머신러닝을 활용한 당뇨병 환자의 관상 동맥 질환 모델 개발

권다운, 장소현, 권보영, 박준영*
울산대학교, 서울과학기술대학교, 고려대학교, 아주대학교*

dobylive01@gmail.com, ish30hwt@seoultech.ac.kr, kby2009@korea.ac.kr, crinexk@gmail.com*

Predictive Study on Coronary Artery Disease in Diabetic Patients using Machine Learning

Kwon Da Woon, Jang So Hyun, Kwon Bo Young, Park June Young*
Ulsan Univ, SeoulTech Univ, Korea Univ, Ajou Univ*

요 약

본 논문은 MIMIC-IV 데이터베이스를 이용하여 당뇨병 환자의 관상 동맥 질환을 예측하는 BalancedBagging 기반의 머신러닝 모델을 설계하고, Feature Selection 을 통해 관상 동맥 질환의 발병에 가장 큰 영향을 미치는 feature 가 무엇인지 분석하였다.

Ⅰ. 서론

심장질환은 2021 년 통계청 사망원인통계에 따르면 사망원인 중 2 위를 차지하였다. 2000 년대 이후 심장 질환에 의한 사망률은 지속해서 상승 추세를 보이고 있다. 특히, 관상 동맥 질환은 다른 심장 질환과의 높은 연관성을 갖고 있으며, 발병 원인은 다양한 요인에 의해 영향을 받는다고 알려져있다.

흡연, 비만과 같은 행동적 요인은 관상 동맥 질환의 발병에 분명한 영향을 주는 것으로 알려져 있으나, 동반질환과 같은 의학적 요인에 따른 발병 예측은 복잡성을 가진다. 특히, 동반질환 중 하나인 당뇨병은 관상 동맥과의 높은 연관성을 보인다.

본 논문에서는 당뇨병 환자 데이터를 기반으로 관상 동맥 질환 발병 위험성 예측을 위한 머신러닝 모델을 제안한다. 또한, 발병에 큰 영향을 미치는 특징들에 대한 탐색적 분석 결과를 포함한다.

Ⅱ. 본론

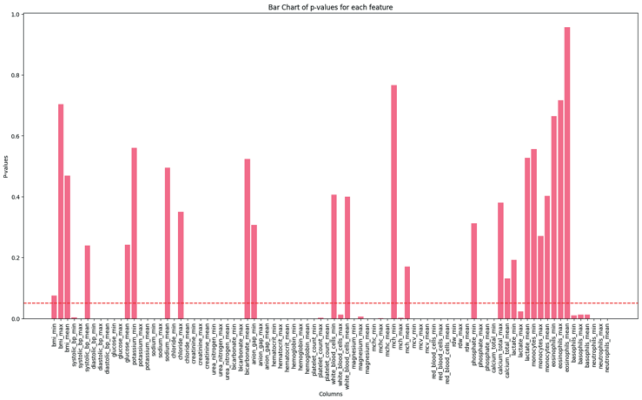
2.1 MIMIC-IV 데이터셋 추출

모델 훈련을 위해 사용된 데이터셋은 MIMIC-IV 데이터베이스에서 추출하여 사용하였다. 본 연구에서는 진단 데이터에서 ICD-9, ICD-10 같은 국제질병분류를 통해 당뇨병 환자들을 필터링 하였으며, 관상 동맥 질환 환자인지의 여부 또한 국제질병분류를 통해 추출하였다. 인구통계학 데이터에서는 환자의 나이, 성별, 인종과 같은 개인적인 정보를 나타내는 데이터를 가져왔으며, Online Medical Record 에서는 환자의 혈압, 체중, 신장, BMI 를 나타내는 데이터 중 혈압과 BMI 를 추출했다. 특히, 혈압은 관상 동맥 질환의 위험요소이므로 수축기 혈압과 이완기 혈압으로 분류하는 과정을 거쳤다. 또한, 표본 검사 (lab events) 데이터에서는 20 개의 검사 결과를 추출했다. 그러나 검사 결과는 관상 동맥 질환 진단에 있어 선후 관계를 파악하는데 모호성이 있었기 때문에 최솟값, 최댓값, 평균으로 나누어 데이터를 구성하였다. 마지막으로, 약물 복용 데이터는 횟수보다 복용 여부에 초점을 두어 이진 데이터로 구성하였다. 약물 복용 데이터는 관련 연구 자료를 기반으로 당뇨병, 관상 동맥 질환과 연관된 약물들을 채택하였다. 결과적으로, 5 가지 종류 (진단, 인구통계학, OMR, 표본 검사, 약물 처방)를 통해 추출한 데이터셋은 39 개의 feature 와 19,400 명의 환자를 포함하고 있으며, 이 중 3,239 명이 관상 동맥 질환을 진단받은 환자이다.

