NHANES 데이터를 이용한 생활 패턴 기반의 우울증 예측 머신러닝 기법

김용호, 김지하, 우기문, 김형빈, 박현희*
명지대학교
{yhkim98, yaki5896, gmwoo, hbkim, hhpark}@mju.ac.kr

Life pattern-based depression prediction machine learning method using NHANES dataset

Yongho Kim, Jiha Kim, Gimoon Woo, Hyungbin Kim, Hyunhee Park*
Myongji Univ.

요 약

본 논문은 NHANES(National Health and Nutrition Examination Survey) 에서 제공되는 설문조사 데이터 세트를 이용하여 PHQ-9 척도의 우울증 점수를 예측하는 방법을 제안한다. 일반 사용자가 접하기 어려운 의료 설문을 통해서 우울증을 진단받는 것 보다 쉽게 수집할 수 있는 생활 패턴 기반의 설문 응답을 가지고 예측을 할 수 있다는 장점이 있으면서. 본 논문에서 제안하는 기법은 다중 클래스 분류에서는 DNN 의 F1 score 가 60%로 가장 높은 성능이 나왔고 이진 클래스 분류에서는 Decision Tree F1 score 가 71%로 가장 높은 성능을 보였다.

Ⅰ. 서론

현재 대한민국은 해마다 우울증 환자가 증가하는 추세이다. 게다가 작년에는 코로나 19 감염병의 확산으로 코로나 블루로 인한 우울증이 더해져 우울증을 겪는 사람들이 더 늘어났다. 하지만 우울증을 진단하기 위한 의료진의 수는 제한이 있으며 모두 진료를 하기에는 현실적인 어려움이 있다[1]. 그런 문제를 해결하기 위해 현제 머신러닝 혹은 딥러닝을 이용한 정신 질환 검사 챗봇이 만들어 지고 있다[2]. 하지만, 학습을 위해서는 우울증이 있는 환자들의 데이터를 수집해야 한다. 일반적으로 우울증 환자 수가 증가하고 있더라도 정상인의 비율로 보면 그림 1 과 같이 심한 데이터 불균형을 보여준다. 따라서 본 논문에서는 간단한 생활 패턴 설문 데이터로 우울증을 예측하여 우울증 검사에 부담을 줄일 수 있는 방법을 제안 한다.

Ⅱ. 본론

1. 데이터 세트 수집

표 1. 학습에 사용한 설문 코드 및 내용

설문 코드	설문 내용
ALQ151	Ever have 4/5 or more drinks every day?
DBQ700	How healthy is the diet
OCD150	Type of work done last week
PAQ605	Vigorous work activity
SLQ050	Ever told doctor had trouble sleeping?
WHQ030	How do you consider your weight
WHQ070	Tried to lose weight in past year
HOD050	Number of rooms in home
HOQ065	Home owned, bought, rented, other

데이터 세트는 미국인의 건강과 영양을 설문조사를 통하여 데이터를 수집하는 기관인 NHANES[3]에서 데이터 세트를 받아 분석에 사용했다. NHANES 에는 다양한 종류의 설문 데이터들이 많은데 그 중에서 생활 패턴과 우울증을 연관 지어서 분석하기 위해 표 1 에 있는 설문 내용을 특성 데이터로 사용하고 분류 데이터로는 PHQ-9[4]의 척도를 기준으로 각 우울증 정도를 사용했다.

2. 데이터 세트 확인

모델을 학습하기 전에 데이터의 내용을 확인하여 데이터 전처리 작업을 진행했다. 설문에 응답을 거부하거나 하지 않은 데이터들을 제거하니 14372 개의 데이터가 남게 되었다. 그 후 데이터의 클래스 별 분포를 확인하니 그림 1 과 같이 데이터의 불균형한 분포를 확인할 수 있었다. 불균형한 데이터 세트에서 PHQ-9 5 단계의 우울 정도를 모델이 학습하여 예측하기가 쉽지 않을 것이라 생각하여 Normal 과 Mild 클래스를 Normal 클래스 하나로 묶고 Moderate, Moderately Severe, Severe 클래스를 Depression 클래스 하나로 묶어 이진 클래스 분류 데이터 세트를 만들어 다중 클래스 분류와 이진 클래스 분류의 성능을 여러 모델을 사용해보며 각각 측정해 보았다.

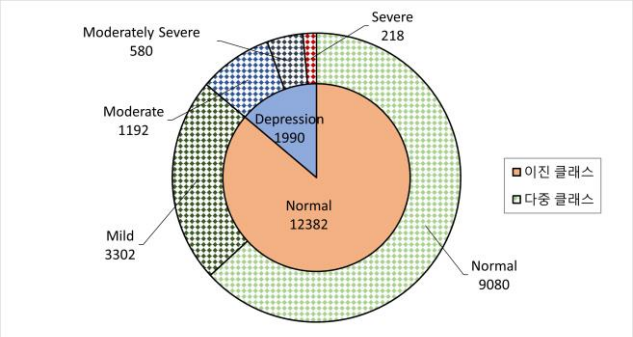

그림 1. 다중 클래스 분류 데이터 세트의 분포 그래프

3. 학습에 사용한 모델

학습에 사용된 모델들은 DNN(Deep Neural Network), DT(Decision Tree), KNN(K-Nearest Neighbor), MLP(Multi-Layer Perceptron), GB(Gradient Boosting), SVM(Support Vector Machine), RF(Random Forest)를 사용했다.

