Journal of the Korean Data Analysis Society (April 2018)
Vol. 20, No. 2, pp. 647-656

머신러닝을 활용한 심근경색증/협심증 예측 및 주요 위험요인 선별

임화경[1]

요 약

본 연구에서는 국민건강영양조사 제6기(2013-2015) 원시자료에 포함되어 있는 인구사회학적특성, 개인과거병력, 가족병력, 건강설문 그리고 건강검진 자료들을 활용하여, 심근경색증/협심증 발생을 예측하고 동시에 심근경색증/협심증 발생에 영향을 미치는 여러 가지 요인들 중 가장 주요한 위험요인들을 찾기 위해서 B-LASSO모형을 제안하였다. 훈련자료(training data)를 이용하여 모형을 구축하고 검증자료(test data)에서 모형들의 예측성능을 AUC를 이용하여 평가한 결과, B-LASSO모형의 예측성능이 LASSO, Random Forest, CART, SVM 모형들에 비하여 우수하였다. B-LASSO모형으로부터 추정된 심근경색증/협심증 발생에 영향을 미치는 주요 위험요인들의 중요도 순위는 연령, 고혈압 의사진단받음, 이상지질혈증 의사진단받음, 허혈성심장질환 가족력 순이었으며 남자가 여자보다 심근경색증/협심증 발생 위험이 더 높은 것으로 나타났다. 남자의 경우 당뇨병 의사진단을 받았거나 일상 활동에 지장이 있거나 빈혈이 있는 사람이 그렇지 않은 사람에 비해 심근경색증/협심증 발생 위험도가 높았으며, 여자의 경우 뇌졸중 의사진단을 받았거나 스트레스를 많이 받는 사람이 그렇지 않은 사람보다 심근경색증/협심증 발생 위험도가 높았다.

주요용어 : Bagging, LASSO, SVM, Random Forest, 국민건강영양조사.

1. 서론

최근에 우리나라에서도 식습관의 서구화, 평균 수명의 증가에 따른 고혈압 및 당뇨병의 증가로 인하여 심근경색증, 협심증과 같은 허혈성심장질환이 급증함에 따라 예방 및 치료에 관심이 높아지고 있다(Kim, 2015; Bae, Lee, 2016; Chang et al., 2018). 심근경색증은 미국인의 사망원인 1위를 차지하고 있는 질환으로 우리나라에서도 그 유병률이 빠른 속도로 증가하고 있으며 허혈성심장질환이 성인 사망률의 주요한 원인이 되고 있다(Kook et al., 2014). 최근 통계청 2016년 사망원인통계 보도자료에 의하면 2006년에는 인구 10만 명당 사망률이 암(134.0), 뇌혈관질환(61.3), 심장질환(41.1), 자살(21.8), 폐렴(9.3) 순이었는데, 2016년에는 암(153.0), 심장질환(58.2), 뇌혈관질환(45.8), 폐렴(32.2), 자살(25.6) 순으로 암 사망률 다음으로 심장질환 사망률이 가장 높았으며 심장질환 사망률은 증가 추세를 보이고 있다(Statistics Korea, 2016).

심장질환 사망률의 증가를 둔화시키기 위해서는 심장질환에 영향을 미치는 주요 위험요인을 발견하고 개선하여 예방 및 건강유지를 할 필요가 있다. 광범위한 역학적 연구의 결과로는 성별, 연령, 고혈압, 고지혈증, 흡연, 당뇨병, 운동부족, 비만, 허혈성심장질환의 가족력, 지방분포의 이상, 심전도상 좌심실비대, 스트레스, 성격 등이 허혈성심장질환의 주된 원인이 되는 죽상동맥경화증의 위험요인으로 알려져 있다(Cho et al., 1991). 일반적으로 심장질환 발생 또는 사망에 관한 예측모형을 고려할 때 로지스틱 회귀모형이 가장 많이 사용되지만, 많은 위험요인들이 있는 경우에 로지스

[1]03080 서울시 종로구 대학로 103, 서울대학교 의과대학 의료관리학연구소 연구원.
E-mail : hklim0830@gmail.com
[접수 2018년 3월 20일; 수정 2018년 4월 17일; 게재확정 2018년 4월 20일]

틱 회귀모형을 사용하게 되면 과대적합(overfitting) 문제가 발생하여 새로운 데이터에 대한 예측력이 떨어질 수 있다. 본 연구에서는 국민건강영양조사 제6기(2013-2015) 자료에서 심근경색증/협심증 발생에 영향을 미치는 여러 위험요인들 중 가장 주요한 요인을 찾기 위해 LASSO(least absolute shrinkage and selection operator)모형을 고려한다. LASSO모형은 중요하지 않은 위험요인들의 회귀계수를 정확히 0이 되도록 만들기 때문에 과대적합 문제를 해결할 수 있다.