2.2 데이터 전처리

2.2.1 열 제거

데이터의 신뢰성을 확인하기 위하여, CAD 환자군와 non-CAD 환자군으로 분류한 데이터의 각 열에 대하여 Student's T-Test 를 수행하여 p-value 을 구해 값이 0.05 이상인 경우 데이터에서 제외시켰다.



<그림 1>

위 전처리에 따라 10 개의 feature 가 제거되었다. 또한 추출한 데이터에서 결측치가 30% 이상인 칼럼은 학습 데이터셋에서 제외시키기로 하였다. 그 결과 age, sex, race 를 포함하여 총 38 개의 feautre 를 학습에 사용하였다.

2.2.2 인코딩

학습 데이터셋을 구성하기 위해 범주형 변수의 인코딩을 진행하였다. 성별 등 이진 범주형 변수의 경우에는 이진 인코딩을 진행하였다. Race 데이터의 경우, 30여개 이상의 범주로 구성되어 있다. 따라서 원-핫 인코딩을 사용할 경우 데이터의 차원이 지나치게 커질 가능성이 있어 해당 카테고리에 해당하는 데이터가 전체 데이터에서 차지하는 비율로 인코딩하는 방식인 빈도 인코딩을 수행했다.

2.3 모델

2.3.1 모델 훈련 및 성능 비교

위 전처리를 수행한 데이터를 활용한 예측 모델로 Random Forest, RUSBoost, EasyEnsemble, BalancedRandomForest, BalancedBagging 총 5 가지 모델을 사용하였다. Random Forest 모델을 제외한 나머지 네 모델은 모두 불균형한 클래스 분포를 가진 데이터셋에서 분류 성능을 향상시키기 위해 사용되는 모델로, 본 실험 데이터셋에 적합한 모델이다. 또한 오버샘플링 기법 사용 시 분류 성능을 높일 수는 있지만, 의료 데이터 특성상 데이터의 왜곡이 생길 수 있으므로 언더샘플링 기반의 모델을 사용하였다. 모델 성능 평가는 accuracy 와 AUC score 를 이용했고, ROC curve 를 시각화하여 확인했다. Random Forest 는 다수의 결정 트리들을 학습하는 앙상블 학습 방법으로, 훈련 과정에서 구성한 다수의 결정 트리로부터 분류 또는 평균 예측치를 출력함으로써 동작하는 모델이다. 다른 모델들은 학습 과정에서 자체적으로 언더샘플링을 시행하기 때문에, 해당 모델에서는 데이터셋에 직접 언더샘플링을 시행한 후 학습시켰다. <표 1>은 5 개 모델에서 트리의 개수를 100 개로 통일한 후 성능 비교를 한 결과이다.

