

심혈관질환 위험 예측을 위한 비용민감 학습 모델*

Cost-Sensitive Learning for Cardio-Cerebrovascular Disease Risk Prediction

이유나¹ · 이경희² · 조완섭^{3*}

LINEWALKS Inc.¹, (주)빅데이터랩스², 충북대학교 경영정보학과³

요약

본 연구에서는 기계 학습을 사용하여 심혈관 질환 예측 모델을 제안한다. 먼저 두 집단간에 다양한 차이를 다차원분석하고 그 결과를 시각화한다. 특히, 질환과 같이 정상집단과 환자집단 간에 높은 클래스 불균형이 존재하는 경우에 대하여 민감도를 향상시킬 수 있는 비용 민감 학습을 사용하는 예측 모델을 제안한다. 본 연구에서는 대표적인 머신러닝 기술인 CART와 XGBoost를 사용하여 예측모델을 개발하고, 심혈관 질환 환자 데이터를 대상으로 예측하고 성능을 비교한다. 연구결과에 따르면 CART가 XGBoost 보다 더 높은 정확도와 특이도를 보였으며, 정확도는 약 70%~74%로 나타났다.

- 중심어 : 심혈관질환예측모델, 비용민감학습, CART, XGboost

Abstract

In this study, we propose a cardiovascular disease prediction model using machine learning. First, a multidimensional analysis of various differences between the two groups is performed and the results are visualized. In particular, we propose a predictive model using cost-sensitive learning that can improve the sensitivity for cases where there is a high class imbalance between the normal and patient groups, such as diseases. In this study, a predictive model is developed using CART and XGBoost, which are representative machine learning technologies, and prediction and performance are compared for cardiovascular disease patient data. According to the study results, CART showed higher accuracy and specificity than XGBoost, and the accuracy was about 70% to 74%.

- Keyword : Cost-Sensitive Learning, Cardio-cerebrovascular, CART, XGBoost

2021년 11월 29일 접수; 2021년 12월 09일 수정본 접수; 2021년 12월 14일 게재 확정.

* 본 연구는 참고문헌[10]을 토대로 작성되었으며, 정부(식품의약품안전처) 출연연구사업 지원을 받아 수행된 연구임(과제고 유번호: KMDF-RnD 21163수입안517-1)

† 교신저자 (wscho@chungbuk.ac.kr)

I. 서론

심혈관 질환은 전 세계적으로 암질환 다음으로 사망률이 높은 질환이다. WHO(2018) 자료에 따르면 2016년 전세계 사망자 5,690만명 중에서 1,520만명의 사망원인이 허혈성 심장질환과 뇌졸중이며, 이는 다른 질환보다 월등히 높은 수치이다[10].

심혈관 질환을 예측하기 위한 연구도 오랜 기간 진행되어 왔다. 참고문헌 [1]은 연령, 성별, 혈압, 총콜레스테롤, 혈청 고밀도지단백 콜레스테롤, 흡연, 당뇨병 유무, 좌심실비대유무를 포함한 관상동맥질환에 대한 예측모형을 제안했다. 국내에서는 [2]의 연구에서 연령, 총콜레스테롤, 수축기혈압, 흡연, 당뇨병 여부를 포함한 허혈성 심장질환 발생 위험도를 예측하는 모형을 만들었으며, [3]에서 60세 이상 지역사회 노인의 심혈관 질환을 예측하는 모형을 만들었다. 대부분의 심혈관 질환의 예측모형 연구는 통계 기법을 이용한 연구이며, 최근에는 머신러닝 알고리즘과 인공지능을 이용한 연구들도 많이 진행되고 있다 [10].

머신러닝 알고리즘으로 예측을 수행하는 경우에 데이터의 균등한 분포가 매우 중요하다. 클래스 분포의 불균형이 심한 경우, 분류기는 다수 클래스에 압도되어 소수 클래스를 올바르게 분류하지 못하기 때문이다. 하지만 질환군과 대조군을 나누는 경우처럼 실세계 데이터는 대부분 불균형이 심한 데이터가 많으며 이와 같은 경우에 소수클래스를 정확하게 예측할 수 있는 기법이 필요하다. 비용민감 학습(CSL: Cost-Sensitive Learning) 기법은 불균형 데이터에 대하여 분류기의 민감도를 향상시킬 수 있는 기법으로 알려져 있다[4].

