http://dx.doi.org/10.15207/JKCS.2015.6.2.025

인공신경망 분석과 결정트리 융합에 의한 금연 프로그램 참여 결정 요인

변해원[1*]
[1]남부대학교 언어치료청각학과

The Factors of Participating in a Smoking Cessation Program using Integrated Method of Decision Tree and Neural Network Algorithm

Haewon Byeon[1*]
[1]Department of Speech Language Pathology & Audiology, Nambu University

요 약 이 연구는 신뢰성 있는 국가통계 데이터를 이용하여 지난 1년 간 금연 시도 경험 결정 요인 모형을 구축하고 개발된 모형을 근거로 금연 프로그램 참여 표적 집단 예측에 관한 기초 자료를 제공하였다. 분석대상은 2010년 서울시복지패널조사를 완료한 19세 이상 흡연자 1,326명이다. 결과변수는 지난 1년간 금연 시도 경험으로 정의하였고, 설명변수는 연령, 성, 최종학력, 현재 취업 상태, 가구 월 평균 총 소득, 배우자 유무, 음주 여부, 주관적 건강상태, 정기적인 운동 여부, 지난 한 달 간 우울증상 여부, 현재 순환기, 내분비계, 근골격계, 호흡기계, 이비인후 질환, 간질환, 비뇨기계질환 등 질병 여부로 설정하였다. 분석방법은 인공신경망 분석과 결정트리모형을 이용하였다. CART 알고리즘을 이용한 금연 프로그램 참여 모형을 구축한 결과, 유의미한 요인은 질병 여부, 주관적 건강 상태, 가구 월 평균 총소득이었다. 이 결과를 기초로 금연 프로그램의 성공적인 시행을 위해서 표적 대상의 특성을 고려한 프로그램 개발 및 교육이 요구된다.

• **주제어** : 데이터마이닝, 인공신경망, 금연 프로그램, 의사결정나무모형, 융합 과학

Abstract The purpose of this study was to analyze the factors that affects the participating in a smoking cessation program. Data were from the A Study on the Seoul Welfare Panel Study 2010. Subjects were 1,326 smokers aged 19 and older living in the community. Dependent variable was defined as experience of smoking cessation. Explanatory variables were included as age, gender, level of education, employment status, household income, marital status, drinking, self-reported health status, depression, disease, and physical activity. A prediction model was developed by the use of a Decision Tree and Neural Network Algorithm. In the Prediction model, self reported health status, disease, income, household income were significantly associated with participating in a smoking cessation program. Based this study, systematic education and development of programs are required.

• **Key Words** : Datamining; Neural Network; Smoking Cessation Program; Decision tree; Convergence Science

*교신저자 : 변해원(byeon@nambu.ac.kr)
접수일 2015년 1월 28일 수정일 2015년 3월 12일 게재확정일 2015년 4월 20일

1. 서론

2015년 1월 1일부터 담뱃값 평균 2,000원 인상과 더불어 실내 전면 금연 구역 지정, 금연광고 확대 등 금연 정책이 강화되었다. 이 같은 정책의 변화로 인해 2015년 흡연률은 전년도에 비해 감소될 것으로 기대하고 있다. 이 같은 감소폭을 반영하더라도 우리나라 흡연율은 2013년 기준으로 28%로 여전히 높은 실정이며 경제협력개발기구(OECD) 가입 국가 중 스페인에 이어 2번째로 흡연율이 높다[1].

금연 정책은 비흡연 인구의 흡연인구로의 진입을 최소화하기 위한 금연 홍보의 다양화와 흡연 인구를 대상으로 흡연율 감소를 위한 교육/상담/치료 등 금연프로그램으로 구성된다[2, 3]. 그러나 흡연자를 대상으로 한 금연프로그램의 경우 담배에 포함되어 있는 니코틴이 약물중독의 증상을 유발하는 정신활성 물질(psychoactive agent)이기 때문에 금연에 어려움을 호소한다[4, 5]. 이 같은 이유로 흡연은 국제질병분류(ICD-10)에서 'Tobacco dependence'로 분류하고 있으며[6], 미국 정신의학회의 정신 질환 진단 및 통계 편람(DSM-V)에서도 'Tobacco Use Disorder'라는 약물 중독으로 규정된다[7].

