한국시뮬레이션학회 논문지
Vol. 29, No. 1, pp. 11-21 (2020. 3)
http://doi.org/10.9709/JKSS.2020.29.1.011
ISSN 1225-5904

특징집합 IG-MLP 평가 기반의 최적화된 특징선택 방법을 이용한 질환 예측 머신러닝 모델

김경륜 · 김재권 · 이종식 †

Optimized Feature Selection using Feature Subset IG-MLP Evaluation based Machine Learning Model for Disease Prediction

Kyeongryun Kim · Jaekwon Kim · Jongsik Lee†

ABSTRACT

Cardio-cerebrovascular diseases (CCD) account for 24% of the causes of death to Koreans and its proportion is the highest except cancer. Currently, the risk of the cardiovascular disease for domestic patients is based on the Framingham risk score (FRS), but accuracy tends to decrease because it is a foreign guideline. Also, it can't score the risk of cerebrovascular disease. CCD is hard to predict, because it is difficult to analyze the features of early symptoms for prevention. Therefore, proper prediction method for Koreans is needed. The purpose of this paper is validating IG-MLP (Information Gain – Multilayer Perceptron) evaluation based feature selection method using CCD data with simulation. The proposed method uses the raw data of the 4^{th} ~ 7^{th} of The Korea National Health and Nutrition Examination Survey (KNHANES). To select the important feature of CCD, analysis on the attributes using IG-MLP are processed, finally CCD prediction ANN model using optimize feature set is provided. Proposed method can find important features of CCD prediction of Koreans, and ANN model could predict more accurate CCD for Koreans.

Key words : Cardio-cerebrovascular Disease(CCD), Information Gain(IG), Artificial Neural Network (ANN), Feature Selection, The Korea National Health and Nutrition Examination Survey (KNHANES)

요 약

암을 제외한 한국인의 가장 높은 사망원인은 심뇌혈관질환으로 사망원인의 24%를 차지한다. 현재 국내 환자의 심혈관질환의 위험도 산출은 프레밍험 위험지수를 기반으로 하지만, 국외의 가이드라인에 의존하고 있어 정확도가 떨어지는 편이며, 뇌혈관질환의 예측에 대한 위험도는 산출할 수 없다. 심뇌혈관질환은 예방을 위한 조기증상들의 특징 분석이 어려워 질환 예측이 힘들며, 한국인에 적합한 예측 방법이 필요하다. 본 연구의 목적은 심뇌혈관질환 데이터를 이용하여, 특징집합 IG-MLP 평가 기반의 특징선택 방법론을 시뮬레이션 하여 검증하는 것이다. 제안하는 방법은 제4~7기 국민건강영양조사 원시자료를 이용한다. 심뇌혈관질환의 예측에 중요한 특징들을 선별하기 위해, 속성들의 심뇌혈관질환에 대한 정보이득-다층신경망을 이용한 분석을 실시하며, 최종적으로 선별된 특징을 이용한 심뇌혈관질환 예측 모델을 제공한다. 제안하는 방법으로 한국인의 심뇌혈관질환에 관련된 중요한 특징들을 찾을 수 있으며, 최적화된 특징들로 구성된 예측 모델은 한국인에 대해 더욱 정확한 심뇌혈관 예측을 할 수 있다.

주요어 : 심뇌혈관질환, 정보이득, 인공신경망, 특징선택, 국민건강영양조사

* This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (NRF-2018R1D1A1B07044571)

Received: 6 September 2019, **Revised**: 13 November 2019, **Accepted**: 20 December 2019

† **Corresponding Author**: Jongsik Lee
E-mail: jslee@inha.ac.kr
Department of Computer Engineering, Inha University

1. 서론

현대 사회는 초고령화의 진행과 함께 질병으로 인한 사망률이 증가하고 있다. 2017년 통계청의 한국인 사망원인에 대한 발표에 따르면, 암을 제외한 사망원인 1,2위는 각각 심혈관질환과 뇌혈관질환으로, 총 24.3%를 차지

한다(통계청, 2017). 심뇌혈관 질환은 협심증과 심근경색과 같은 심혈관질환, 그리고 뇌졸중, 뇌경색과 같은 뇌혈관질환이 포함된다. 국민관심질병통계에 따르면 근 5년간의 심혈관질환과 뇌혈관질환의 환자는 2014년에 비해 각각 16.4%, 13.9% 증가했다(건강보험심사평가원, 2018). 또한, 심뇌혈관질환은 예방 및 초기대응에 대한 인지도가 낮고, 증상 인지율 또한 20% 미만으로 낮다(오경재, 2016).

본 연구의 목적은 IG-MLP 특징선택 방법을 심뇌혈관질환 데이터를 이용하여 검증하는 것에 있으며, 검증을 위해 인공신경망을 이용한 심뇌혈관질환 예측 모델에 중요한 특징들을 정보이득(Information gain, IG) – 다층신경망(Multilayer perceptron, MLP) 학습 기반 특징선택 방법 통해 선택한다. 특징집합의 최적화를 통해 만들어진 심뇌혈관질환 예측 인공신경망 모델의 정확도를 높일 수 있다(Karegowda et al., 2010; Hu et al., 2015).

제안하는 방법은 질병관리본부에서 발표하는 국민건강영양조사의 제4기~제7기 원시자료를 이용하며(질병관리본부, 2007-2017), 데이터 셋에서 심뇌혈관질환의 예측에 중요한 특징들을 선별한다. 또한 선별된 특징들로 심뇌혈관질환을 예측하는 인공신경망을 제공한다. 특징들의 선별을 위해, 먼저 특징들의 클래스에 대한 IG의 계산을 통해 특징들을 필터링한다. 선별된 특징들과 나머지 특징들로 생성된 특징부분집합들을 특징선택을 위한 MLP에 입력하고, 입력된 특징부분집합의 학습, 테스트를 통한 평가 결과의 단계별 시뮬레이션 비교를 통해 최적화된 특징들을 찾아낼 수 있다.