개발된 머신러닝 모델은 향후 식품과 건강의 관련성을 예측하는 등 헬스케어 분야에서 활용할 예정이다. 헬스케어 관련 실제 데이터에는 대부

분의 경우 환자 데이터의 수가 정상인 데이터에 비해 훨씬 적은 클래스 불균형 문제가 발생하므로 제안된 기법은 유용하게 사용될 수 있을 것이다.

II. 관련연구

2.1 질병예측 연구

기존의 통계기법을 활용한 질병예측 연구는 최근 들어 대규모 빅데이터와 머신러닝 혹은 딥러닝 기법을 활용한 연구로 발전하고 있다. 심혈관 질환의 경우 사망원인이 높은 만큼 국내외에서 다양한 예측모형들이 활발하게 연구되고 있다. 대표적으로, 미국의 Framingham Study에서 개발한 모형이 있으며, 1998년 이후에는 인터넷에서 누구나 손쉽게 심장질환 10년 위험도를 계산해 볼 수 있다[5]. 국내에서는 한국인에 맞는 심혈관 질환 예측 모형을 개발하기 위해 한국인 모집단을 대상으로 한 연구 자료를 이용하고 있으며, 콕스비례위험, 생존분석 등과 같은 통계 기법을 주로 사용한다.

본 연구와 유사한 사례로 [6]의 연구결과가 있다. 이 연구에서는 표본 코호트 데이터를 사용하여 5년 이내 심혈관 질환에 걸릴 위험을 예측하였다. 데이터 불균형 문제를 해결하기 위해 컨트럴 표본을 언더 샘플링하여 총 17,420건의 데이터가 사용되었으며, 그 중 심혈관계 질환에 걸렸다고 판단되는 케이스 표본은 8,710건, 컨트럴 표본은 8,710건이다. K-Nearest Neighbors 등 7가지 머신러닝 알고리즘과 Deep Belief Network(DBN) 등 2가지 인공지능 기법을 사용하여 예측하며, DBN이 77.5%로 가장 높은 정확도를 보여주었고, 랜덤포레스트는 67.8%로 가장 낮았다. [6]의 연구는 본 연구와 마찬가지로 표본 코호트를 원천 데이터로 사용하고 있으며, 추적관찰 기간 또한 5년으로 본 연구와 유사하지만 심혈관계 질환

발생 유무를 심사평가원 데이터를 사용하여 심근 경색, 협심증, 뇌졸중, 그리고 심혈관계 죽음으로 정의하였다는 점에서 차이가 있다. 본 연구에서는 건강보험공단 데이터를 사용하고 있다. 또 다른 차이는 클래스 불균형을 처리하기 위한 언더 샘플링 방식을 적용한 후 질병예측 학습 모형을 개발하였다는 점이다. 언더 샘플링의 경우 잠재적으로 유의미한 데이터가 제거될 수 있다는 단점이 있다. 본 연구에서는 클래스 불균형이 매우 심한 데이터에서 샘플링 기법으로 클래스 불균형을 해소하는 대신에 비용민감학습 방식을 사용하여 모형의 민감도와 정확도를 향상시킨다.

[7]의 연구에서는 서울아산병원의 심혈관질환자 데이터를 대상으로 의사결정트리, 랜덤포레스트, 서포트벡터기계, 나이브베이즈 분류기법 등을 사용하여 예측모형을 개발하였으며, 정확도는 66%에서 73% 정도로 평가되었다.