머신러닝(machine learning)은 인공지능의 한 분야로 이미지인식, 음성인식, 강수량예측 등 다양한 분야에서 활발히 연구가 진행 중이며 최근 의학 분야에서도 위험 예측의 정확성을 향상시키기 위해 머신러닝을 이용한 연구가 활발하게 진행되고 있다: 방사선 판독(Wang, Summers, 2012), 뇌종양 탐지(Sharma, Kaur, Gujral, 2014), 위암 진단(Kourou et al., 2015), 심혈관 위험예측(Weng et al., 2017).

머신러닝 기법 중에 하나인 앙상블(ensemble)은 주어진 자료로 여러 개의 예측모형들을 만든 후 결합하여 하나의 최종 예측모형을 만드는 방법으로 예측성능을 향상시키기 위해 널리 사용되고 있다. 본 연구에서는 심근경색증/협심증 발생의 예측 성능을 높이고 동시에 강건한(robust) 변수선택을 위해서 앙상블 기법 가운데 하나인 배깅(bagging)과 LASSO를 결합한 B-LASSO모형을 제안하며 다양한 머신러닝 방법들(LASSO, CART, Random Forest, SVM)과 예측성능을 비교한다.

본 논문의 구성은 다음과 같다. 2절에서는 분석에 사용된 국민건강영양조사 자료의 변수들을 설명하고, 3절에서는 변수선택 및 예측을 위한 회귀모형으로 B-LASSO모형을 소개한다. 4절에서는 훈련자료(training data)로 모형을 구축하고 검증자료(test data)에서 모형의 예측성능을 다른 방법론들과 비교하여 제안된 모형의 우수성을 보이고, 전체 자료에서 B-LASSO모형의 예측성능 및 심근경색증/협심증 발생에 주요인으로 선별된 변수들을 살펴본다. 마지막으로 5절에서는 본 논문의 결론을 맺는다.

2. 분석 자료

본 연구에서는 국민건강영양조사 제6기(2013-2015) 복합표본설계(complex sampling design)에 의한 원시자료를 이용하여, 건강설문조사와 검진조사에 참여한 만40세~80세 성인들 중에서 조사 문항 중 심근경색증/협심증 의사진단 여부에 응답을 한 사람들을 분석 대상으로 하였다(Korean Centers for Disease Control and Prevention, 2015). 심근경색증/협심증 발생에 영향을 주는 요인으로는 인구사회학적특성, 개인과거병력, 가족병력, 건강설문 그리고 건강검진 결과 등의 위험요인이 모두 모형에 반영될 수 있도록 하였고, Rao-Scott χ^2-test를 이용하여 심근경색증/협심증 발생에 유의한 영향을 주는 변수들을 선정하여 분석모형에서 설명변수로 사용하였으며 그 변수들에 대한 정보는 Table 1에 있다.

국민건강영양조사 자료는 복합표본설계에 의해 추출된 자료이므로 층화변수, 집락변수, 가중치를 고려하여 분석을 수행하였다. 다만 자료의 가중치의 합이 우리나라 전체 인구 수이므로 가중치의 합을 조사대상자로 조정한 표준가중치를 계산하여 사용하였다. 이는 일반적인 가설검정에서 자료의 가중치의 수가 크면 정보량을 과대 인식하여 귀무가설을 쉽게 기각하는 성질을 가지고 있기 때문이다(Lee, 2017).

무응답이 있는 케이스는 분석에서 제외하였으며 분석에 사용된 자료는 전체 9332명으로 남자는 4000명, 여자는 5332명이다. 가중치를 적용한 빈도(%)분석 결과, 심근경색증/협심증 의사진단을 받은 사람은 313명(2.73%)이었고, 남자 중에 심근경색증/협심증 의사진단을 받은 사람은 151명(2.87%), 여자 중에 심근경색증/협심증 의사진단을 받은 사람은 162명(2.61%)이었다.

Table 1. Variable description

	Variable	Description
Response variable	Doctor-diagnosed myocardial infarction or angina	Yes(1), No(0)
Socio-demographic characteristics	Gender	Female(1), Male(0)
	Age	40-49(1), 50-59(2), 60-69(3), 70-80(4)
	Town	Eup/Myeon(1), Dong(0)
	Education level	Elementary graduate(1), Middle school graduate(2), High school graduate(3), University graduate and higher(4)
	Recipient of national basic living	Yes(1), No(0)
	Employment status	Unemployed(1), Employed(0)
Disease history	Doctor-diagnosed hypertension	Yes(1), No(0)
	Doctor-diagnosed dyslipidemia	Yes(1), No(0)
	Doctor-diagnosed stroke	Yes(1), No(0)
	Doctor-diagnosed diabetes	Yes(1), No(0)
Family history	Ischemic heart diseases	Yes(1), No(0)
	Stroke	Yes(1), No(0)
Health questions	Having trouble walking	Yes(1), No(0)
	Having a problem with my usual activities	Yes(1), No(0)
	Having Pain/discomfort	Yes(1), No(0)
	Having anxiety or depression	Yes(1), No(0)
	A lot of stress	Yes(1), No(0)
Current morbidity status	Hypertension	Yes(1), No(0)
	Obesity	Yes(1), No(0)
	Diabetes	Yes(1), No(0)
	Hypercholesteolemia	Yes(1), No(0)
	Anemia	Yes(1), No(0)