2. 관련연구

한국인의 심혈관질환 위험도는 프레밍험위험지수(Framingham Risk Score, FRS)를 사용했지만(Wilson et al., 1998), 미국인을 대상으로 한 비교적 오래된 연구이고, 한국인에 대해서 과대평가 된다고 알려져 있다(Ahn et al., 2006). 또한 FRS는 관상동맥질환의 발생 가능성에 대해서만 예측이 가능하다는 제한점이 있다. 최근에는 유럽의 SCORE산정법(Systematic Coronary Risk Ealuation)도 이용하며, ASCVD(Atherosclerotic Cardiova-scular Disease) 위험지수도 사용되고 있지만, 아시아인에 대해 과대평가 된다고 알려져 있다(Cho, 2018).

우리나라의 보건복지부에서는 심뇌혈관질환 관리 계획 등을 수립하여 인프라를 구축 중이며, 그에 따라 개인의 건강관련 데이터 또한 매년 증가하고 있다(Jeong et al., 2016).

국내에서는 국민건강영양조사를 이용한 한국인의 심혈관질환 요인들에 대한 통계분석 연구들이 진행되었으며(Bae, 2016; Kang et al., 2017), 국내 병원의 환자들에 대한 뇌혈관질환의 위험요인 분석에 대한 연구들 또한 진행되었다(Kim et al., 2010).

3. 방법

특징 선택을 위한 방법으로 Fig. 1의 의사코드의 단계를 따른다. 실험을 위해 본 논문에서는 11개의 속성을 가지는 후보 특징집합을 정했고, 다층신경망을 이용한 특징집합의 평가 과정의 계산량을 줄이기 위해, 통계적 분석인 정보이득의 계산을 진행한다. 정보이득 계산을 통해, 선택된 초기 특징들은 다층신경망을 이용한 특징 후보 세트의 평가 과정에 항상 포함되는 특징들이다. 또한 적은 데이터양을 보완하기 위해, 10 fold 교차검증을 진행하며, 각 fold별 검증과정을 통해 선택된 특징 부분집합들의 특징들의 빈도를 계산하여 최종 특징 세트를 결정할 수 있도록 컴퓨터를 이용하여 시뮬레이션 한다.

Pseudo code for the feature selection.
Candidate feature set : $X=\{x_1,x_2,\cdots,x_n\}$.
Calculate entropy of the class using the equation (1).
Calculate information gain between the class and the features using the equation (3)
Select the features which have a high value.
Filtered feature set : $X^{IG}=\{x_1^{IG},x_2^{IG},\cdots,x_m^{IG}\}$
For $i=1$ to (# of the folds of cross validation)
  while X''s evaluation doesn't increase
    Do heuristic search to select the features except the X^{IG}'s element.
    $X'=\{x'_1, x'_2, \cdots, x'_k\}$
    Evaluate the X' using MLP.
  }
}
Select the features which has high proportion.

Fig. 1. Pseudo code of the method.

4. 데이터 모델

본 논문에서는 실험을 위해, 제4기~제7기 국민건강영양조사 원시자료 데이터 셋을 사용한다. 국민건강영양조사는 대한민국의 보건복지부 산하의 질병관리본부에서 수행되었고, 보건정책의 기초자료로 활용하기 위한 통계이다. 또한 데이터 셋에는 국민의 건강수준, 건강행태, 식

품 및 영양섭취 실태 등이 포함 되어있으며, 이는 국가 및 시도 단위의 대표서와 신뢰성을 갖춘 통계자료이다.

데이터 셋을 연구에 이용하기 위해, 연도별 원시자료를 통합한 뒤에, 데이터 전처리, 데이터 샘플링을 진행하였다. 최종적으로 선택된 데이터 셋은 총 9252개의 레코드로 이뤄져 있으며, 실험을 위해 트레이닝 셋은 6478개(70%)의 레코드로, 테스트 셋은 2774개(30%)의 레코드로 구성된다.

데이터의 통합 과정에서 심뇌혈관 예측을 위한 속성들의 선택은 기존의 질환 진단 방법인 FRS와 ASCVD의 입력변수를 기반으로 참고하였다. FRS와 ASCVD의 입력변수는 성별, 나이, 수축기혈압, 총콜레스테롤, HDL콜레스테롤, 고혈압여부, 흡연여부, 당뇨병유무 등으로 구성되어 있으며, 그 외에도 심뇌혈관 질환의 중요인자로 판단되는 이상지질혈증여부, 이완기혈압, BMI(Body Mass Index) (김상현, 2016; 김태년, 2015) 등도 포함하였다.

5. 데이터 전처리

본 연구에서, 데이터 전처리의 주요 목적은 국민건강영양조사 통합 데이터 셋을 심뇌혈관질환 예측 모델의 입력으로 사용할 수 있도록 구성하는 것이다. 데이터 전처리 단계는 데이터 정제(Data cleansing), 데이터 대체(Data imputation), 특징선택(Feature selection) 등으로 이뤄져있다. 통합된 데이터 셋에서 심뇌혈관질환 유병여부에 대해 참값을 포함한 레코드 수는 총 2910개로, 비유병자와의 데이터 균형이 맞지 않는다. 따라서 심뇌혈관질환의 유무에 따라 데이터 셋을 나눈 후, 나눠진 데이터 셋에 대해 데이터 정제와 대체 작업을 진행하고, 추출된 데이터 셋의 통합 이후에 특징선택을 진행한다. 데이터의 전처리 과정은 Fig. 2와 같다.