흡연이 건강에 미치는 악영향은 다수의 연구에서 보고되었는데, 과도한 흡연은 폐암의 직접적인 원인일 뿐만 아니라 최근에는 만성폐쇄성폐질환 등의 호흡기질환과, 관상동맥성, 심잘질환, 뇌졸중, 구강암, 식도암, 후두질환 등의 위험요인으로 알려져 있다[8, 9, 10, 11].

흡연은 수정가능한 건강위험행위이므로 성공적인 금연 정책을 위해서는 흡연자를 대상자별로 세분화 한 맞춤형 프로그램의 시행이 요구된다.

그럼에도 불구하고, 금연 프로그램은 주로 성별이나 연령별로만 정책이 수립되어 왔으며, 금연 프로그램을 주제로 한 국내 연구 또한 주로 인구사회학적 요인에 초점을 맞추어 수행되었다. 표적 집단의 특성을 파악하기 위해서는 금연에 영향을 미칠 수 있는 복합적인 요인의 규명이 필요하다. 따라서 최근 패턴인식분야를 비롯한 금융시장 예측, 교량 수명 예측 등의 융복합 분야에서 이용되는 데이터마이닝 분석 기법은 금연 프로그램의 표적 집단을 예측하는 데에도 유용하게 활용될 수 있다[11].

본 연구는 신뢰성 있는 국가통계 데이터를 이용하여 지난 1년 간 금연 시도 경험 결정 요인 모형을 구축하고, 개발된 모형을 근거로 금연 프로그램 참여 표적 집단 예측에 관한 기초 자료를 제공하였다.

2. 연구 방법

2.1 분석 데이터

본 연구의 분석 데이터는 2010년 6월 1일부터 2010년 8월 31일까지 서울 시민을 대상으로 서울시복지재단에서 조사한 서울시복지패널조사(Seoul Welfare Panel Study)의 원시데이터이다. 서울시복지패널조사는 서울시에 거주하는 가구의 복지수준을 파악하고 복지취약계층의 실태파악 및 복지서비스 수요를 추정하기 위한 목적에서 2009년 통계청의 승인(제20113호)을 받아서 수행되었다[12]. 2005년 인구주택총조사 대상가구 중 조사 시기를 기준으로 한 서울시 소재 가구를 모집단으로 하였고, 표본추출방식은 서울시 25개 구를 대상으로 층화집락추출방법을 이용하였다. 주요 조사항목은 소득, 경제수준, 건강, 생활여건, 복지서비스 수요 등이며, 조사방법은 면접원이 조사 대상 가구를 방문하여 휴대용 컴퓨터에 구조화된 설문에 따라 응답한 내용을 입력하는 컴퓨터를 이용한 대면면접조사(Computer Assisted Personal Interviewing)방법을 이용하였다.

본 연구는 서울시복지패널 완료자 7,761명 중에서 비흡연자 6,435명을 제외한 19세 이상 흡연자 1,326명을 분석대상으로 하였다.

2.2 변수의 측정 및 정의

결과 변수는 지난 1년간 금연 시도 경험(있음, 없음)으로 정의하였다. 설명변수는 연령(20-39세, 40-59세, 60세 이상), 성, 최종학력(초등학교 이하, 중학교, 고등학교, 대학졸업 이상), 현재 취업 상태(취업, 미취업), 가구 월 평균 총 소득(200만원 미만, 200-400만원, 400만원 이상), 배우자 유무(배우자가 있고 함께 살고 있음, 배우자가 있으나 함께 살고 있지 않음, 배우자 없음), 음주 여부(음주자, 비음주자), 주관적 건강상태(좋음, 보통, 나쁨), 정기적인 운동 여부(없음, 있음) 지난 한 달 간 우울증상 여부(없음, 있음), 현재 순환기, 내분비계, 근골격계, 호흡기계, 이비인후 질환, 간질환, 비뇨기계질환 등 질병 여부(없음, 있음)를 포함하였다.

