당뇨병 발생 예측을 위한 다층 스태킹 앙상블 모델 구축 기법

성아영[1], 윤소현[1], 강수연[1], 김건우[1*]
[1]경상국립대학교 컴퓨터과학부
achieve00@gnu.ac.kr, yunsoyun9426@gnu.ac.kr, sillon@gnu.ac.kr, gunwoo.kim@gnu.ac.kr

Automatic Multi-layer Stacking Ensemble Generation Technique for Predicting Diabetes Mellitus Incidence

Ayeong Seong[1], Sohyun Yun[1], Suyeon Kang[1], Gun-Woo Kim[1*]
[1]School of Computer Science, Gyeongsang National University

요 약

최근 현대인의 식습관 및 고령화로 인해 당뇨병 환자의 수가 연간 증가하고 있다. 따라서 현재는 아직 당뇨병이 발생하지 않았더라도 미래에 발생할 가능성 예측의 중요성이 커지고 있다. 기존의 당뇨병 발생 여부 진단 연구는 회귀 분석과 같은 단일 모델을 사용하여 수행된다. 그러나 당뇨병에 영향을 미치는 변수들은 복잡하게 얽혀있어 단일 모델만으로는 패턴을 충분히 학습하기 어렵다. 본 논문에서는 데이터에 적합하게 자동으로 다층 스태킹 앙상블 모델을 구성하는 알고리즘을 이용한 다층 스태킹 앙상블 모델을 제안한다. 제안하는 방법은 성능이 높은 모델들을 기준으로 층을 쌓으며 모델을 구성하며 실험 결과 다른 자동 기계학습 라이브러리와 비교해 F1 score 기준으로 최대 12.89%p의 성능 향상을 보였다.

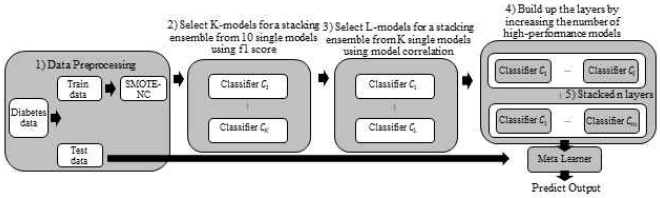
1. 서론

대한당뇨병학회가 발표한 2022년 한국 당뇨병 팩트시트(Diabetes Fact Sheet in Korea 2022)에 따르면 30세 이상 성인 약 10명 중 4명이 당뇨병 전단계에 해당한다. 그에 따라 당뇨병 발생을 예측하는 모델을 개발하고자 하는 연구가 늘어나고 있다. [1], [2] 기존의 당뇨병 발생 여부 진단 연구는 로지스틱 회귀와 같은 간단한 단일 모델을 이용했다. 하지만 다양한 특성으로 이루어져 있고, 이들 간의 복잡한 관계가 형성되는 의료 데이터의 특성상 단일 모델만으로는 데이터의 패턴을 충분히 학습하기 어렵다. 하지만 의료분야에 대한 도메인 지식만으로는 다양한 모델을 이용하는 기계학습 기법을 사용하기에는 어려움이 존재한다. [3]

따라서 본 연구는 데이터에 맞춰 자동으로 다층 스태킹 앙상블 모델을 구성하는 알고리즘을 제안하며 이를 기반으로 스태킹 앙상블 모델을 구축한다.

2. 다층 스태킹 앙상블

본 연구는 국내 코호트 데이터인 한국인유전체역학조사 자료를 활용했다. 이 데이터는 독립변수 118개, 종속변수 1개로 총 119개의 열과 100*30개의 행으로 구성되어 있다. 결측값 처리를 위해 KNN 알고리즘과 평균화 기법을 활용하였으며, 결측값이 50% 이상일 때 해당 열을 삭제했다. 이를 통해 일차적으로 55가지의 변수를 선택했다. 이후 Step-wise Feature Selection을 이용해 사용할 변수를 축소했다. 학습 방법이 다른 네 가지 분류기를 이용해 이 중 세 가지 분류기에서 공통으로 선택된 변수를 선정했다. 최종적으로 선택된 변수는 당뇨병 가족력 유무, 성별 등과 같은 기초자료 4가지, 혈당과 당화혈색소와 같은 당뇨 검사 지표 3가지, 기저 인슐린, 중성지방, 고혈압 여부 등과 같은 검사자료 20가지 총 27가지 변수가 학습에 사용됐다.