5.1 데이터 정제(Data cleansing)

데이터 정제 작업은 데이터 내의 결측값들을 삭제 또는 대체하는 과정이다. 검증되지 않은 품질의 데이터의 사용 시, 모델에 악영향을 끼칠 수 있다. 심뇌혈관질환의 예측을 위한 특징들의 후보는 Table 1과 같다.

데이터 셋 중, 심뇌혈관질환 클래스 값이 참인 데이터에서, 하나의 레코드에 여러 결측값이 있는 경우에는 레코드를 삭제(list wise deletion)처리 하였고, 클래스 값이 거짓인 데이터 셋에서는 결측 값이 없는 레코드만을 추출하였다.

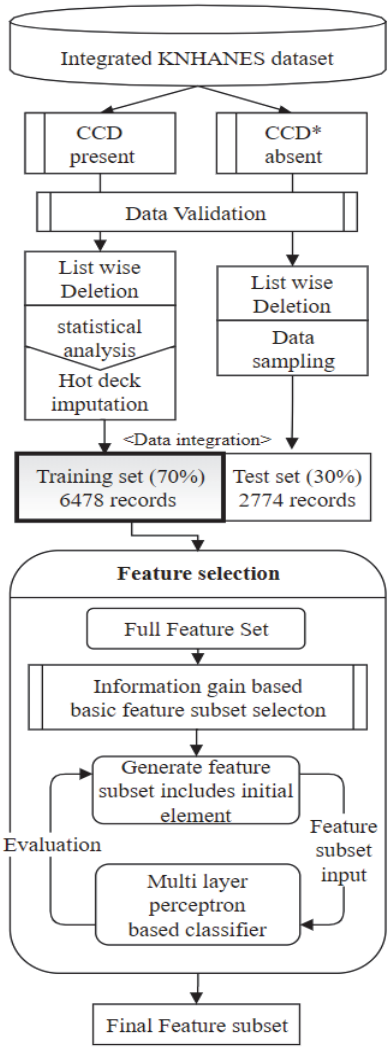

Fig. 2. Architecture of data preprocessing.
*CCD: Cardio-cerebrovascular diseases

5.2 데이터 대체(Data imputation)

결측 값의 처리 방법은 일률적 삭제(list wise deletion), 평균값 대체, 확률적 대체와 같은 단순대체(single imputation), 다중대체(multiple imputation)등의 방법이 있다.

본 연구에서는 데이터 셋 중, 심뇌혈관질환의 유병여부가 참값인 레코드들에 대해서만 데이터 대체를 진행하였다. 의료 데이터의 특성상, 특정 질환을 가진 환자의 수가 소수이기 때문에, 일률적 삭제의 방법보다는 데이터를 대체하는 것이 합리적이라 판단하였다.

결측속성, 결측비율, 대체 방법은 Table 2와 같다. 결측치 비율이 ~10%인 경우 어느 처리 방법을 사용해도

Table 1. Variables of data set

	변수	내용	범위	평균	표준편차
연속형	age	대상자 나이	[19,80]	55.1	16.9
	SBP	수축기 혈압	[79,221]	121.2	17.3
	DBP	이완기 혈압	[25,120]	75.2	10.3
	BMI	체질량 지수	[15,45]	24.1	3.4
	total_CHOL	총 콜레스테롤	[83,384]	185.9	38
	HDL_CHOL	HDL 콜레스테롤	[20,124]	48.9	12.3
				비고	
범주형	sex	성별	{1,2}	1: 남자 2: 여자	
	hypertension	고혈압 유병여부	{0,1,8}	0. 없음 1. 있음 8. 모름	
	dyslipidemia	이상지질혈증 유병여부	{0,1,8}	0. 없음 1. 있음 8. 모름	
	diabetes	당뇨병 유병여부	{1,2,3}	1: 정상 2. 공복혈당장애 3. 당뇨병	
	smoke	흡연 정도	{0,1,2,3}	0. 피우지 않음 2. 가끔 1. 과거에 피움 3. 매일	
	CCD	심뇌혈관질환 유병여부	{0,1}	0. 없음 1. 있음	

Table 2. Missing value and the ratio

속성	결측률	결측 값 대체 방법
SBP	0.275%	고혈압 유무에 따른 평균값 대체
DBP	0.275%	
BMI	0.618%	평균값 대체
total_CHOL	10.271%	이상지질혈증 유무에 따른 Hot deck 대체
HDL_CHOL	10.306%	

큰 문제가 없고, 10~20%인 경우에는 Hot deck 대체, 회귀 대체 등의 방법이 효과적이라 알려져 있다(Hair et al., 2013; Myers, 2011).

Hot deck 대체를 위해 f-검정과 t-검정으로 대체를 위한 기증(donor) 데이터를 검증하고, 결측값을 기증값으로 대체한다. 검정 결과는 Table 3과 같다. 유의수준 0.05에서 고혈압유무에 대한 수축기혈압과 이완기 혈압, 이상지질혈증의 유무에 대한 총콜레스테롤 수치는 두 집단의 평균이 달랐다고 해석할 수 있다. 하지만 이상지질혈증 유무에 따른 HDL콜레스테롤 수치의 평균에는 차이가 없었다고 해석할 수 있다. 결과에 따라, SBP, DBP, total_CHOL의 결측 값은 선택된 기증 데이터 셋에서 랜덤대체하고, HDL콜레스테롤과 BMI의 결측 값은 각각의 평균값으로 대체하였다.

