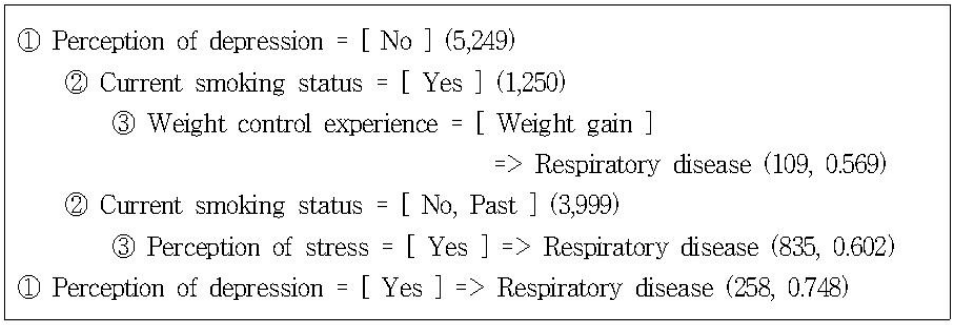

Figure 4.1 Result of CART for respiratory disease

Figure 4.1을 보면 호흡기질환군에 가장 영향을 많이 끼치는 요인으로 먼저 우울감 인지 여부가 나타났다. 다음으로 현재흡연여부로 분화되었고, 스트레스 인지 여부가 세 번째로 분화되었다. 최근 1년 동안에 연속적으로 2주 이상 일상생활에 지장이 있을 정도의 우울감을 경험해 본 사람 (258명, 74.8%)의 경우 호흡기질환군으로 판정되었다. 우울감을 경험해 보지 않은 사람 (5,249명)의 경우에는 현재 흡연자 (1,250명)이며, 최근 1년 동안 체중을 감량하려고 노력을 한 적이 있는 사람 (109명, 56.9%)의 경우 호흡기질환군으로 판정되었다. 그리고 과거에 흡연을 한 경험이 있는 사람을 포함한 현재 비 흡연자 (3,999명)인 경우에는 일상생활에서 스트레스를 많이 느끼는 편인 사람 (835명, 60.2%)의 경우 호흡기질환군으로 판정되었다.

Table 4.1 Odds ratio of risk factors for respiratory disease

Factors		Odds ratio	95% Confidence interval	
			Lower	Upper
Perception of depression	No Yes	3.475	2.827	4.27
Current smoking status	Yes No	1.684	1.506	1.883
Perception of stress	No Yes	1.676	1.512	1.858
Weight control experience	Weight gain Others	1.596	1.336	1.908

4.2. 주요인에 대한 오즈비

Table 4.1은 호흡기질환군의 주요인에 대한 오즈비를 계산해 본 결과이다. 오즈비를 살펴보면 우울감 인지 여부, 현재흡연여부, 스트레스 인지 여부, 체중 조절 시도 여부 순으로 높게 나타났다. 첫 번째 요인인 우울감 인지 여부에서 우울감을 경험해 본 그룹과 그렇지 않은 그룹에 대한 오즈비가 3.475로, 경험 있는 그룹의 경우 경험이 없는 그룹에 비해 호흡기 질환을 가질 위험도가 3.475배 높게 나타났다. 두 번째 요인인 현재흡연여부에서는 비흡연자와 흡연자에 대한 오즈비가 1.684로, 비흡연자의 경우 흡연자보다 호흡기 질환을 가질 위험도가 1.684배 높게 나타났다. 그 다음 요인인 스트레스 인지 여부에서는 평소 일상생활에 지장을 줄 정도의 스트레스를 받은 경험이 있는 그룹과 경험이 없는 그룹에 대한 오즈비가 1.676으로, 경험이 있는 그룹의 경우 없는 그룹에 비해 호흡기 질환을 가질 위험도가 1.676배 높게 나타났다. 마지막으로 체중조절시도에서 체중을 늘리려는 노력을 하는 그룹과 그렇지 않은 그룹에 대한 오즈비가 1.596으로, 노력을 한 그룹이 그렇지 않은 그룹에 비해 호흡기 질환을 가질 위험도가 1.596배 높게 나타났다.

5. 결론

2012년 8월 16일부터 10월 31일까지 2012년 7월 기준으로 시·군·구에 거주하는 만 19세 이상의 성인을 대상으로 하는 지역사회건강조사의 경상북도자료 중 10,481명의 데이터를 데이터 마이닝 기법에 적용하여 호흡기질환의 주요인을 찾아보았다.