(그림 1) Proposed Process

그림 1과 같은 과정을 수행해 적응형 다층 스태킹 앙상블 모델을 도출하고 학습 및 예측을 진행하였다. 본 연구에 사용한 데이터는 종속변수의 클래스가 8:2 정도의 비율 차이가 나는 불균형 데이터이다. SMOTE-NC 알고리즘을 통해 범주형 데이터를 고려하며 학습 데이터를 증강해 데이터 균형을 맞추며 불균형 데이터 문제를 해결하였다.

일반적으로 자주 사용되는 분류 모델 중 10가지 (Naïve Bayes, Logistic regression, SGD Classifier,

* 교신저자(Corresponding Author)

KNN, SVM, Decision tree, Random forest, lightGBM, XGBoost, Catboost)를 선정했다. 모델은 거리기반, 확률기반, 트리기반 같은 학습 방법을 기준으로 선정했다. 이후 3단계를 거쳐 10가지의 분류 모델 중 다층 스태킹 앙상블에 사용할 모델을 선택하는 과정을 자동화했다. 10개의 모델 중 K-fold 교차 검증을 진행해 얻은 F1-score를 기준으로 성능 비교한다. 성능이 낮은 모델을 일차적으로 제외하는 과정을 통해 K개의 모델을 선택한다. 남은 모델들의 예측값을 비교해 모델들 간의 상관관계를 분석하고, 0.7 이상의 높은 상관관계를 보이는 모델 쌍에서 성능이 낮은 모델을 추가로 제외한다. 최종적으로 스태킹 앙상블을 구성하는데 이용할 L개의 모델을 선택한다.

Algorithm 1: Stacking Multi-layer Ensemble Models

Input: Selected Models M, Training set X, Y
Output: best layers

for $l = 1$ **to** L **do**
  Randomly split data into (X_{train}, Y_{train}), (X_{val}, Y_{val});
  for $num = 1$ **to** $len(M)$ **do**
    $one\_layer \leftarrow M[:num]$;
    Make a stacking ensemble classifier with one_layer;
    Train the stacking ensemble classifier on X_{train}, Y_{train};
    if $f1\_score > best\_f1$ **then**
      $best\_f1 \leftarrow f1\_score$;
      $best\_layer \leftarrow one\_layer$;
    end
    else
      $flag+ = 1$;
    end
    if $flag > threshold$ **then**
      break;
    end
  end
  $X \leftarrow concatenate(X, \hat{Y})$;
  if $best\_layer == best\_layers$ **then**
    break;
  end
  $best\_layers \leftarrow concatenate(best\_layers, best\_layer)$;
end
return $best\_layer$;

(그림 2) Stacking Multi-layer Ensemble Models

마지막으로 남은 L개의 모델을 이용해 다층 스태킹 앙상블을 자동으로 구성하는 알고리즘은 그림 2와 같다. 선택된 모델을 성능 순으로 정렬한 후, 모델의 개수를 점진적으로 하나씩 늘려가며 최적의 성능을 가지는 조합을 찾는다. 조합 탐색 과정에서 모델의 수를 추가해도 성능 향상이 없다면 반복을 중단하고 해당 조합을 최적의 단일 층으로 선택한다. 이후 이전 층에서 생성된 출력을 입력으로 사용하여 앞의 단계를 반복하며 최적의 성능을 가지는 조합을 찾고 층을 늘려나간다. 이전 층과 현재 층의 모델 구성이 같다면 반복을 종료한다.