Table 3. F-test and T-test of the attributes

	f-test		t-test		
	H_0	p-value	가정	H_0	p-value
$x_{hyp}{}^{*}, x_{SBP}$	두 집단의 분산이 같음	0.00	이분산	두 집단의 평균이 같음	0.00
x_{hyp}, x_{DBP}		0.00			0.00
$x_{dys}{}^{**}, x$		0.03			0.00
$x_{dys}, x_{HDL}{}^{****}$		0.06	등분산		0.73

*hypertension, **dyslipidemia,
***total_CHOL, ****HDL_CHOL

5.3 특징선택(Feature selection)

특징선택은 모델의 정확도를 향상시키기 위해, 데이터에서 가장 좋은 성능을 보여줄 수 있는 특징부분집합을 찾는 것이다. 특징선택 과정을 통해 모델의 성능과 신뢰도를 높일 수 있고, 특징선택 과정을 통해, 데이터의 차원(dimension)을 감소시켜, 보다 효율적인 학습 연산을 할 수 있다(Guyon et al., 2003).

특징선택에 중요한 것은, 특징부분집합의 생성 알고리즘과, 생성된 부분집합들에 대한 평가 수행이다. 특징부분집합의 평가에는 독립적 분석방법인 통계기반 변수 분석과, 분류기와 같은 러너(learner)를 이용한 종속적 분석방법이 있다(Huda et al., 2016). 본 논문에서는 두 방법을 융합하여 분류 모델의 학습과 성능에 효율적인 특징

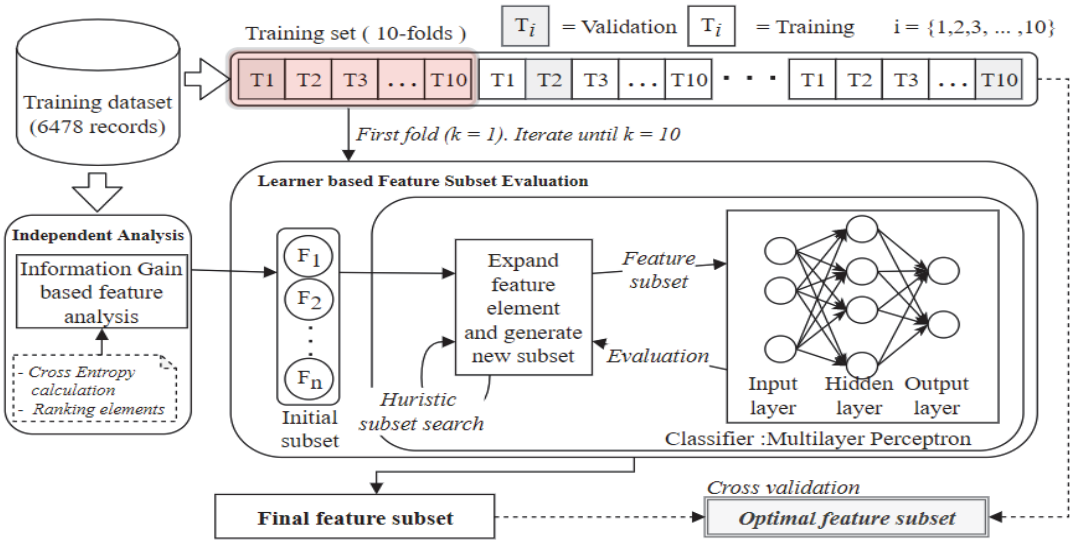

Fig. 3. Architecture of the feature selection process

집합을 선택한다.

첫 번째로, 심뇌혈관질환의 유무 클래스와, 학습을 위해 쓰일 후보 속성들 간의 IG를 계산한다(Kent, 1983). IG 계산 결과에 따라 높은 IG값을 가진 속성은 러너 기반 최적특징집합 선택과정의 최초원소로 채택된다. 러너 기반 특징부분집합의 선택은 k-fold 교차검증을 통해 이뤄진다. 아키텍처는 Fig. 3와 같다.

$$e = H(S) = -\sum_{i=0}^{n} p_i \log_2(p_i) \quad (1)$$

단, $p_i = \frac{freq(C_i, S)}{|S|}$

S : 데이터 집합
C : 특정 레코드 값의 집합
$freq(C_i, S)$: S 중, C_i에 속하는 데이터수

5.3.1 정보 이득(Information Gain)

심뇌혈관질환 예측 모델과 독립적인 특징분석을 위해, IG를 계산한다. IG는 엔트로피 계산을 기반으로 이뤄진다. 엔트로피 e는 식 (1)과 같이 정의된다(Pathria et al., 2011).

Table 4. Entropy of CCD class

S	p_1	p_2	$H(S)$
CCD	2039	4439	0.896

심뇌혈관질환 유병여부에 대한 엔트로피는 Table 4와 같다. CCD 클래스는 범주가 {0, 1}이므로, e의 범위는 $[0, 1]$이다. CCD의 e값은 0.896으로, 정보의 혼잡도가 높다. CCD와 나머지 속성 간의 교차엔트로피는 식 (2)와 같이 계산할 수 있다(Ian et al., 2016).

$$Cross\ Entropy = H(p,q) = -\sum_{i=0}^{n} p_i \log_2(q_i) \quad (2)$$

$$Information Gain(CCD, Attr) = H(CCD) - \sum_{v \in Feature} \frac{|CCD_v|}{|CCD|} H(CCD_v), \quad (3)$$

$(CCD_v = \{ i \in CCD | Attr_i = v \})$

$Attr$: 속성집합
v : $Attr$에 속한 모든 값들의 집합
CCD_v : $Attr$의 값이 v일 때 CCD의 부분집합

교차 엔트로피는 p_x에 대한 추정치 q_x로 p_x를 맞추기 위해 소모해야 하는 정보의 크기라고 해석할 수 있다. 본 연구에서는, 교차 엔트로피 식을, 특징 분석을 위한 IG의 계산에 사용하기 위해 쿨백-라이블러 발산(Kullback-Leibler divergence, KLD)로 정리한 식을 사용하였으며 (Hall, 2000), 식 (3)과 같다.