2.2 비용민감학습

비용민감학습은 오분류 비용을 고려하여 학습하며, 한 클래스의 샘플을 다른 클래스로 분류하는 패널티를 인코딩하기 위해 비용 행렬을 사용한다[8]. 이진 분류 문제일 때 비용 행렬 C 의 구조는 Table 1에서 볼 수 있다. $C(i,j)$ 은 실제 클래스는 j 에서 클래스 i 로 예측한 비용을 나타내며, 양성(소수 클래스)은 1, 음성(다수 클래스)은 0으로 표기하였다. 일반적으로 소수 또는 희귀 클래스는 긍정 클래스로 간주되며, 실제 양성인 값을 음성 클래스로 잘못 분류하는 것이 실제 음성인 값을 양성 클래스로 잘못 분류하는 것보다 비용이 더 크다. 앞선 예제처럼, 암환자를 암이 아닌

라고 잘못 예측하는 비용이 암이 아닌 환자를 암 환자라고 잘못 예측하는 비용보다 더 크다.

비용 행렬이 주어지면 최소예상비용 원칙에 따라 데이터는 최소예상비용을 갖는 클래스로 분류되어야 한다. [9]에 따르면 데이터 샘플 x 을 클래스 i 로 분류하는 예상비용 $R(i|x)$ 은 다음과 같이 구할 수 있다.

$$R(i|x) = \sum_j P(j|x) C(i,j)$$

III. 연구방법

본 장에서는 연구에 사용된 데이터 세트를 설명하고, 다변량 자료의 차원축소를 위해 주성분 분석과 요인분석한 결과를 기술한 후 두 기법을 비교한다.

3.1 연구 데이터 세트

본 연구는 건강보험 데이터를 활용하여 5년 이내 심혈관 질환 발병에 영향을 미치는 위험요인을 규명하기 위해 국민건강보험공단의 건강검진 DB와 표본코호트 DB를 이용한다. 2007년부터 2013년까지 표본코호트 자격을 유지하고 있으며, 2007년 또는 2008년 국가 1차 일반건강검진을 수검한 사람 중 건강검진 기준년도 이전에 심혈관 질환으로 청구된 내역이 없는 사람을 연구대상자로 선정하였다. 또한 심혈관 질환 최초 진단일이 건강검진 년도와 같은 해인 환자와 건강검진 이후 1년 이내 인 환자는 추적 연구 기간이 짧기 때문에 분석대상에서 제외하였다. 연구 대상의 선정 및 제외 기준을 정리하면 Table 2와 같다 [10].

신규 심뇌혈관 질환 환자는 2002년부터 2013년까지 주상병 또는 부상병에 허혈성심장질환(I20-I25) 또는 뇌혈관질환(I60-I69)으로 청구된

(Table 1) 비용 행렬(Cost Matrix)

	actual negative	actual positive
predict negative	$C(0,0) = TN$	$C(0,1) = FN$
predict positive	$C(1,0) = FP$	$C(1,1) = TP$

<Table 2> 연구 대상 선정 및 제외 기준

연구대상 선정기준	기간 : 2007년 1월 1일~2013년 12월 31일 조건 : 표본코호트 자격 정보를 유지하고 있으며, 2007년 또는 2008년 1차 일반 건강검진 수검자
연구대상 제외기준	기간 : 2002년 1월 1일~2006년 12 31일, {건강검진 기준 시점 + 12개월} 이내 조건 : 심혈관 질환으로 청구된 이력이 있는 자

명세서 내역을 시간 순으로 정렬하고 해당 상병 코드가 가장 처음 청구된 시점을 ‘최초 진단일’로 추정하여, 최초 진단일이 건강검진 기저시점으로부터 5년 이내에 해당하는 사람으로 정의하였다.