2.3 분석 방법

2.3.1 인공신경망 분석

금연 프로그램 참여 요인 탐색은 인공신경망(Neural Network)을 이용하여 분석하였다. 인공신경망 분석은 인

간 두뇌의 신경망을 모방해서 실제 데이터로부터의 반복적인 학습 과정을 거쳐 데이터에 숨어 있는 패턴을 찾아내는 데이터마이닝 모델링 기법으로서 복잡한 구조를 가지는 자료에서 예측문제를 해결하기 위해 사용하는 유연한 비선형 모형(nonlinear models)이다.

인공신경망은 계층구조를 갖는 수많은 프로세싱 요소로 이루어진 수학모형으로서 과거의 입력 데이터 값과 해당 출력 데이터 값들을 통해 가중치들이 반복적으로 조정되어 결국 입력 및 출력간의 관계가 학습되는 구조이다. 인공신경망의 구조는 각 입력변수에 대응되는 마디들로 구성된 입력층(Input Layer)과 여러 개의 은닉마디로 구성된 은닉층(Hidden Layer)으로 구성된다. 은닉마디는 입력층으로부터 전달되는 변수 값들의 선형결합을 비선형 함수로 처리하여 출력층 또는 다른 은닉층에 전달한다.

본 연구에서는 은닉층의 결합함수로 원형기준함수(Radial Basis Function; RBF)를 사용하는 RBF 신경망을 이용하였다. 원형기준함수는 식(1)과 같다.

$$H_i = \exp\left(-\frac{(x_1-c_{1i})^2+(x_2-c_{2i})^2+\cdots+(x_k-c_{ki})^2}{r_i^2}\right) \quad (1)$$

인공신경망 분석에서 입력변수 중요도(relative importance of inputs)가 0.1 이상인 변수는 결과변수의 결정에 영향을 주는 주요 설명변수로 간주하여 결정트리 모형에 포함하였다.

2.3.2 결정트리모형

결정트리모형은 CART(Classification And Regression Tree) 알고리즘을 이용하여 구축하였다. CART는 지니 계수(Gini Index)를 이용하여 불순도를 측정하며, 부모마디로부터 자식마디가 2개만 형성되는 이진분류 기반 알고리즘이다[13].

지니 계수는 n개의 원소 중에서 임의로 2개를 추출하였을 때, 추출된 2개가 서로 다른 그룹에 속할 수 있는 확률을 의미하며, 도출과정은 각 마디에서 도수가 가장 많은 목표변수의 오분류 확률을 식(2)의 통계식을 이용하여 산출한다[14].

$$Gini\ Index(t) = 1 - \sum_j [P(j/t)]^2 \quad (2)$$

다음으로, 지니계수의 감소량이 계산되면, 알고리즘의 마지막 과정으로 지니 계수를 가장 감소시켜 주는 분류변수와 최적 분리를 자식 마디로 선택한다.

연구의 모형에서 CART 알고리즘에 대한 의사결정규칙의 분리 및 병합 기준값은 0.05로 설정 하였고, 부모마디의 수는 200명, 자식마디 수는 100명, 분지가지 개수는 5개로 제한하였다. 최종 모형의 타당성 평가는 10배 교차검증법(K-fold cross-validation method)을 이용하여 평가하였다[15].

분석은 MINITAB version 13(Minitab Inc., State College, Pennsylvania, USA)과 Decision Tree version 20.0(IBM Inc., Chicago, Illinois, USA)을 이용하였다.

3. 결과

3.1 대상자의 일반적 특성

연구 대상의 일반적 특성은 <Table 1>에 제시하였다. 흡연자 1,326명 중에서 지난 1년간 금연 경험이 있는 사람은 527명(39.7%) 이었고, 금연 경험이 없는 사람은 799명(60.3%) 이었다. 교차검정 결과, 금연 경험이 있는 사람과 금연 경험이 있는 사람은 최종학력, 배우자 유무, 주관적 건강상태, 정기적인 운동 여부, 지난 한 달간 우울증상, 현재 질병 여부에서 통계적으로 유의미한 차이가 있었다($p<0.05$).

지난 1년간 금연 경험률은 고등학교 졸업자(66.0%), 미혼 또는 배우자가 없는 사람(67.7%), 주관적 건강상태가 좋은 사람(66.5%), 운동을 안 하는 사람(63.1%), 지난 한 달간 우울증상이 없는 사람(61.5%), 현재 질병이 있는 사람(64.1%)이 높았다.