식(3)의 계산으로, 심뇌혈관실환을 예측하는 후보 특징들에 대해, CCD 클래스의 엔트로피 감소에 대한 정량

적 비교를 할 수 있다. IG의 값이 클수록 CCD의 분류에 큰 영향을 미치며, Table 5는 각 속성들에 대해 IG 값을 계산한 결과이다.

계산 결과, age와 hypertension 속성이 유의미한 IG 값을 가졌고, 혈압을 나타내는 SBP와 DBP는 IG 값이 차이가 났다. 실제 임상에서도 나이는 심뇌혈관질환의 위험도에 가장 큰 변수로 알려져 있다.

Table 5. IG results of attributes and CCD

Attribute	IG
age	0.2094
hypertension	0.1157
diabetes	0.0530
dyslipidemia	0.0529
SBP	0.0415
total_CHOL	0.0369
HDL_CHOL	0.0286
smoke	0.0152
BMI	0.0128
DBP	0.0037
sex	0.0004

5.3.2 다층신경망 기반 특징부분집합 선택

심뇌혈관질환 예측 모델의 특징선택을 위해, 특징부분집합의 평가에 MLP를 이용한다. 목적 모델과 동일한 분류기를 사용하는 특징선택 방법은 성능이 좋다고 알려져 있으나, 속성 개수에 따라 연산이 지수적으로 증가하여 효율이 떨어진다. 그러므로 본 논문에서는 IG 분석을 통한 특징부분집합의 최초 원소의 선택 후, MLP 기반 특징선택을 진행한다.

IG 계산결과, 나이와 고혈압여부가 심뇌혈관질환의 유무의 분류에 높은 영향을 미치기 때문에, 초기 특징부분집합의 원소로 채택한다. MLP 기반 특징선택의 핵심은 특징부분집합의 생성과 평가에 있다. 특징부분집합의 생성은 최초 특징부분집합 {age, hypertension}의 확장으로 이뤄진다. 본 논문에서는 추가할 속성의 선택을 위해서 휴리스틱 탐색(Heuristic search)중 하나인 Beam search를 사용한다. 평가를 위한 분류기는 MLP를 이용하며, 생성된 특징부분집합의 평가는 F-score 계산으로 이뤄진다.

Beam search 기법은 Best-first search (BFS) 알고리즘에서 종료 조건을 추가한 탐색 기법이다. BFS는 탐색 중, 노드를 모두 기억하지만, Beam search 방법은 기억할 노드의 수를 제한하여 탐색 효율성을 개선하며, MLP 기반 특징선택에 사용 시, 효율이 좋다고 알려져 있다 (Kohavi and John, 1997).

특징부분집합의 평가를 위해, MLP는 특징부분집합 생성부로부터 n개의 속성원소를 가진 데이터를 입력받는다. MLP는 2개의 히든 레이어(hidden layer)를 가지며, 히든레이어의 노드 수는 각각 $n/2$, $n+2$로 설정하였다. 입력된 특징부분집합의 평가는 F1-score 값의 비교로 이뤄지며, 계산은 Table 6과 같다.

Precision은 예측된 참값에 대한 실제 참값의 비율이다. Recall은 실제 참값에 대한 예측 참값의 비율이다. F1-score는 precision과 recall의 조화평균이며, 불균형한 데이터에 대해서도 precision과 recall 값의 균형 있는 평가가 가능하다.

Table 6. Confusion matrix

Confusion Matrix		Predicted	
		Positive	Negative
Observed	Positive	TP	FN
	Negative	FP	TN

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F1-score = 2 \times \frac{\mathrm{Precision} \times Recall}{\mathrm{Precision} + Recall}$$

$TP(True\ Positive)$: 참 값을 참으로 예측
$FP(False\ Positive)$: 거짓 값을 참으로 예측
$TN(True\ Negative)$: 거짓 값을 거짓으로 예측
$FN(False\ Negative)$: 참 값을 거짓으로 예측

Table 7. Frequency of selected elements of the feature subset.

Attribute	Frequency (10 folds)	비고
sex	2	성별
age	10	나이
hypertension	10	고혈압유무
dyslipidemia	10	이상지질혈증유무
smoke	7	흡연여부
SBP	3	수축기혈압
DBP	3	이완기혈압
BMI	7	체질량지수
diabetes	10	당뇨병여부
total_CHOL	10	총콜레스테롤
HDL_CHOL	6	HDL콜레스테롤

본 논문에서는 10-fold 교차검증(cross validation, CV)을 통해 최종 특징선택을 진행하였으며, 총 10번의 CV에서 선택된 특징부분집합의 원소의 선택 빈도는 Table 7과 같다. 기본원소로 채택한 age와 hypertension을 제외하고서라도 dyslipidemia, diabetes, total_CHOL의 속성은 10/10 빈도로 선택되었고, 다음으로는 smoke, BMI, HDL _CHOL,이 유의미하게 빈도가 높았으며, SBP와 DBP는 낮은 빈도를 가졌다.

6. 평가

6.1 측정기준

$$Sensitivity: \frac{TP}{TP+FN} \qquad Specificity: \frac{TN}{TN+FP}$$

$$PPV: \frac{TP}{TP+FP} \qquad NPV: \frac{TN}{TN+FN}$$

$$LR+: \frac{Sensitivity}{1-Specificity} \qquad LR-: \frac{1-Sensitivity}{Specificity}$$

특징선택 과정에서 선택된 특징부분집합의 빈도수가 낮은 속성들은 10번의 CV 중 최종 특징부분집합에 포함된 비율이 적었다고 해석 가능하며. 이를 기반으로, 빈도가 낮은 속성들을 제거하면서 특징집합의 최적화를 진행한다. 본 연구에서는 검증을 위해, 아래와 같은 측정지표의 계산을 수행한다.