<Table 3> 심혈관 질환 코드표

허혈성 심장질환명	뇌혈관 질환명
I20 협심증1	I60 거미막하출혈
I21 급성 심근경색증	I61 뇌내출혈
I22 후속심근경색증	I62 기타 비외상성 두 개 내 출혈
I23 급성 심근경색증 후 특정 현종 합병증	I63 뇌경색증
I24 기타 급성 허혈성심장질환	I64 출혈 또는 경색증으로 명시되지 않은 뇌졸중
I25 만성 허혈심장병	

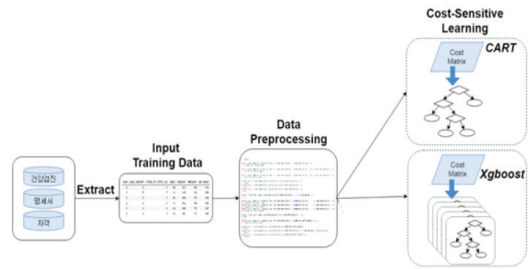
본 연구의 반응변수인 ‘심뇌혈관 질환 유무’는 신규 심뇌혈관 질환자이면 1, 아니면 0으로 코딩하였고, 건강검진의 검사항목(혈액검사, 뇨검사, 계측검사), 문진항목(본인질환력, 가족력, 흡연, 음주, 운동습관)과 인구통계정보(연령대, 소득분위, 성별, 거주지역)을 예측모델의 설명변수로 하였다.

본 연구에서는 2007년부터 2013년까지 6년간 표본코호트 자격 정보를 유지한 사람 중 2007년 또는 2008년 1차 일반건강검진 수검자 273,586명의 건강검진 데이터를 추출하였다. 선택된 사람에서 건강검진 이전에 심뇌혈관 질환이 있는 사

람과 건강검진일 이후 1년 이내 심뇌혈관 질환으로 청구된 이력이 있는 사람을 포함하여 38,286명과 건강검진데이터에 결측치가 있는 29,017명을 제외하여 206,283명을 분석대상으로 하였다 [10].

3.2 연구 내용 및 방법

전처리가 완료된 데이터를 학습용 데이터와 시험용 데이터로 나누고, 학습용 데이터로 모델을 훈련하며, 시험용 데이터로 모델을 평가한다. Fig.1은 연구방법을 도식화한 것이다. 건강검진과 명세서 및 자격 DB를 통합하여 훈련 데이터셋을 만들고, 이를 전처리한 후에 CART와 XGBoost 기법으로 예측모델을 생성한다[10].



<Fig. 1> 연구모형

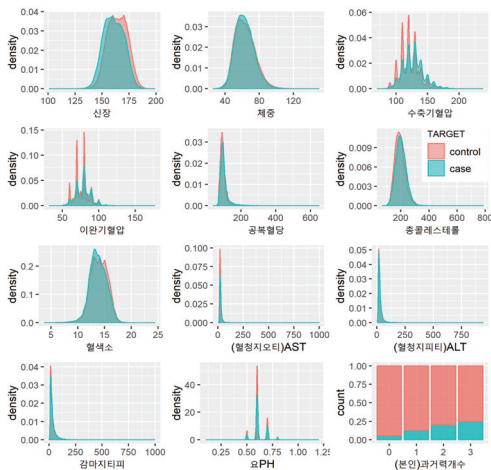
CART와 XGBoost 모델은 비용행렬을 적용하여 학습하며, 비용행렬을 통해 소수클래스(환자군)와 다수 클래스(대조군)의 가중치를 다르게 설정하여 소수 클래스에 대한 민감도를 향상시킨다. CART와 XGBoost는 트리 기반의 모형이며, 로지스틱 회귀분석과 같이 독립변수와 종속변수에 대한 직접적인 연관관계는 볼 수 없지만, 모형 내에서 소수 클래스와 다수 클래스 분포의 균형을 조절하여 다수 클래스의 영향력을 완화시킬 수 있는 장점이 있다. 부스팅 기법은 이전 분류기에서 잘못 분류한 데이터에 가중치를 더 부여하여 약한 분류기를 더 강한 분류기로 만드는 방법이며, 정확도를 향상시킬 수 있다는 장점이 있다.

IV. 분석결과

4.1 환자군-대조군의 분포 시각화

모형을 학습하기 전 분석에 사용하는 변수별 분포를 커널밀도곡선으로 연속형 변수를 시각화한 것이다. 그래프에서 심혈관 질환 환자군은 녹색, 대조군은 적색으로 구분된다. 본 연구에서 사용한 데이터는 환자군과 대조군의 빈도수 차이가 크므로 환자군과 대조군의 분포 차이를 쉽게 인지할 수 있도록 커널밀도곡선으로 표현하였다. 대부분의 변수는 환자군과 대조군의 분포가 비슷하며, 신장, 혈색소, 수축기혈압과 이완기혈압에서 차이가 있다[10].