3. 변수선택 및 예측을 위한 회귀모형: B-LASSO

$\{(Y_i, \boldsymbol{X}_i), i=1,\cdots,n\}$를 n개 관측치의 자료라고 하자. 여기서 Y_i는 0 또는 1을 갖는 이항 반응변수이고, $\boldsymbol{X}_i=(X_{i1},\cdots,X_{ip})'$는 반응변수와 관련이 있는 p차원 설명변수들의 벡터이다. 로지스틱 회귀모형은 다음과 같다.

$$logit(\pi_i)=\log\left(\frac{\pi_i}{1-\pi_i}\right)=\sum_{j=1}^{p}\beta_j X_{ij}=\boldsymbol{X}_i'\boldsymbol{\beta},\quad i=1,\cdots,n, \tag{1}$$

여기서 $\pi_i=P(Y_i=1|\boldsymbol{X}_i)$이고, $\boldsymbol{\beta}=(\beta_1,\cdots,\beta_p)'$는 회귀계수들의 벡터이다. 많은 설명변수들이 있는 경우에 식 (1)의 로지스틱 회귀모형은 과대적합이 발생하므로 변수선택 방법으로 LASSO(least absolute shrinkage and selection operator)모형을 고려할 수 있다(Tibshirani, 1996). LASSO 추정치 $\hat{\beta}$는 다음과 같이 주어진다.

$$\hat{\beta}=\arg\min\sum_{i=1}^{n}\{-Y_i(\boldsymbol{X}_i'\boldsymbol{\beta})+\log(1+\exp(\boldsymbol{X}_i'\boldsymbol{\beta}))\}+\lambda\sum_{j=1}^{p}|\beta_j|, \tag{2}$$

여기서 λ는 음이 아닌 튜닝파라미터(tuning parameter)이다. λ가 증가함에 따라 LASSO모형은 일부 계수를 정확하게 0으로 줄이고 소수의 변수에서만 0이 아닌 회귀 계수를 갖도록 함으로써 모형의

복잡도를 감소시킨다. 하지만, 기존의 LASSO 모형은 매우 작은 회귀계수가 영으로 수렴하는 속도가 매우 빠르진 않아서 영이 아닌 회귀계수가 과대 선택되는 경향이 있고(Huang et al., 2008), 유의미한 변수를 선택할 때 과도한 편향(bias)을 일으키고 변수 선택 측면에서 일치성이 없다(Fan, Li, 2001; Leng et al., 2006).

Bach(2008)는 주어진 자료의 여러 개의 붓스트랩(bootstrap) 표본을 사용하여 LASSO 모형을 실행하고 일치된 모형 선택 결과를 얻기 위해 LASSO 모형으로부터의 영이 아닌(non-zero) 계수 추정치의 여러 세트들의 교집합을 계산하도록 제안했다. Guo et al.(2015)은 먼저 단일 출력을 생성하기 위해 여러 붓스트랩 표본의 LASSO 추정치를 평균화하는 배깅(bagging)(Breiman, 1996) 기법을 사용하고, 여러 번의 반복을 통해 0이 아닌 평균 LASSO 추정치의 교집합을 구하여 강건한(robust) 변수 선택 결과를 얻도록 하였다. Lim et al.(2017)은 이벤트 시퀀스 자료에서 불량을 조기발견하기 위한 통계적 공정관리(statistical process control)모형을 개발하고 불량예측 성능을 향상시키기 위해 배깅과 LASSO를 결합한 방법론을 제안하였다. 본 연구에서는 예측성능 향상뿐만 아니라 강건한 변수 선택을 위해 다음의 방법론을 고려한다. 이 방법론을 B-LASSO 모형이라고 하자. B-LASSO의 상세한 절차는 다음과 같다.