Sensitivity는 recall과 동일한 값이고, 실제 질병을 가진 사람들이 검사를 통해 양성이라 판단될 확률이다. Specificity는 질병이 없는 사람들이 검사를 통해 음성이라 판단될 확률이며 sensitivity와 반대이다. PPV는 검사 후, 결과가 양성인 사람이 실제로 질병을 가질 확률이며, NPV는 결과가 음성인 사람이 실제 질병이 없을 확률이다. LR+는 양성 가능도비라 해석되고, 검사결과의 전반적인 판별력 지표라 여겨진다. 또한, 질환에 대해 무병한 사람의 본 검사 결과가 양성일 확률에 비해, 유병한 사람의 검사 결과가 양성으로 나올 확률이 얼마나 큰지 알려주는 지표이며, 1 보다 클수록 판별력이 우수하다고 볼 수 있다. LR-는 음성 가능도비라 해석되고, 질환에 대해 무병한 사람의 검사 결과가 음성이 나올 확률에 비해, 유병한 사람의 검사 결과가 음성으로 나올 확률이 얼마나 더 큰지 알려주는 지표이며, 1 보다 작은 값을 가지며, 값이 작을수록 좋다.

6.2 인공신경망 모델

심뇌혈관질환 예측 인공신경망 모델은 MLP를 사용한다. 전체 데이터 셋의 70%인 6478개의 트레이닝 데이터로 선택된 특징집합을 사용한 10-fold CV를 통해 학습되었으며, 평가는 30%의 2774개의 테스트 데이터를 통해 진행된다.

6.3 실험 결과 및 분석

본 연구에서 제안하는 IG-MLP기반 최적 특징선택의 실험결과는 Table 8과 같으며, 각 실험의 성능 평가는 ROC curve의 면적의 크기인 AUROC의 비교를 통해 이뤄진다. AUROC는 X_9, X_8의 특징부분집합을 사용하여 학습한 모델이 0.861로 값이 가장 높았다. 모든 속성을 포함한 X_{11}에서 sex와 DBP, SBP를 뺀 특징을 사용했을 때까지 AUROC가 상승하였였다. 또한, IG-MLP의 결과의 빈도가 10/10이었던 값들의 집합인 X_5을 사용한 모델까지는 AUROC의 값이 완만하게 감소하는 추세였지만, 그 이후의 특징부분집합으로 학습한 모델들의 AUROC 값은 급격히 떨어졌다.

X_{11}을 사용한 모델은 민감도가 0.691, 특이도는 0.836이었다. PPV값은 0.658이며 NPV 값은 0.856이었다.

X_{11}에서 sex속성을 제외한 X_{10}를 사용한 모델은 민감도가 0.767, 특이도는 0.791, PPV는 0.626, NPV는 0.882로 sex속성은 CCD 유무의 분류에 우선도가 낮다고 할 수 있다.

X_{10}에서 DBP 속성을 제외한 X_9를 사용한 모델은 민감도가 0.767, 특이도는 0.799였다. PPV는 0.634이며, NPV는 0.882였다. 또한, 심뇌혈관질환의 검사를 진행했을 때, LR+는 3.809, LR-는 0.292였다.

X_9에서 SBP 속성을 제외한 X_8을 사용한 모델은 민감도가 0.784, 특이도는 0.784로 민감도와 특이도가 비슷했다.PPV는 0.624이며, NPV는 0.889였다. 또한, LR+는 3.630, LR-는 0.275였다.

X_7은 X_8에서 $total\_CHOL$ 속성을 제외한 특징부분집합이며, 이후의 속성들은 최종 특징부분집합의 원소로 채택된 빈도가 6/10 이상이었기 때문에, X_8을 이용한 모델 보다는 성능이 떨어질 것이라 예상되었고, 실제로도 재현율은 0.788로 소폭 상승하였으나, 특이도는 0.784로 재현율에 비해 상대적으로 낙폭이 컸다. PPV또한 0.612로 감소하였고 NPV값은 0.889로 동일하였다. F1-score 값도 확연하게 떨어졌으며, AUROC 또한 떨어졌다.

검증을 위해, 특징부분집합은 IG를 통해 필터링 된 {age, hypertension}의 X_2에 도달할 때 까지 원소를 하나씩 제거하며 진행되었다.

Table 8. Evaluation of the feature subset

	Sensitivity	Specificity	PPV	NPV	LR+ (95% CI)	LR- (95% CI)	AUROC
$X_{11} = ALL^*$	0.691	0.836	0.658	0.856	4.222 (3.780-4.718)	0.370 (0.334-0.409)	0.858
$X_{10} = X_{11} - sex$	0.767	0.791	0.626	0.882	3.675 (3.342-4.040)	0.294 (0.260-0.333)	0.859
$X_9 = X_{10} - DBP$	0.767	0.799	0.634	0.882	3.809 (3.458-4.196)	0.292 (0.258-0.330)	**0.861**
$X_8 = X_9 - SBP$	0.784	0.784	0.623	0.889	3.630 (3.310-3.981)	0.275 (0.242-0.313)	**0.861**
$X_7 = X_8 - HDL\_CHOL$	0.788	0.773	0.612	0.889	3.469 (3.171-3.795)	0.275 (0.241-0.313)	0.858
$X_6 = X_7 - BMI$	0.798	0.767	0.609	0.893	3.420 (3.132-3.735)	0.263 (0.230-0.301)	0.858
$X_5 = X_6 - smoke$	0.830	0.750	0.602	0.907	3.334 (3.067-3.625)	0.226 (0.194-0.262)	0.859
$X_4 = X_5 - dyslipidemia$	0.837	0,743	0.597	0.909	3.252 (2.997-3.529)	0.219 (0.188-0.255)	0.854
$X_3 = X_4 -$	0.828	0.727	0.580	0.903	3.031 (2.800-3.281)	0.236 (0.204-0.274)	0.840
$X_2 = X_3 - diabetes$	0.822	0.733	0.583	0.901	3,081 (2.843-3.340)	0.242 (0.209-0.280)	0.841