범주형 변수는 Fig.2의 가장 오른쪽 하단 그래프와 같이 비율누적막대그래프를 이용하여 확인하였으며, 환자군이 요당과 요단백, 흡연기간, 본인질환력 변수가 환자군에서 높았다. 또한, 연령이 높아질수록 환자군의 비율이 월등히 증가했으며, 소득이 낮은 그룹과 높은 그룹의 환자군 비율이 다른 그룹에 비하여 높게 나타났다[10].



〈Fig. 2〉 커널밀도곡선과 비율누적막대그래프

4.2 심혈관 질환 예측 모형과 평가

본 연구에서는 비용 행렬을 적용한 CART 및 XGBoost 기반의 모형을 사용한 딥러닝 기반의 예측 모형을 제시한다.

먼저 비용민감학습 방식을 적용한 CART 알고리즘을 사용하기 위해, 모형을 적합하기 이전 오분류에 대한 비용을 고려하여 비용 행렬을 생성하였다. Ling(2008)에 의하면 소수클래스에 대한 오분류 시 비용 행렬에 부여하는 값은 $N(\text{다수클래스})/N(\text{소수클래스})$ 로 한다. 본 연구에서는 대조군의 수를 환자군의 수로 나눈 값은 13이다. 비용을 크게 하면 민감도는 올라가지만 반대로 특이도는 떨어진다. 민감도와 특이도의 트레이드 오프(Trade-Off) 때문에 적절한 비용을 부과하는 것이 중요하다. 일반적으로 비용 행렬에 대한 비용은 주로 연구자의 주관 또는 경험치에 따른다. 본 연구에서 비용행렬은 Table 4와 같이 연구자의 주관에 따라 민감도는 조금 내려가더라도 정확도와 ROC가 향상될 수 있도록 값을 조정하여 대조군을 환자군으로 오분류하는 비용을 상대적으로 높은 10으로 정하였다[10].

〈Table 4〉 환자군과 대조군의 비용 행렬

	실제 대조군	실제 환자군
예측 대조군	0	1
예측 환자군	10	0

CART 모형은 10번의 교차검증을 통해 학습하였으며, 환자군은 환자군으로, 대조군은 대조군으로 정확하게 예측한 비율인 정확도는 74.2%, 실제 환자군 중에서 정확하게 환자군으로 예측한 비율인 민감도는 61.8%, 실제 대조군 중에서 대조군으로 정확하게 예측한 비율인 특이도는 75.1%로 나타났다. CART 모형에 시험용 데이터를 적용하여 구한 혼동 행렬은 Table 5와 같다.

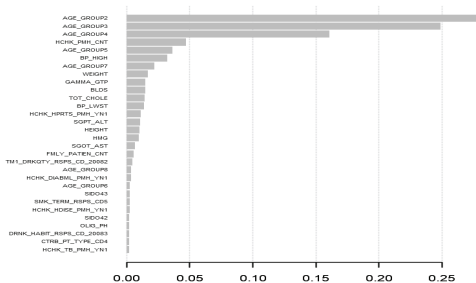
〈Table 5〉 CART 모형의 혼동 행렬

예측값		실제값	
		대조군	환자군
대조군	대조군	28,836	1,100
	환자군	9,539	1,780

XGBoost 알고리즘으로 모형을 학습하기 위해서는 파라미터를 연구자가 직접 조정해야 하며, 일반 파라미터, 부스터 파라미터와 태스크 파라미터에 대한 정의가 필요하다. 일반 파라미터는 어떤 부스터 구조를 사용할지 정하며, 트리 기반의 모형과 선형 함수 기반의 모형을 사용할 수 있다. 부스터 파라미터는 모형에 대한 부스팅 방법을 설정하며, 마지막으로 태스크 파라미터에서는 학습 과정을 위한 파라미터를 설정하며 목적 함수와 모형의 평가 함수를 설정할 수 있다.