3.2 인공신경망 분석을 이용한 요인 탐색

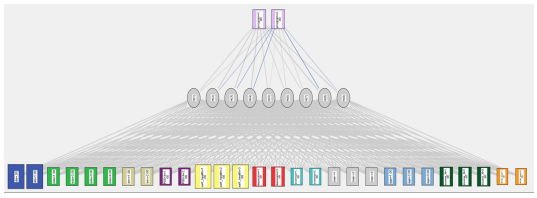
훈련 표본 669명(61.0%), 검정 표본 315명(28.7%), 검증 표본 113명(10.3%)에 대한 인공신경망 분석 결과, 검정 데이터의 오차를 가장 적게 발생시키는 은닉층의 수는 9개가 도출되었고, 제곱합 오차는 9.44, 분류정확도는 훈련 표본 98.5%, 검정 표본 96.8%, 검증 표본 98.2%로 도출되었다. ROC 곡선 면적(AUROC)은 0.802로 분류모형의 적합도와 설명력은 우수한 것으로 평가되었다.

연령, 성, 최종학력, 현재 취업 상태, 가구 월 평균 총소득, 배우자 유무, 음주 여부, 주관적 건강상태, 정기적

인 운동 여부, 지난 한 달 간 우울증상 여부, 현재 순환기, 내분비계, 근골격계, 호흡기계, 이비인후 질환, 간질환, 비뇨기계질환 등 질병 여부를 입력변수로 선정하여 다층신경망모형을 구축한 시냅스 가중값이 주어진 네트워크 다이어그램의 결과는 [Fig. 1]에 제시하였다.

〈Table 1〉 Characteristics of the subjects, n (%)

Variables	Experience of smoking cessation		p
	No (n=799)	Yes (n=527)	
Age			0.267
20-39	218 (37.3)	366 (62.7)	
40-59	218 (41.3)	310 (58.7)	
60 ≤	91 (42.5)	123 (57.5)	
Sex			0.766
Male	491 (39.9)	741 (60.1)	
Female	36 (38.3)	58 (61.7)	
Education			0.017
Elementary school	46 (44.2)	58 (55.8)	
Middle school	50 (45.0)	61 (55.0)	
High school	156 (34.0)	303 (66.0)	
Collage	275 (42.2)	377 (57.8)	
Employment status			0.540
Employment	362 (40.3	536 (59.7)	
Unemployed	165 (38.6)	263 (61.4)	
Household income			0.685
2,000,000 ≥	175 (37.3)	294 (62.7)	
2,000,000-4,000,000	190 (39.3)	294 (60.7)	
4,000,000 ≤	59 (41.0)	85 (59.0)	
Marital status			<0.001
Married	370 (43.8)	475 (56.2)	
Separated	9 (39.1)	14 (60.9)	
Unmarried	148 (32.3)	310 (67.7)	
Drinking			0.937
No	100 (39.5)	153 (60.5)	
Yes	427 (39.8)	646 (60.2)	
Self reported health status			<0.001
Good	210 (33.5)	416 (66.5)	
Normal	201 (43.1)	265 (56.9)	
Bad	116 (49.6)	118 (50.4)	
Physical activity			0.005
No	314 (36.9)	536 (63.1)	
Yes	213 (44.7)	263 (55.3)	
Depression			0.029
No	434 (38.5)	693 (61.5)	
Yes	93 (46.7)	106 (53.3)	
Disease			<0.001
No	175 (50.6)	171 (49.4)	
Yes	352 (35.9)	628 (64.1)	



[Fig. 1] Synaptic weighted network diagram

모형 훈련 후의 네트워크 다이어그램에서 시냅스 가중값은 주어진 레이어와 다음 레이어 사이의 관련성을시각적으로 제시하며, 연결 가중치가 높을수록 레이어간의 선이 굵게 제시된다. 본 모형에서 현재 질병 여부, 주관적 건강 상태, 지난 한 달 간 우울증상, 가구 평균 월 총소득은 금연 경험의 가중치가 높은 주요 변수로 가정되었다. 입력 변수중요도와 입력 변수를 영향 범주별로 영향 정도의 단위를 동일하게 변환한 정규화중요도(normalized importance)는 <Table 2>와 같다.