Abbreviations : PPV, Positive Predictive value; NPV, Negative Predictive Value; LR+, Positive Likelihood Ratio; LR-, Negative Likelihood Ratio; CI, Confidence Interval, AUROC, Area Under ROC curve

특징부분집합 X_7 이후로는 Table 8과 같이 성능이 떨어짐을 확인할 수 있었다. 결론적으로, X_9, X_8의 특징부분집합을 사용하여 학습한 심뇌혈관질환 예측 모델은 거의 성능이 비슷하다고 볼 수 있으나, 더 작은 특징부분집합으로 비슷한 성능을 낸 X_8의 특징부분집합이 더 우수하다고 볼 수 있다.

6.4 실험 비교

제안하는 방법을 통해 최종적으로 선택된 인공신경망 기반의 심뇌혈관질환 예측 모델을 위한 최적화된 특징부분집합은 {나이, 고혈압, 당뇨병, 총콜레스테롤, 이상지질혈증, 흡연여부, BMI, HDL 콜레스테롤}이다.

본 연구에서 제안한 특징선택을 이용한 심뇌혈관질환 예측모델과, FRS로 계산한 결과의 ROC curve는 Fig. 4와 같으며, 제안된 모델의 정확도가 나은 성능을 보였다.

Table 9는 알려져 있는 특징선택 방법들을 사용하여 심뇌혈관질환에 관련된 속성들을 선택한 뒤, 인공신경망을 학습시켜 테스트 셋을 통해 도출된 결과표이다. FRS의 경우 기존의 프레밍험 위험지수의 도출을 위한 속성들을 사용하였으며, PC는 피어슨 상관계수를 이용한 클래스-속성 관계를 기반으로 상관지수가 0.2 이상인 속성들만을 사용하였다. IG는 클래스-속성간의 정보이득 지수를 통해 속성을 선택하였고, CFS는 클래스-특징부분집

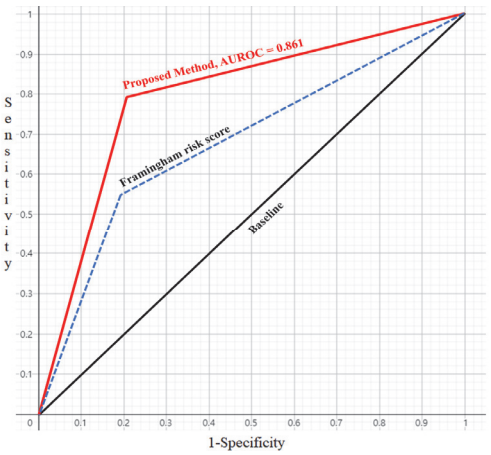

Fig. 4. ROC curve comparison of proposed model vs FRS

Table 9. Comparison with other feature selection methods.

	Sensitivity	Specificity	PPV	NPV	AUROC
IG-MLP	0.784	0.784	0.623	0.889	0.861
FRS	0.799	0.751	0.594	0.892	0.852
PC*	0.775	0.760	0.595	0.881	0.841
IG	0.744	0.787	0.614	0.871	0.852
CFS**	0.822	0.742	0.592	0.902	0.852

Abbreviations : PC, Pearson's correlation between the class and attributes; CFS, Pearson's correlation between the class and the feature subsets

Feature Subsets : **FRS** {sex, age, hypertension, smoke, SBP, total_CHOL, HDL_CHOL}; **PC** {age, hypertension, diabetes, SBP, total_CHOL, HDL_CHOL}; **IG** {age, hypertension, dyslipidemia, diabetes, SBP, total_CHOL}; **CFS** {age, hypertension, Smoke, total_CHOL, HDL_CHOL}

합의 피어슨 상관계수를 이용한 통계적 계산을 통해 특징을 선택하였다(Hall, 2000).

AUROC의 비교를 통한 분류기의 성능은 본 논문에서 제시한 IG-MLP기반 특징선택방법을 이용한 모델이 가장 성능이 좋았으며, Sensitivity의 경우는 CFS의 특징들을 이용한 모델의 예측 결과가 0.822로 IG-MLP 방법보다 0.038 정도 높았으나, 그 신뢰도와 특이도가 IG-MLP보다 각각 0.03, 0.042 낮았다. FRS 또한 Sensitivity가 IG-MLP에 비해 0.015 높았으나, 나머지가 0.03정도 낮았다.

7. 결론

최근 우리나라는 사망으로 이를 수 있는 질병의 발병률이 증가하고 있다. 심혈관질환과 뇌혈관질환의 발병률은 매해 증가 추세를 보이며, 그에 따라 보건복지부, 질병관리본부에서는 발병률을 줄이기 위한 정책들이 다수 진행 중에 있다. 심뇌혈관질환에 대한 관심이 높아짐에 따라, 관련 질환의 건강 데이터가 증가하고 있으며, 한국인의 사망에 이르는 고위험 질환의 예방을 위해 건강 빅 데이터와 인공지능과의 융합 학문이 필요하다.

본 논문에서는 IG-MLP기반의 최적 특징부분집합 선택 방법을 제안하고 검증을 위해 특징 선택 과정을 시뮬레이션 하였다. 실험을 위해, 우리나라의 보건복지부가 작성한 통계인 제4기~제7기 국민건강영양조사 통계 데이터를 이용하였으며, 기존의 특징선택 방법들과 비교했다. 또한 선별된 최적 특징부분집합을 이용한 인공신경망의 분류 모델을 제안하였고, 이를 통해 우리나라에 맞는 심뇌혈관질환의 예측을 위한 효율적인 속성들을 선택했다. 만들어진 모델은 개인의 심뇌혈관질환을 예측, 예방할 수 있다.