단계 1. 훈련자료(training data) $\{(Y_i, \boldsymbol{X}_i), i=1,\cdots,n\}$ (행렬 형식으로는 $(\boldsymbol{Y},\boldsymbol{X}) \in R^{n\times(1+p)}$)에 대하여, 자료 $\boldsymbol{D}=(\boldsymbol{Y},\boldsymbol{X})$는 두 부분으로 $(\boldsymbol{D}_1, \boldsymbol{D}_0)$로 나누어진다. 여기서 $\boldsymbol{D}_1 \in R^{n_1\times(1+p)}$와 $\boldsymbol{D}_0 \in R^{(n-n_1)\times(1+p)}$는 각각 심근경색증/협심증 있음$(Y=1)$과 없음$(Y=0)$을 포함하는 자료 행렬을 나타내고, n_1은 심근경색증/협심증이 있는 사람의 수이다. 자료 $\boldsymbol{D}$에 심근경색증/협심증이 없는 사람들이 많기 때문에, 크기 $k\times n_1$의 표본 $\boldsymbol{D}_0^s$를 $\boldsymbol{D}_0$로부터 랜덤하게 추출(random sampling)한다. 즉, $n(Y=1):n(Y=0)=1:k$. 그런 다음 $\boldsymbol{S}_1=(\boldsymbol{D}_1, \boldsymbol{D}_0^s)$을 크기가 $n_1\times(1+k)$인 붓스트랩 표본이라 놓자. 이런 방식으로 여러 붓스트랩 표본들 $\boldsymbol{S}_b, b=1,\cdots,B$을 생성한다.

단계 2. 각 붓스트랩 표본 $\boldsymbol{S}_b$에 대하여 LASSO 회귀계수 추정치를 구하며 최적의 튜닝파라미터 λ는 K-fold 교차타당성(cross-validation)을 통해 선택한다. B개의 붓스트랩 표본들로부터 구한 LASSO 회귀계수 추정치들의 평균을 각 변수별로 구한다. 즉, $\overline{\mathrm{B}_1}=(\overline{\beta_1},\cdots,\overline{\beta_p})'$.

단계 3. 평균 회귀계수가 0이 아닌 설명변수를 선택하여 $J_1=\{j, \overline{\beta_j}\neq 0\}$의 형태로 저장한다.

단계 4. 강건한 추정치를 얻기 위해 단계 1부터 단계 3까지의 과정을 R번 반복한 다음 $J_1,\cdots,J_R$의 교집합을 계산한다. 즉, $\boldsymbol{\Lambda}=\bigcap_{r=1}^{R} J_r$이 중요변수들의 집합이다.

단계 5. 평균 붓스트랩 추정치 $\overline{\mathrm{B}_1},\cdots,\overline{\mathrm{B}_R}$의 각각으로부터 중요변수들에 해당하는 LASSO 추정치들을 추출한 후 그것의 평균을 구하여 B-LASSO 회귀계수 추정치를 얻는다.

단계 6. 검증자료(test data)에 대한 예측은 $\overline{\mathrm{B}_1},\cdots,\overline{\mathrm{B}_R}$의 평균을 사용하여 수행한다.

4. 심근경색증/협심증 예측모형들의 성능 비교 및 주요 위험요인 선별

4.1. 분석방법

B-LASSO 모형의 예측성능은 LASSO, CART(classification and regression tree), Random Forest, SVM(support vector machine) 회귀모형들과 비교되었다. CART는 의사결정나무(decision tree)로 표시

될 수 있는 예측 모형으로 이진 분할(binary split)에 의해 반복적으로 데이터 셋을 분할하면서 최적의 의사결정나무를 찾는 방법이고, Random Forest는 다수의 의사결정나무를 결합하여 하나의 모형을 생성하는 앙상블 기법 중의 하나이다. SVM은 판별경계로부터 각 집단 사이의 거리(margin)를 최대로 하는 판별경계를 찾아 오분류를 최소화하기 위한 방법론이다. 모형들의 예측성능을 평가하기 위해서 분석 자료를 훈련자료 및 검증자료로 구분하였다. 훈련자료와 검증자료에서 심근경색증/협심증의 비율이 유사하도록 만들기 위해 층화 무작위 샘플링(stratified random sampling) (층: Y=0/1)을 사용하였고 훈련자료 (70%)와 검증자료 (30%)로 나누었다. 그리고 분석 자료를 훈련 및 검증 자료로 분할하는 과정을 20번 반복하여 특정 분할에 민감하지 않도록 하였다. 훈련자료는 모형을 적합하기 위해 사용되었고 검증자료는 모형의 예측성능을 평가하는데 사용되었다. B-LASSO, LASSO, Random Forest, CART 및 SVM 모형의 예측성능은 ROC(receiver operating characteristic) 곡선 아래의 영역 AUC(area under the curve)를 사용하여 평가되었다. B-LASSO 및 LASSO 모형의 경우 튜닝파라미터는 10-fold 교차타당성을 통해 선택되었고, B-LASSO 모형의 경우 B-LASSO 알고리즘 절차에 필요한 모수들은 k=3, B=10, R=20으로 설정하였다. CART에서는 의사결정나무가 과대적합 되지 않도록 post-pruning 방법을 사용하였고, SVM에서는 RBF(radial basis function) 커널을 사용하였으며 튜닝파라미터는 교차타당성을 통해 선택하였다.