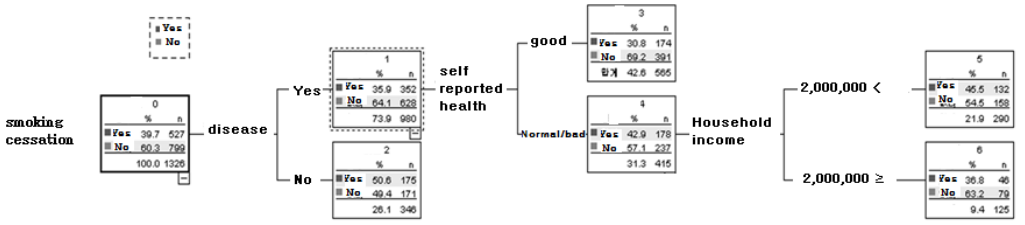
〈Table 2〉 Relative importance of inputs

Inputs	Relative importance	Normalized importance
Disease	0.193	100.0%
Education	0.074	38.5%
Sex	0.070	36.3%
Employment status	0.064	33.0%
Self reported health status	0.179	92.8%
Depression	0.102	52.6%
Physical activity	0.054	27.8%
Marital status	0.067	34.8%
Age	0.060	31.2%
Household income	0.096	49.5%
Drinking	0.041	21.4%

3.2 CART 알고리즘의 금연 프로그램 결정 요인

CART 알고리즘을 이용한 금연 프로그램 결정 요인은 [Fig. 2]에 제시하였다. 인공신경망 분석을 이용하여 금연 경험의 가중치가 높은 주요 변수들을 예측 모형에 포함한 후 CART 알고리즘을 이용하여 통계학적 분류모형을 구축한 결과, 유의미한 영향을 미치는 분류 변수는 현재 질병 여부, 주관적 건강 상태, 가구 월 평균 총소득이었다.

금연 프로그램 참여에서 가장 우선적으로 관여하는 예측 요인은 현재 질병 여부였다. 두 번째 분류 변수는



[Fig. 2] Prediction model for participating in a smoking cessation program

〈Table 3〉 Gains chart of predictor variable by CART algorithm

Node no	Node n (%)[1]	Gain n (%)[2]	Response %[3]	Gain Index %[4]	Characteristic
2	346 (26.1)	175 (33.2)	50.6	127.3	Disease-no
5	290 (21.9)	132 (25.0)	45.5	114.5	Disease-yes; Self reported health status-normal or bad; Household income-≥ 2,000,000 won

[1] Node n(%); node number, % to 1,326

[2] Gain n(%); gain number, % to 39.7

[3] Response (%): The fraction of the participating in a smoking cessation program

[4] Gain index (%):=127.3 in total 4 node

현재 질병이 있는 집단의 주관적 건강 상태였다. 세 번째는 주관적 건강상태가 나쁜 집단에서 가구 월 평균 총 소득이 분류 변수였다.

<Table 3>은 금연 프로그램 참여의 유의미한 경로를 이득율이 높은 순서 데로 제시한 CART 알고리즘기반 예측모형의 이익 도표이다. 마디번호는 최종마디의 번호이고, Gain Index (%)는 최종 노드에 대한 이익지표이다.

이익지표를 도출했을 때, 총 4개의 경로 중에서 유의미한 경로로 2번 노드와 5번 노드가 확인 되었다. 먼저, 금연 프로그램 참여에 있어서 이익지표 값이 가장 큰 제1경로는 현재 질병이 없는 사람으로 33.2%가 금연 프로그램 참여자로 분류되었고, 이익지표는 127.3%이었다. 제2경로는 현재 질병이 있고, 자신의 건강 수준을 보통 이하로 인지하고, 가구 평균 월 소득이 200만원을 초과하는 사람으로 25.0%가 금연 프로그램 참여자로 예측되었고, 이익지표는 114.5%이었다.

개발된 예측 모형의 평가는 10-fold 교차타당성 검정을 이용하였다. 도출된 모형의 안정성을 비교한 결과, 위험지수는 크로스 분류모형의 위험지수는 0.280, 오분류율은 28%로 도출되어, 예측모형의 위험지수 0.270 및 오분류율 27%와 동일하였다.