본 연구에서는 그리드 서치를 사용하여 가장 성능이 좋은 파라미터 값들의 조합을 찾았다. 그리드 서치 결과, 트리개수는 100개, 트리의 최대 깊이는 4, 학습률(Eta)은 0.1, 감마는 5로 조정하였다. 또한, 소수클래스의 오분류에 대한 높은 비용을 부과하기 위해 CART 모형과 동일한 비용 행렬(Table 4)을 사용하였다[10].

그리드 서치로 찾아낸 파라미터의 값으로 모형을 튜닝한 결과, 학습 모형의 변수 중요도는 Fig. 3과 같다. 상위 30개의 변수 중 가장 많은 영향력을 미친 변수는 연령대와 (본인) 과거력 개수이며, 수축기혈압, 체중, 감마지티피, 공복혈



〈Fig. 3〉 XGBoost 변수중요도

당, 총콜레스테롤 등 순으로 높게 나왔으며, 본 연구에서 사용한 본인 과거력 개수의 중요도가 높게 나와 만성질환의 개수가 심혈관 질환 발생에 유의미한 영향을 미치는 것으로 나타났다.

XGBoost 모형의 학습 결과 정확도는 73.9%, 실제 환자군은 환자군으로 정확하게 예측한 비율인 민감도는 64.5% 특이도는 74.6% 이다. CART 모형과 비교하여 민감도는 상승하였고 정확도와 특이도는 약간 하락하였다. XGBoost 모형에 시험용 데이터를 적용하였을 때 혼동행렬은 Table 6과 같다.

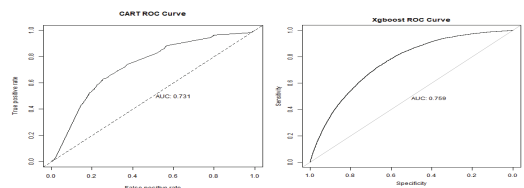
〈Table 6〉 XGBoost 모형의 혼동 행렬

예측값		실제값	
		대조군	환자군
대조군	대조군	28,630	1,021
	환자군	9,745	1,859

두 모형에 대한 평가는 Table 7과 같다. CART 알고리즘 모형의 정확도는 74.2%, 민감도는 61.8%, 특이도는 75.1%이며 AUC는 0.731로 본 연구에서 개발한 모형은 설명력이 있다고 볼 수 있다. XGBoost 모형의 정확도는 73.9%, 민감도는 64.5%, 특이도는 74.6%이며, AUC는 0.759로 Xgboost의 모형 또한 설명력이 있으며, 두 모형

〈Table 7〉 평가 지표 비교

모형	정확도	민감도	특이도
CART	74.2%	61.8%	75.1%
Xgboost	73.9%	64.5%	74.6%



〈Fig. 4〉 CART와 XGBoost 모형의 ROC Curve

을 비교해보면, 정확도와 특이도는 CART 모형이 더 높으며, 민감도와 AUC는 XGBoost 모형이 더 높았다(Fig.4).

본 논문에서 비용민감학습을 적용한 CART 및 XGBoost 모형은 전체적으로 70% 이상의 정확도를 갖는다.

당 연구팀에서는 본 연구의 결과를 사용하여 식품과 건강 데이터부터 관련성을 예측하는 문제에 적용하고 있으며, 실제 환경의 데이터에서 클래스 불균형 문제를 해결하는데 도움이 될 것이다.

V. 결론

본 논문에서는 건강보험 데이터를 이용하여 5년 이내 심혈관 중증질환인 허혈성 심장질환(I20-I25)과 뇌혈관질환(I60-69) 발병을 예측하는 모형을 구축하였다. 총 41개의 변수로 데이터셋을 구성하여 CART와 XGBoost를 적용한 모형을 생성하였으며, 평가결과 70% 이상의 정확도를 나타냈다.