4.2. 분석결과

Figure 1은 20개의 검증자료에서 5가지 예측모형(B-LASSO, LASSO, Random Forest, CART, SVM)들에 대한 AUC 값을 상자 그림(box plot)으로 나타낸 것이다. AUC는 민감도(sensitivity)와 특이도(specificity)를 동시에 고려하여 모형의 성능을 평가하기 위한 지표로서 'AUC=0.5이면 예측 성능 없음', '0.7 <= AUC < 0.8이면 허용가능', '0.8 이상이면 모형예측 성능이 우수', '0.9 이상이면 모형예측 성능이 매우 우수'라고 평가하며 간편하게 모형들의 성능을 비교할 수 있어 널리 사용되고 있다. 검증자료에서 모형들의 예측 성능을 살펴보면, 전체 예측모형 및 남자, 여자 예측모형 모두에서 B-LASSO모형이 다른 모형들에 비해 가장 우수한 예측성능을 보였다.

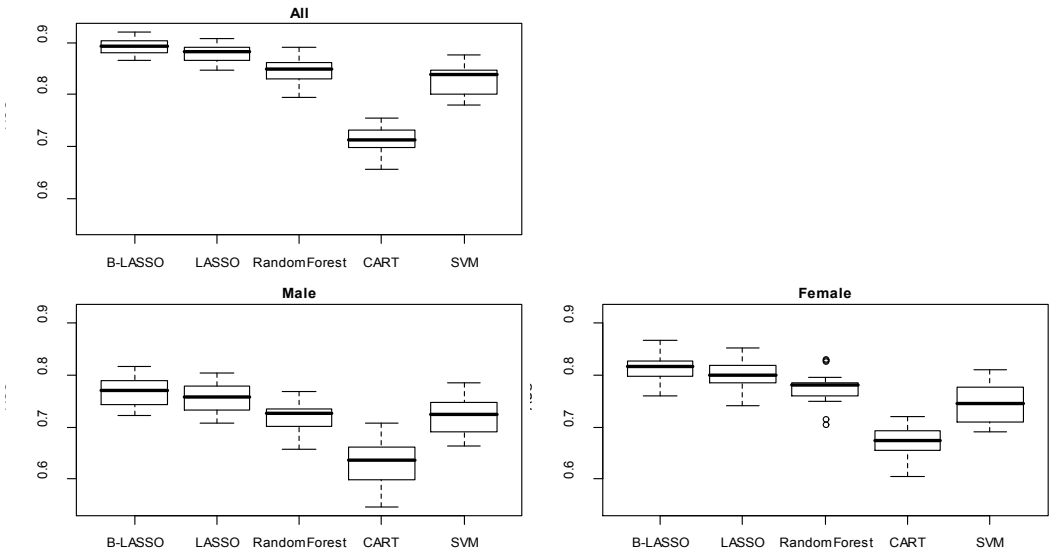

Figure 1. Box plots of the AUC values for the five models in test data sets

본 연구에서 제안하는 B-LASSO모형을 국민건강영양조사 원시자료에 적용하여 리프트 도표(lift chart), ROC 곡선, 심근경색증/협심증 발생에 영향을 미치는 주요 위험요인들에 대해 알아보았다.

Table 2의 리프트 도표는 추정된 사후확률(posterior probability)의 분위수에 따른 반응률(%Response)을 도표화 한 것으로, 리프트는 모형을 적용하지 않았을 경우에 비해 모형을 적용하였을 경우에 예측력이 향상된 정도를 나타내며, 상위 분위수에서의 리프트가 클수록 모형의 성능이 우수함을 나타낸다. 상위 10% 수준에서 리프트 값을 보면 전체 예측모형에서는 4.35, 남자와 여자의 예측모형에서는 각각 4.04와 4.82를 보여, 심근경색증/협심증 발생 예측모형을 이용하여 상위 10%를 관리할 경우 전체 집단에 비해 각각 4.35배, 4.04배, 4.82배의 예측력 향상을 기대할 수 있을 것이다. Figure 2는 원시자료에서의 ROC 곡선으로, ROC 곡선의 아래면적을 나타내는 AUC값은 전체 예측모형에서 0.819, 남자와 여자의 예측모형에서는 각각 0.801, 0.838로 상당히 좋은 예측성능을 보인다.