전체 데이터를 훈련용 데이터와 검증용 데이터를 50%, 50%로 나누어 데이터마이닝 기법인 인공신경망, 로지스틱 회귀모형, 베이지안 네트워크, C5.0, CART를 호흡기질환 자료에 적용했을 때, 모형의 정확도가 가장 높게 나온 기법인 CART를 최종모형으로 선택하여 분석하였다. 그 결과 호흡기질환에 가장 영향을 많이 끼치는 요인은 우울감 인지 여부였으며, 그 다음으로 현재흡연여부, 스트레스 인지 여부 순으로 선택되었다. 일상생활에 지장이 있을 정도의 우울감을 경험한 사람의 경우 대부분 호흡기질환을 가지고 있는 것으로 판정되었다. 그러나 우울감을 경험하지 않은 사람의 경우에는 현재흡연여부에 따라 나뉘는데, 과거에 흡연을 했으나 현재는 흡연을 하지 않는 사람을 포함한 현재 비 흡연자이면서 평소 일상생활에서 스트레스를 많이 느끼는 사람의 경우 호흡기질환자로 판정되었다. 그리고 현재 흡연을 하고 있는 사람의 경우에는 체중 감량을 시도한 적이 있는 경우 호흡기질환자로 판정되었다. 호흡기질환에 가장 영향을 많이 끼치는 주요인의 영향력을 알아보기 위해 개별적인 오즈비를 계산해본 결과 우울감을 경험해 본 그룹이 그렇지 않은 그룹에 비해서 호흡기질환을 가질 위험도가 3.475배 높게 나타났다.

결과적으로 우울감의 경험 여부가 호흡기질환에 가장 영향을 많이 미친다는 것을 보였다. 우울감은 삶의 질을 떨어뜨리고 잠재적으로는 치명적일 수 있으며, 만성 호흡기질환자에게서 흔히 관찰된다 (Ryu 등, 2010). 그리고 우울 및 공황장애의 치료제로 사용되는 항정신성 약물이 심한 중증의 만성 폐쇄성 폐질환자 치료에 효과가 있다는 연구 결과가 있어 호흡병태생리와 우울 및 불안 등의 정신심리학적 관련성

을 시사하고 있다 (Ryu 등, 2007; Yohannes 등, 2000). 따라서 호흡기질환을 판단할 때 우울증 인지 여부를 확인 하는 것이 중요하며, 또한 호흡기질환이 더 악화되지 않도록 잘 관리하기 위해서는 우선적으로 우울감 및 스트레스를 극복 할 수 있도록 하는 방안을 마련해 주는 것이 필요하다.

본 연구는 다음과 같은 제한점이 있다. 첫째, 질환에 관한 연구는 비교적 긴 관찰기간이 필요하나 연구에 사용된 지역사회건강조사는 단면조사 연구로써 변수간의 시간적인 선후관계의 파악이 어려우며, 만성 질병변수와 건강위험행위변수의 인과관계의 파악이 어렵다는 단점이 있었다. 둘째, 연구에 사용된 스트레스, 우울증 등은 자가 기입 설문지의 형태이므로 주관적인 판단으로 이루어졌기 때문에 전문가의 정확한 진단을 대신하여 설명이 될 수 있는지에 대한 의문이 있다. 이러한 점들을 보완하기 위해서 향후 연구에서는 시간 경과에 따른 변화를 연구할 수 있는 코호트연구 또는 패널연구와 같은 종단적인 연구자료와 주관성을 보완할 수 있는 자료를 이용할 필요가 있을 것으로 생각된다.

References

Berry, M. and Linoff, G. (1997). *Data mining techniques: For marketing, sales and customer support*, Wiley, New York.

Berry, M. and Linoff, G. (2011). *Data mining techniques: For marketing, sales and customer relationship management*, Wiley, New York.

Berson, A., Smith, S. and Thearling, K. (2000). *Building data mining applications for CRM*, McGraw-Hill, New York.

Breiman, L., Friedman, J. H., Olshen, R. and Stone, C. J. (1984). *Classification and regression tree*, Chapman & Hall, New York.

Freund, Y. and Mason, L. (1999). The alternating decision tree learning algorithm. *Proceedings of the Sixteenth International Conference on Machine Learning*, **99**, 121-133.