본 연구의 한계와 향후 연구 과제는 다음과 같다. 본 연구는 의학적 사전지식보다는 건강보험 데이터를 기반으로 심혈관 질환 예측 모형을 개발하였으며, 연구결과는 심혈관 질환 전문가와 상의하여 더욱 신뢰성 있는 모형을 개발하는 것이 필요하다[10].

본 연구에서 심혈관 질환 예측모형에 적용한 XGBoost와 같은 앙상블 기법은 로지스틱 회귀분석과 같이 독립변수가 종속변수에 미치는 영향을 직접적으로 구하기 어려운 점이 있다. 따라서 심혈관 질환 발생과 예측변수들의 관계를 명확히 설명해줄 수 있는 분류기법을 고안할 필요가 있다. 또한, 비용민감학습 방식을 사용한 CART와 XGBoost 두 모형은 트리 기반으로 추후 연구에서는 다른 종류의 분류기를 적용하여 비교해 볼 필요가 있다.

참 고 문 헌

- [1] Bosma, H., Peter, R., Siegrist, J., & Marmot, M. (1998). Two alternative job stress models and the risk of coronary heart disease. *American Journal of Public Health*, 88(1), 68 - 74.
- [2] 지선하, 송지원, 조흥근, 김상연, 장양수, & 김정희.(2004), “허혈성심질환 발생예측모형 (health risk appraisal) 개발 연구,” 한국지질 동맥경화학 회지, 14(2), 153-168.
- [3] 홍새미, 변해원, 김정순 and 문순희. (2015). 지역 사회 거주 노인의 심뇌혈관질환 예측 모형. *예술 인문사회 융합 멀티미디어 논문지*, 5(1), 37-46.
- [4] Chawla, N. V., Japkowicz, N., & Kotcz, A.(2004), “Special issue on learning from imbalanced data sets,” *ACM Sigkdd Explorations Newsletter*, 6(1), 1-6.
- [5] Burke, G. M., Genuardi, M., Shappell, H., D’Agostino Sr, R. B., & Magnani, J. W. (2017). Temporal associations between smoking and cardiovascular disease, 1971 to 2006 (from the Framingham Heart Study). *The American journal of cardiology*, 120(10), 1787-1791.
- [6] 이동훈, 김민호, 김영원, 한영웅, 임명은, 김대희, 정호열, 최재훈.(2017), “Deep Belief Network를 이용한 심혈관계 질환 위험 예측,” 제 29회 영상 처리 및 이해에 관한 워크샵 IPIU 2017, Jeju, 1-3.
- [7] 강길원 등(2019), 빅데이터 연계를 통한 심혈관 질환 예측 모형 개발, 보건복지부 보건의료기술 연구개발사업 보고서, 충북대학교 산학협력단.
- [8] Elkan, C.(2001), “The foundations of cost-sensitive learning,” *International Joint Conference on Artificial Intelligence*. Lawrence Erlbaum Associates Ltd, 17(1), 973-978.
- [9] Ling, C. X., & Sheng, V. S.(2008), “Cost-Sensitive Learning and the Class Imbalance Problem; 2011,”

Encyclopedia of Machine Learning, berlin, Springer.

- [10] 이유나(2019), Cost-Sensitive Learning을 활용한 심뇌혈관 질환 발생 예측 모형 개발, 충북대학교 석사학위논문.

저자 소개



이유나(Yu Na Lee)

- 2017년 : 충북대 통계학과 (학사)
- 2019년 : 충북대 빅데이터학과 석사
- 현소속 : LINEWALKS Inc.
- 관심분야 : 빅데이터분석



이경희(Kyung-Hee Lee)

- 2004년 : 충북대 컴퓨터과학과 (박사)
- 2016년~2020년 : 충북대 빅데이터학과 초빙교수
- 2020년~현재 : (주)빅데이터랩스
- 관심분야 : 빅데이터, 알고리즘



조완섭(Wan-Sup Cho)

- 1987년: KAIST 전산학과 (박사)
- 1996년~현재: 충북대학교 교수
- 관심분야 : 빅데이터, 블록체인, 빅데이터거버넌스