Table 2. Lift chart of the B-LASSO predictive model for myocardial infarction or angina

	All				Male				Female			
Grade	Freq	CR(%)	RE(%)	Lift	Freq	CR(%)	RE(%)	Lift	Freq	CR(%)	RE(%)	Lift
10%	136	43.5	14.6	4.35	61	40.4	15.3	4.04	78	48.1	14.6	4.82
20%	63	20.1	6.8	2.01	24	15.9	6.0	1.59	36	22.2	6.8	2.22
30%	45	14.4	4.8	1.44	23	15.2	5.8	1.52	16	9.9	3.0	0.99
40%	27	8.6	2.9	0.86	18	11.9	4.5	1.19	15	9.3	2.8	0.93
50%	14	4.5	1.5	0.45	11	7.3	2.8	0.73	4	2.5	0.8	0.25
60%	10	3.2	1.1	0.32	6	4.0	1.5	0.40	5	3.1	0.9	0.31
70%	10	3.2	1.1	0.32	3	2.0	0.8	0.20	5	3.1	0.9	0.31
80%	4	1.3	0.4	0.13	3	2.0	0.8	0.20	2	1.2	0.4	0.12
90%	2	0.6	0.2	0.06	1	0.7	0.3	0.07	1	0.6	0.2	0.06
100%	2	0.6	0.2	0.06	1	0.7	0.3	0.07	0	0.0	0.0	0.00

1) CR(Captured Response)(%)=(The number of myocardial infarction or angina cases in each grade/the number of myocardial infarction or angina cases in total)x100; 2) RE(Response)(%)=(The number of myocardial infarction or angina cases in each grade/the number of cases in each grade)x100; 3) Baseline lift(%)=(the number of myocardial infarction or angina cases in total/the number of cases in total)x100; 4) Lift=RE(%)/Baseline lift(%)=Incidence rate of myocardial infarction or angina in each grade/Incidence rate of myocardial infarction or angina in total.

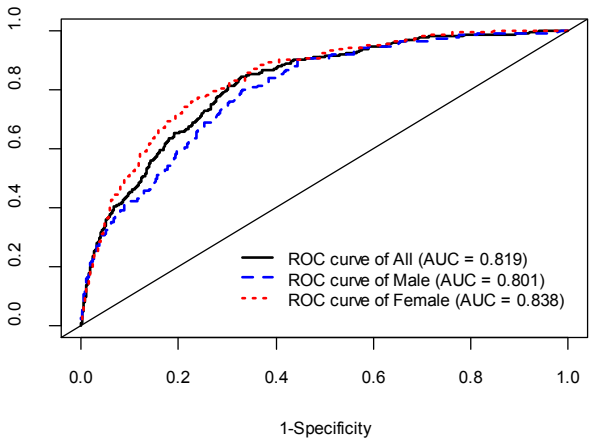

Figure 2. ROC curve of the B-LASSO predictive model for myocardial infarction or angina

Table 3. Selected important variables from the B-LASSO predictive model for myocardial infarction or angina

	All			Male			Female		
Variables	Coef	OR	Ranking	Coef	OR	Ranking	Coef	OR	Ranking
Gender (Female)	-0.496	0.61	6						
Age 50-59 (ref: 40-49)	0.636	1.89		0.333	1.40		0.146	1.16	
Age 60-69 (ref: 40-49)	1.440	4.22	1	1.200	3.32	1	0.610	1.84	1
Age 70-80 (ref: 40-49)	2.124	8.36		1.504	4.50		1.538	4.66	
Elementary graduate (ref: Univ)	0.300	1.35		0.121	1.13		0.250	1.28	
Middle school graduate (ref: Univ)	0.624	1.87	5	0.323	1.38	7	0.398	1.49	7
High school graduate (ref: Univ)	0.340	1.40		0.249	1.28		0.003	1.00	
Doctor-diagnosed hypertension	1.087	2.97	2	0.648	1.91	2	0.944	2.57	3
Doctor-diagnosed dyslipidemia	0.892	2.44	3	0.616	1.85	3	1.059	2.88	2
Doctor-diagnosed stroke	0.476	1.61	7				0.672	1.96	5
Doctor-diagnosed diabetes				0.558	1.75	4			
Ischemic heart disease family history	0.881	2.41	4	0.521	1.68	5	0.753	2.12	4
Stroke family history				0.239	1.27	9	0.236	1.27	10
Having a problem with my usual activities				0.519	1.68	6			
A lot of stress	0.397	1.49	8				0.523	1.69	6
Obesity	0.352	1.42	10	0.198	1.22	10	0.253	1.29	9
Diabetes							0.380	1.46	8
Anemia	0.379	1.46	9	0.248	1.28	8			