Han, S. H., Park, J. S., Seo, S. H., Yoon, J. E. and Jee, S. H. (2005). Factors affecting the morbidity related to respiratory diseases in urban Korea. *Journal of the Korea Gerontology Society*, **28**, 205-217.

Heo, M. H. and Lee, Y. G. (2008). *Data mining modeling and example*, Hannarae, Seoul.

Kim, N. M., Lee, W. K. and Park, J. Y. (2013). The ecological analysis of asthmatic occurrence in patients: Using the national health insurance data. *Journal of the Korean Data & Information Science Society*, **24**, 679-688.

Kim, T. H. and Kim, Y. H. (2013). A study on the analysis of customer loan for the credit finance company using classification model. *Journal of the Korean Data & Information Science Society*, **24**, 411-425.

Kim, W. J., Bae, H. S., Choi, B. K., Hwang, J. M., Shin, K. H., Kim, M. H., Lee, K. H., Kim, K. U., Jeon, D. S., Park, H. K., Kim, Y. S., Lee, M. K. and Park, S. K. (2010). Depressive conditions in relation to asthma severity and control. *Tuberculosis and Respiratory Diseases*, **69**, 265-270.

Lee, J. W., Park, M. R. and Yoo, H. N. (2005). *Statistical methods for life science research*, Free Academy, Seoul.

Park, I. S., Han, J. T., Sohn, H. S. and Kang, S. B. (2011). Developing the administrative model using the data mining technique for injury in National Health Insurance. *Journal of the Korean Data & Information Science Society*, **23**, 467-476.

Oraka, E., King, M. E. and Callahan, D. B. (2010). Asthma and serious psychological distress: Prevalence and risk factors among US adults, 2001-2007. *CHEST Journal*, **137**, 609-616.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*, Morgan-Kaufmann Publishers, San Mateo, CA.

Ryu, Y. J., Chun, E. M., Sim, Y. S. and Lee, J. H. (2007). Depression and anxiety in outpatients with chronic obstructive pulmonary disease. *Tuberculosis and Respiratory Diseases*, **62**, 11-8.

Ryu, Y. J., Chun, E. M., Lee, J. H. and Chang, J. H. (2010). Prevalence of depression and anxiety in outpatients with chronic airway lung disease. *Korean Journal of Internal Medicine*, **25**, 51-57.

Sarle, W. S. (1994). Neural networks and statistical models. *Proceedings of the 19th Annual SAS Users Group International Conference*, 1-13.

Tan, P., Steinbach, M. and Kumar, V. (2006). *Introduction to data mining*, Addison Wesley Longman, California, USA.

Yohannes, A. M., Baldwin, R. C. and Connolly, M. J. (2000). Depression and anxiety in elderly outpatients with chronic obstructive pulmonary disease : Prevalence, and validation of the BASDEC screening questionnaire. *International Journal of Geriatric Psychiatry*, **15**, 1090-1096.

Journal of the Korean Data & Information Science Society
2014, **25**(2), 373–384

http://dx.doi.org/10.7465/jkdi.2014.25.2.373
한국데이터정보과학회지

Identification of major risk factors association with respiratory diseases by data mining

Jea-Young Lee[1] · Hyun-Ji Kim[2]

[12]Department of Statistics, Yeungnam University

Received 12 February 2014, revised 5 March 2014, accepted 10 March 2014

Abstract

Data mining is to clarify pattern or correlation of mass data of complicated structure and to predict the diverse outcomes. This technique is used in the fields of finance, telecommunication, circulation, medicine and so on. In this paper, we selected risk factors of respiratory diseases in the field of medicine. The data we used was divided into respiratory diseases group and health group from the Gyeongsangbuk-do database of Community Health Survey conducted in 2012. In order to select major risk factors, we applied data mining techniques such as neural network, logistic regression, Bayesian network, C5.0 and CART. We divided total data into training and testing data, and applied model which was designed by training data to testing data. By the comparison of prediction accuracy, CART was identified as best model. Depression, smoking and stress were proved as the major risk factors of respiratory disease.

Keywords: Data mining, decision tree, respiratory disease.

[1] Corresponding author: Professor, Department of Statistics, Yeungnam University, Kyungsan 712-749, Korea. E-mail: jlee@yu.ac.kr
[2] Graduate student, Department of statistics, Yeungnam University, Kyungsan 712-749, Korea.