Ⅲ. 실험

모델을 학습하고 학습한 모델의 성능을 측정하기 위해 학습 데이터 세트와 시험 데이터 세트를 8:2 비율로 나누었다. 각 모델을 학습 시킨 후 학습한 모델로 시험 데이터를 예측한 결과로 정확도, 정밀도, 재현율, F1 score 를 측정하여 성능을 비교했다. 딥러닝 모델로는 표 2 와 같은 구조의 하이퍼 파라미터를 사용했다.

표 2. 딥러닝 모델 하이퍼파라미터

계층	노드	활성 함수
입력층	9	-
은닉층 1	128	ReLU
은닉층 2	256	ReLU
은닉층 3	128	ReLU
출력층	5	Softmax

해당 모델에서 다중 클래스 분류를 학습하기 위한 손실 함수는 Sparse Categorical Crossentropy 를 사용하였으며, 이진 분류를 위한 손실 함수는 Binary Crossentropy 를 사용하였다. 최적화를 위한 알고리즘으로는 adam 을사용하였다.

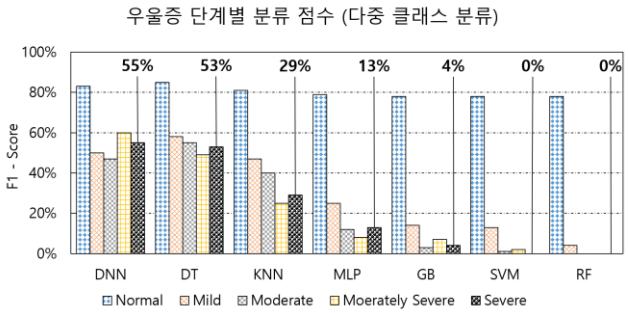

그림 2. 다중 클래스 분류에서의 모델 성능 그래프

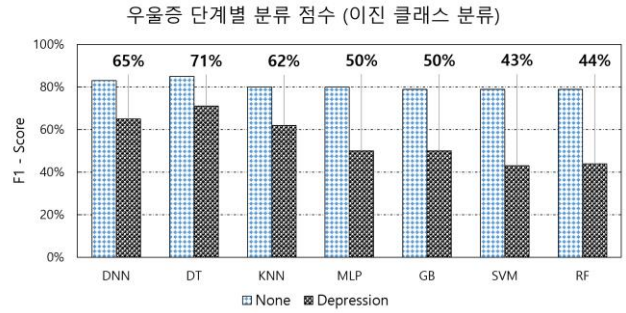

그림 3. 이진 클래스 분류에서의 모델 성능 그래프

Ⅳ. 결론

여러 모델로 다중 클래스 우울증과 이진 클래스 우울증 분류 성능을 측정한 결과 그림 2 와 그림 3 과 같은 결과가 나왔다. 다중 클래스 분류에서는 DNN 이 55%의 F1 score 로 가장 높은 결과를 보이고 이진 클래스 분류에서는 Decision Tree 가 71%의 F1 score 로 가장 높은 결과를 보였다. 이와 같은 결과로 보아 단순한 생활 패턴만으로도 우울증을 예측할 수 있다는 결과를 알 수 있었다.

향후 불균형한 데이터 세트의 문제를 해결하기 위하여 다양한 기법을 적용해 볼 것이다. 또한 예측을 원하는 데이터의 클래스가 비정상적인 비율을 가질 때 발생하는 문제를 해결하기 위한 데이터 증강 기법을 연구할 계획이다.

ACKNOWLEDGMENT

이 논문은 2021 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2021-0-00990, 설명가능한 인공지능 기반 무선랜 네트워크 시스템 고도화 핵심 기술 연구)

참 고 문 헌

[1] "정신과의사 부족, 병상수 증가때문." 의사신문. 2008 년 06 월 12 일 수정, 2021 년 05 월 23 일 접속, http://www.doctorstimes.com/news/articleView.html?idxno=11125

[2] "정신건강, 인공지능 챗봇으로 지킨다." Ai 타임스. 2019 년 04 월 08 일 수정, 2021 년 05 월 23 일 접속, http://www.aitimes.com/news/articleView.html?idxno=47015

[3] Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, [2021.05.26][https://wwwn.cdc.gov/nchs/nhanes/Default.aspx].

[4] Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validaity of a brief depression severity measure. J Gen Intern Med 2001;16:606-613.