Table 3은 B-LASSO모형에 의해 추정된 심근경색증/협심증 발생에 영향을 미치는 주요 위험요인들 중에 중요도가 높은 상위 10개 요인을 나타낸 것이다. 전체 모형에서 인구사회학적 요인으로 성별의 경우는 여자가 남자보다 심근경색증/협심증 발생 가능성이 더 낮은 것으로 나타났다. 연령의 경우에는 나이가 많을수록 심근경색증/협심증 발생위험이 높았다. 40대에 비해 50대의 경우 심근경색증/협심증 발생 위험도가 1.89배, 60대는 4.22배, 70대 이상은 8.36배 높은 것으로 나타났고, 학력은 대졸이상에 비해 중졸의 경우에 상대적으로 심근경색증/협심증 발생 가능성이 1.87배로 높았으며 초졸은 1.35배, 고졸은 1.40배였다. 개인과거병력 요인의 경우에 고혈압 의사진단(doctor-diagnosed hypertension) 받은 경우는 받지 않은 사람들보다 심근경색증/협심증 발생 위험도가 2.97배 높았고, 이상지질혈증 의사진단(doctor-diagnosed dyslipidemia) 받은 경우는 2.44배, 뇌졸중 의사진단(doctor-diagnosed stroke) 받은 경우는 위험도가 1.61배 높았다. 가족병력의 경우에는 허혈성심장질환 가족력(ischemic heart disease family history)이 있는 사람이 그렇지 않은 사람들보다 심근경색/협심증 발생 위험도가 2.41배 높았다. 건강설문에서는 스트레스 과다, 일상활동 지장있음이 심근경색증/협심증 발생에 영향을 주었으며 특히 남자의 경우에 일상활동 지장 있는 경우가 그렇지 않은 경우보다 심근경색증/협심증 발생 위험도가 1.68배로 높았고, 여자의 경우에는 스트레스를 많이 받는 경우는 받지 않은 사람들보다 심근경색증/협심증 발생 위험도가 1.69배로 높았다. 건강검진 요인 중에는 비만(obesity), 당뇨병(diabetes), 빈혈(anemia)이 있는 사람이 심근경색증/협심증 발생 위험도가 높았으며, 남자의 경우에 특히 빈혈이 있는 사람의 위험도는 그렇지 않은 사람들보다 심근경색증/협심증 발생 위험도가 1.28배로 높았으며, 여자의 경우는 당뇨병이 있는 사람이 그렇지 않은 사람들보다 심근경색증/협심증 발생 위험도가 1.46배로 상대적으로 높았다. Table 3에 오즈비(odds ratio, OR)의 크기에 따라 심근경색증/협심증 발생에 영향을 미치는 주요 위험요인들의 중요도 순위를 매겼다. 심근경색증/협심증 발생 예측에 가장 큰 영향을 미치는 요인은 연령이었고, 다음으로

고혈압 의사진단받음, 이상지질혈증 의사진단받음, 허혈성심장질환 가족력, 학력(중졸), 성별(남자) 순으로 나타났다.

5. 결론

본 연구에서는 국민건강영양조사 제6기(2013-2015) 원시자료를 이용하여 심근경색증/협심증 발생 예측 및 주요 위험요인을 선별하기 위해 B-LASSO모형을 제안하였다. 훈련자료를 이용하여 모형을 구축하고 검증자료에서 모형들의 예측성능을 AUC를 이용하여 평가한 결과, B-LASSO모형의 예측성능이 LASSO, Random Forest, CART, SVM모형에 비하여 우수하였다. B-LASSO모형을 원시자료에 적용하여 추정된 심근경색증/협심증 발생에 영향을 미치는 주요 위험요인들로는 연령, 고혈압 의사진단받음, 이상지질혈증 의사진단받음, 허혈성심장질환 가족력, 학력(중졸), 성별(남자) 순이었으며 남자가 여자보다 심근경색증/협심증 발생 가능성이 더 높은 것으로 나타났다. 남자의 경우 당뇨병 의사진단을 받았거나 일상활동에 지장이 있거나 빈혈이 있는 사람이 그렇지 않은 사람에 비해 심근경색증/협심증 발생 위험도가 높았으며, 여자의 경우 뇌졸중 의사진단을 받았거나 스트레스를 많이 받는 사람이 그렇지 않은 사람보다 심근경색증/협심증 발생 위험도가 높았다.

국민건강영양조사 원시자료에서 심근경색증/협심증 발생에 영향을 주는 요인으로 인구사회학적 특성, 개인 과거병력, 가족병력, 건강설문, 건강검진 결과 등의 위험요인이 모두 모형에 반영될 수 있도록 하였고, Rao-Scott 카이제곱검정을 이용하여 심근경색증/협심증 발생에 유의한 영향을 주는 변수들을 선택하여 모형의 설명변수로서 사용하였다. 그러나 심근경색증/협심증 발생과 연관성이 높지만 결측 비율이 높은 변수들은 분석에서 제외하였기 때문에 유의한 변수들을 찾는데 한계가 있었다.

질병발생에 관한 연구는 비교적 긴 추적관찰 기간이 필요하나 본 연구에서 사용된 국민건강영양조사는 횡단면 자료(cross-sectional data)로서 시간 경과에 따른 변화를 살펴볼 수가 없고 질병발생 추이 및 건강위험요인들 간의 인과관계 파악이 어렵다는 단점이 있다. 향후 연구에서는 국민건강보험공단 자료를 이용하여 시간 경과에 따른 변화를 연구하고자 한다.

References

Bach, F. R. (2008). Bolasso: model consistent Lasso estimation through the bootstrap. In *Proceedings of the 25th International Conference on Machine Learning,* 33-40.