References

건강보험심사평가원(보건의료빅데이터개방시스템, 국민관심질병통계), 2019. 03

김상현. (2016). 이상지질혈증 진료지침의 최신지견. *Journal of the Korean Medical Association*, 59(5), 349-351.

김영은, 김일화, 문아지, 김남권, 이성근, 이기상. (2010). 뇌혈관질환 위험요인과의 분석을 통한 EAV(MERIDIAN) 활용에 관한 연구. *대한한의학회지*, 31(5), 136-145.

김태년. (2015). 심혈관대사질환 예측인자로서 허리둘레/신장비의 유용성. *대한비만학회지*, 24(2), 92-94.

오경재. (2016). 권역심뇌혈관질환센터 예방관리사업의 지역사회 성과. 2017년 대한예방의학회 가을학술대회 심포지엄 발표자료

정일영, 김석관, 이다은, 이유현. (2016). 데이터 기반 헬스케어 혁신의 부상과 대응전략. 정책연구, 1-204.

Ahn, K. A., Yun, J. E., Cho, E. R., Nam, C. M., Jang, Y., & Jee, S. H. (2006). Framingham Equation Model Overestimates Risk of Ischemic Heart Disease in Korean Men and Women. *Korean Journal of Epidemiology*, 28(2), 162-170.

Cho, Y. G. (2018). Cardiovascular Risk Prediction in Korean Adults. *Korean journal of family medicine*, 39(3), 135-136.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar), 1157-1182.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2013). Multivariate data analysis: Pearson new international edition. Pearson Higher Ed.

Hall, M. A. (2000). Correlation-based feature selection of discrete and numeric class machine learning.

Hu, Z., Bao, Y., Xiong, T., & Chiong, R. (2015). Hybrid filter–wrapper feature selection for short-term load forecasting. *Engineering Applications of Artificial Intelligence*, 40, 17-27.

Huda, S., Abawajy, J., Alazab, M., Abdollalihian, M., Islam, R., & Yearwood, J. (2016). Hybrids of

support vector machine wrapper and filter based framework for malware detection. *Future Generation Computer Systems*, 55, 376-390.

Ian G,, Yoshua B., and Aaron C. (2016). Deep Learning. MIT Press.

Karegowda, A. G., Manjunath, A. S., & Jayaram, M. A. (2010). Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management*, 2(2), 271-277.

Kent, J. T. (1983). Information gain and a general measure of correlation. Biometrika, 70(1), 163-173.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2), 273-324.

KOSIS(통계청, 2017년 사망원인통계), 2018. 09. 19.

Myers, T. A. (2011). Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data. *Communication Methods and Measures*, 5(4), 297-310.

Pathria, R. K.; Beale, Paul (2011). Statistical Mechanics (Third Edition). Academic Press. p. 51. ISBN978-0123821881.

The Fifth Korea National Health and Nutrition Examination Survey (KNHANES V), 2010-2012, Korea Centers for Disease Control and Prevention.

The Fourth Korea National Health and Nutrition Examination Survey (KNHANES IV),2007- 2009, Korea Centers for Disease Control and Prevention.

The Seventh Korea National Health and Nutrition Examination Survey (KNHANES Ⅶ 1-2), 2016-2017, Korea Centers for Disease Control and Prevention.

The Sixth Korea National Health and Nutrition Examination Survey (KNHANES Ⅵ), 2013-2015, Korea Centers for Disease Control and Prevention.

Wilson P, D'Agostino R, Levy D, Belanger A, Silbershatz H, Kannel W (1998) Prediction of coronary heart disease using risk factor categories. Circulation 97(18), 1837-1874.

Yeonhee Bae, Kowoon Lee. (2016). Risk Factors for Cardiovascular Disease in Adults Aged 30 Years and Older. *Journal of Korean Society of Integrative Medicine*, 4(2), 97-107.

Young Mi Kang, Hyun Jin Kim, Tae-yong Lee, Bon Jeong Ku. (2017). The Relationship between Death and Clinical Risk Factors in Korean: Community Cohort Study. The Journal of the Korean Public Health Association, 43(3), 81-90.

김 경 륜 (ORCID : https://orcid.org/0000-0003-0876-4948 / secretppul@gmail.com)

2018 인하대학교 컴퓨터공학과 학사
2018~ 인하대학교 전기컴퓨터공학과 석사과정

관심분야 : 인공지능, 시스템 모델링 & 시뮬레이션, 디지털 트윈

김 재 권 (ORCID : https://orcid.org/0000-0001-9982-5413 / jaekwonkorea@naver.com))

2011 가천의과학대학교 정보처리과 학사
2013 인하대학교 컴퓨터정보공학과 석사
2019 인하대학교 컴퓨터정보공학과 박사
2019~ 가톨릭대학교 의과대학 의료정보학교실 연구교수

관심분야 : 인공지능, 의료정보, 모델링 & 시뮬레이션

이 종 식 (ORCID : https://orcid.org/0000-0002-3913-372X / /jslee@inha.ac.kr)

1993 인하대학교 전자공학과 학사
1995 인하대학교 전자공학과 석사
2001 미국 애리조나대 전기·컴퓨터공학과 박사
2001~2002 캘리포니아 주립대학교 전기·컴퓨터공학과 전임강사
2002~2003 클리블랜드 주립대학교 전기·컴퓨터공학과 조교수
2003~인하대학교 컴퓨터공학과 교수

관심분야: 시스템 모델링 & 시뮬레이션, 디지털 트윈, 소프트웨어공학, 머신 러닝