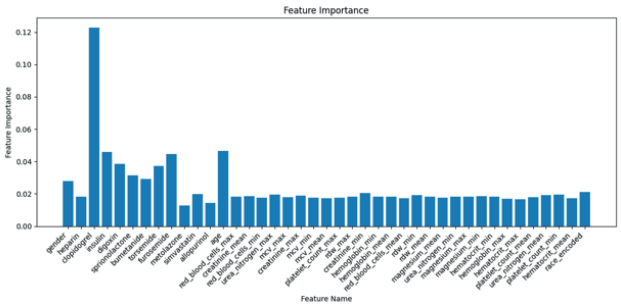
모델	Accuracy	AUC
RF	0.6969	0.7692
RUSBoost	0.7101	0.7585
EasyEnsemble	0.7072	0.7724
BalancedRF	0.6987	0.7719
BalancedBagging	**0.7704**	**0.7714**

<표 1>

결과를 살펴보면 accuracy 가 나머지 모델들보다 높고, AUC 도 높은 점수를 보인 BalancedBaggin 모델의 하이퍼 파라미터를 튜닝해 모델 성능을 높인다. n_estimators=400, max_features=0.75, sampling_strategy=0.로 설정하였을 때 accuracy 는 0.8116, AUC 는 0.7756 이라는 향상된 결과를 얻었다. 이후 내부 분류기인 DecisionTreeClassifier 의 하이퍼파라미터를 조정하여 accuracy 가 0.8188, AUC 가 0.7826 이라는 가장 좋은 성능을 보였다.

2.3.2 Feature Selection

이후 BalancedBagging 의 성능을 더욱 높이기 위하여 feature selection 을 진행했다. Feature importance 와 SHAP value 값을 모두 추출하여 상위 값들을 복합적으로 고려해 feature 를 선택했다. <그림 2>는 feature importance 를 시각화한 것이다.



<그림 2>

Feature importance 상위 4 개 값인 clopidogrel, age, furosemide, insulin 을 고정으로 두고 SHAP value 상위 10 개 feature 의 조합을 다양하게 하여 성능 확인을 진행했다. 결과적으로 모든 feature 를 사용하는 것보다 clopidogrel, age, insulin, furosemide, gender, urea_nitrogen_max, hematocrit_mean 의 총 7 가지 feature 를 사용했을 때 accuracy 가 0.8273, AUC 가 0.7828 이라는 더 나은 결과를 얻을 수 있었다.

Ⅲ. 결론

본 논문에서는 MIMIC-IV 데이터베이스를 활용하여 당뇨병 환자의 CAD 동반 질환을 예측하기 위한 연구를 진행하였다. MIMIC 데이터베이스를 구축하여 연구의 방향과 일치하는 데이터셋을 추출하고 정제 과정을 통해 관련성이 부족한 feature 를 제거하였다. 사용한 여러 분류 모델 중에서 불균형 데이터셋에 특화된 BalancedBagging 모델이 가장 좋은 성능을 보였다. 또한 하이퍼파라미터 튜닝을 통해 추가적으로 성능을 개선하고자 하였다. Feature importance 를 추출한 결과, 항혈소판제 (Clopidogrel), 나이, 이뇨제 (Furosemide) 등의 순으로 나왔으며 약물이 관상 동맥 질환 발병과 많은 연관성을 가진다는 것을 알 수 있었다. Feature selection 과정을 거쳐 전체 중 7 개 feature 만을 선택하여 훈련시켰을 때 최상의 예측 성능을 보임을 알 수 있었다.

본 연구의 한계점은 MIMIC-IV 데이터셋의 결측치 문제로 더 다양한 feature 를 사용하지 못했으며, LAB 표본 검사와 CAD 진단 사이의 인과관계가 모호하다는 점이 있었다. 또한 일반 환자 수가 CAD 진단 환자보다 압도적으로 많은 불균형 클래스 데이터였기에, 언더샘플링 기반 모델로 성능을 개선하는 데 한계가 존재한다.

참 고 문 헌

[1] Katarzyna Nabrdalik, "Machine Learning Predicts Cardiovascular Events in Patients With Diabetes: The Silesia Diabetes-Heart Project", 2023, ScienceDirect, (https://doi.org/10.1016/j.cpcardiol.2023.101694)

[2] Chris Seiffert, “RUSBoost: Improving Classification Performance when Training Data is Skewed”, 2008, IEEE Xplore (https://doi.org/10.1109/ICPR.2008.4761297)