본 연구에서 생성하고자 하는 모델은 다층 스태킹 앙상블이다. 이에 따라 모델의 복잡도가 상당히 증가할 것이 예상된다. 따라서 모델이 데이터에 과적합 되는 것을 방지하고자 메타 학습기로는 Logistic regression을 선택했다.

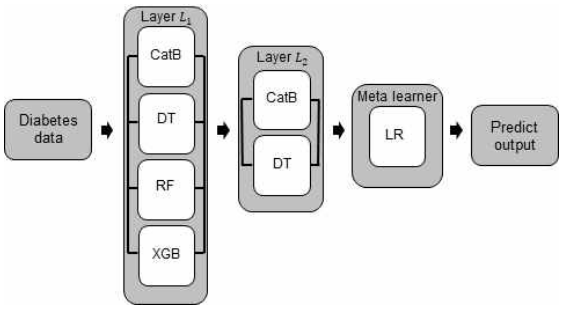
3. 실험

학습 데이터와 평가 데이터는 데이터의 클래스 분포를 고려하며 7:3의 비율로 나누어 실험을 진행했다. 표 1은 전체 데이터에 대한 요약이다.

<표 1> Summary of Train and Test sets

Class	Train set	Test set	Total
정상	6004	2573	8577
당뇨	1017	436	1453

사용한 데이터의 당뇨 발생 비율이 불균형한 데이터임을 고려하여 Accuracy 외에도 F1-score와 ROC AUC를 추가로 모델 평가 지표로 이용했다.



(그림 3) stacking ensemble model using proposed method

실험 결과 제안하는 방법을 통해 도출된 다층 스태킹 앙상블 모델은 그림 3과 같다. 첫 번째 층은 CatBoost, Decision tree, Random forest, XGBoost 두 번째 층은 CatBoost, Decision tree, 메타 학습기는 Logistic regression을 사용했다.

성능 비교 실험에는 자동 기계학습 라이브러리 중 TPOT, Pycaret, Autogluon 세 가지를 사용해 진행했다. 표 2는 각 라이브러리에서 도출된 최적의 조합 및 성능 비교이다.

<표 2> Performance comparison with other methods

Methods	Accuracy	F1-score	ROC AUC
TPOT (ExtraTree)	0.752	0.416	0.698
Pycaret (CatB, XGB, LightGBM)	0.863	0.3439	0.609
Autogluon (WeightedEnsemble_L2)	**0.867**	0.3789	0.625
Proposed Model	0.851	**0.4728**	**0.700**

4. 결론

본 연구에서는 당뇨병 발생을 예측하기 위해 다층 스태킹 앙상블 모델을 제안했다. 모델은 F1 score 기준으로 최대 12.89%p의 성능 향상을 보였다. 다른 자동 기계학습 라이브러리와 비교했을 때, 다른 autoML 라이브러리와 다르게 하이퍼 파라미터 튜닝을 진행하지 않았음에도 다른 방법과 비교해 비슷하거나 높은 성능을 보였다. 따라서 후속 연구로 하이퍼 파라미터 튜닝을 수행하는 알고리즘을 추가한다면 성능을 더 높일 수 있을 것으로 기대된다.

ACKNOWLEDGMENT

본 연구는 2023년도 동남권 이공계 여성인재 양성사업단의 지원으로 수행되었습니다.

참고문헌

[1] 이용호, 김대중. "한국인에서 당뇨병 위험 예측 모델." Journal of Korean Diabetes, Vol.14(1) : 6-10, 2013

[2] Nam-Kyoo Lim, et al. .“A Risk Score for Predicting the Incidence of Type 2 Diabetes in a Middle-Aged Korean Cohort” Circulation Journal 76.8, 2012.

[3] Rashidi, Hooman H, et al. "Machine learning in health care and laboratory medicine: General overview of supervised learning and Auto ML." International Journal of Laboratory Hematology 43, 15-22. 2021.