Bae, Y. H., Lee, K. W. (2016). The relative factors of cardiovascular disease among young and middle aged adults in Korea: analysis of 2012 and 2014 Korean national health and nutrition examination survey, *Journal of the Korean Data Analysis Society*, 18(4B), 2199-2214. (in Korean).

Breiman, L. (1996). Bagging predictors, *Machine Learning,* 24(2), 123-140.

Chang, K., Lim, J., Lee, S. (2018). The effect of a health keeper's cardiocerebrovascular disease prevention activity on elders' physical fitness, BMI and physiologic parameters, *Journal of the Korean Data Analysis Society*, 20(1), 487-499. (in Korean).

Cho, K. W., Park, J. C., Kang, J. C. (1991). Comparative study on the risk factors of atherosclerosis in cerebral infarction and myocardial infarction, *Korean Journal of Medicine*, 41(4), 469-480. (in Korean).

Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association,* 96(456), 1348-1360.

Guo, P., Zeng, F., Hu, X., Zhang, D., Zhu, S., Deng, Y., Hao, Y. (2015). Improved variable selection algorithm using a lasso-type penalty, with an application to assessing hepatitis B infection relevant factors in community residents, *PLoS ONE*, 10(7), e0134151.

Huang, J., Ma, S., Zhang, C.-H. (2008). The iterated lasso for high-dimensional logistic regression, *Technical Report 392,* Department of Statistics and Actuarial Science, University of Iowa.

Kim, A. L. (2015). Depression, uncertainty, patient provider relationship and compliance of health behavior of myocardial infarction patient, *Journal of the Korean Data Analysis Society*, 17(1B), 423-437. (in Korean).

Kook, H. Y., Jeong, M. H., Oh, S., Yoo, S. H., Kim, E. J., Ahn, Y., Kim, J. H., Chai, L. S., Kim, Y. J., Kim, C. J., Cho, M. C. (2014). Current trend of acute myocardial infarction in Korea (from the Korea Acute Myocardial Infarction Registry from 2006 to 2013), *The American Journal of Cardiology,* 114(12), 1817-1822.

Korean Centers for Disease Control and Prevention (2015). *Korea national health and nutrition examination: the sixth (2013-2015) guide*. (in Korean).

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction, *Computational and Structural Biotechnology Journal,* 13, 8-17.

Lee, E. (2017). Association between body shape perception and stress, depression, suicidal ideation, *Journal of the Korean Data Analysis Society,* 19(6B), 3331-3343. (in Korean).

Leng, C., Lin, Y., Wahba, G. (2006). A note on the Lasso and related procedures in model selection, *Statistica Sinica,* 16(4), 1273-1284.

Lim, H. K., Kim, Y., Kim, M.-K. (2017). Failure prediction using sequential pattern mining in the wire bonding process, *IEEE Transactions on Semiconductor Manufacturing*, 30(3), 285-292.

Sharma, K., Kaur, A., Gujral, S. (2014). Brain tumor detection based on machine learning algorithms, *International Journal of Computer Applications*, 103(1), 7-11.

Statistics Korea (2016). *Cause of death statistics in 2016*. http://kostat.go.kr/portal/korea/kor_nw/2/6/2/index.board.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society*, Series B, 58(1), 267-288.

Wang, S., Summers, R. M. (2012). Machine learning and radiology, *Medical Image Analysis,* 16(5), 933-951.

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data?, *PLoS ONE,* 12(4), e0174944.

Prediction of Myocardial Infarction/Angina and Selection of Major Risk Factors Using Machine Learning

Hwa Kyung Lim[1]

Abstract

In this study, the B-LASSO model was proposed to predict the incidence of myocardial infarction/angina using the Korea national health and nutrition examination survey (KNHANES 2013-2015) data and to select the most important risk factors among the factors affecting myocardial infarction/angina. The prediction performance of the B-LASSO model was superior to that of the LASSO, Random Forest, CART, and SVM models in test data. The major risk factors affecting myocardial infarction/angina incidence estimated from the B-LASSO model were age, doctor-diagnosed hypertension, doctor-diagnosed dyslipidemia, family history of ischemic heart disease, and gender. Men had a higher risk of myocardial infarction/angina than women. The risk of myocardial infarction/angina was higher in men who had doctor-diagnosed diabetes or anemia than in those who did not. The risk of myocardial infarction/angina was higher in women with a doctor-diagnosed stroke or a lot of stress than in those without.

Keywords : Bagging, LASSO, SVM, Random Forest, KNHANES.

[1]Researcher, Institute of Health Policy and Management, Seoul National University College of Medicine, 103 Daehak-ro, Jongno-gu, Seoul, 03080, Korea. E-mail : hklim0830@gmail.com
[Received 20 March 2018; Revised 17 April 2018; Accepted 20 April 2018]