4. 결론

본 연구는 데이터마이닝의 신경망 분석과 결정트리모형을 결합한 통계적 예측모형을 이용하여 금연 프로그램 참여의 결정 요인을 분석하였다.

이 연구에서 현재 질병이 없는 사람과 현재 질병이 있고, 자신의 건강 수준을 보통 이하로 인지하고, 가구 평균 월 소득이 200만원을 초과하는 사람은 금연 프로그램 참여 가능성이 높은 집단이었다.

본 결과를 기초로 금연 프로그램의 성공적인 시행을 위해서 표적 대상의 특성을 고려한 프로그램 개발 및 교육이 요구된다.

REFERENCES

[1] Organization for Economic Cooperation and Development, "Health Data 2013", Paris, Organisation for Economic Cooperation and Development. 2014.

[2] M. Seo, "Current situation and policy goods of anti-smoking policy: with a focus on Health Plan

2010", Health and Welfare Forum, Vol. 175, pp. 74-83, 2011.

[3] K. Cho, "National tobacco control programmes", Health and Welfare Forum, Vol. 105, pp. 1226-3648, 2005.

[4] P. S. Hendricks, J. W. Ditre, D. J. Drobes, T. H. Brandon, "The early time course of smoking withdrawal effects", Psychopharmacology, Vol. 187, pp. 385-396, 2006.

[5] W. Fernando, R. J. Wellman, J. R. DiFranza, "The relationship between level of cigarette consumption and latency to the onset of retrospectively reported withdrawal symptoms", Psychopharmacology, Vol. 188, pp. 335-342, 2006.

[6] K. Wilhelm, P. Mitchell, T. Slade, Brownhill, G. Andrews, "Prevalence and correlates of DSM-IV major depression in an Australian national survey", J Affect Disord, Vol. 75, pp. 155-162, 2003.

[7] D. Kandel, C. Schaffran, P. Griesler, J. Samuolis, M. Davies, R. Galanti, "On the measurement of nicotine dependence in adolescence: comparisons of the mFTQ and a DSM-IV-Based Scale", J Pediatr Psychol, Vol. 30, pp. 319-332, 2005.

[8] J. Banoczy, C. Squier, "Smoking and disease", European Journal of Dental Education, Vol. 8, pp. 7-10, 2004.

[9] H. Byeon, "The trend of the association between amount of smoking and self-reported voice problem", Journal of the Korea Academia-Industrial cooperation Society, Vol. 13, No. 3, pp. 1246-1254, 2012.

[10] H. Byeon, "Analysis of Linear Trend Relationships Between Self-reported Voice Problems and Dysphonia", Journal of speech & hearing disorders, Vol. 21, pp. 89-102, 2012.

[11] H. Byeon, "The Prediction Model for Self-Reported Voice Problem Using a Decision Tree Model", Journal of the Korea Academia-Industrial cooperation Society, Vol. 14, pp. 3368-3373, 2013.

[12] Seoul Welfare Foundation, Seoul, "Seoul Welfare Panel Study 2010", Seoul Welfare Foundation. 2010.

[13] L. Brieman, J. Friedman, R. A. Olshen, C. J. Stone, New York, "Classification and Regrression Trees", Chapman & Hall. 1984.

[14] H. Byeon, "The factors that affects the experience of discrimination in children in multi-cultural families using QUEST algorithm : focusing on Korean language education", Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology, Vol. 4, pp. 303-312, 2014.

[15] L. Breiman, "Bagging predictors", Machine Learning, Vol. 24, No. 2, pp. 123-140, 1996.

저자소개

변 해 원(Haewon Byeon) [정회원]

- 2009년 8월 : 단국대학교 대학원 언어병리학 (이학석사)
- 2013년 2월 : 아주대학교 대학원 의학과 사회보건학 (이학박사)
- 2011년 3월 ~ 2013년 2월 : 대림대학교 언어재활과 조교수
- 2013년 3월 ~ 현재 : 남부대학교 언어치료청각학과 조교수 및 언어치료센터장

<관심분야> : 건강 예측 모형, 음성 의학, 노년기 의사소통장애