한국컴퓨터정보학회 하계학술대회 논문집 제30권 제2호 (2022. 7)

혈액 유전자 발현을 이용한 기계학습 기반 인지장애 예측

이승은[O], 주 우[*], 강경태[*]
[O]한양대학교 컴퓨터공학과 바이오인공지능융합전공,
[*]한양대학교 컴퓨터공학과 바이오인공지능융합전공
e-mail: {seungeunlee, zhouyu, ktkang}@hanyang.ac.kr[O*]

Prediction of Cognitive Impairment Using Blood Gene Expression Based on Machine Learning

Seungeun Lee[O], Yu Zhou[*], Kyungtae Kang[*]
[O]Dept. of Computer Science, Maj. in Bio Artificial Intelligence, Hanyang University,
[*]Dept. of Computer Science, Maj. in Bio Artificial Intelligence, Hanyang University

● 요 약 ●

알츠하이머성 치매는 현존하는 치료법이 없어 경도인지장애 단계에서의 예방이 중요하다. 지금까지의 알츠하이머 연구는 대부분이 뇌영상 마커와 뇌척수액 마커에 집중되어 있었으며, 경도 인지 장애 단계에서의 탐색은 더욱 적었다. 이러한 점에서 혈액 유전자 발현을 이용한 경도 인지장애 단계 예측은 인지 능력에 따른 관련 유전자 식별과 접근 가능한 진단 및 치료 바이오 마커 탐색에 기여할 수 있다. 그러나 유전자 발현 데이터의 경우 환자 수에 비해 높은 차원을 가지기 때문에 과적합을 막고 질병 관련 유전자를 식별하기 위해서는 데이터에서의 의미 있는 차원만을 뽑아내는 차원 축소가 선행되야 한다. 본 연구는 유전자 발현데이터에서의 인지장애 분류를 위해 차원 축소기법과 신경망을 적용하여 인지 장애 정도를 예측하였다. 그 결과, Lasso 이용 차원축소와 신경망을 이용하여 97%의 정확도로 정상과 조기 경도 인지장애, 후기 경도 인지장애 환자를 분류 할 수 있었으며, 더 적은 차원에서도 분류가 가능했다. 이는 혈액 유전자 발현을 이용해 경도 인지장애 단계를 예측한 첫 번째 연구이며, 인지능력 저하에 따른 혈액 유전자 발현의 연관성을 확인하고 향후 조기 진단, 치료 표적 탐색에 기여한다.

키워드: 유전자 발현(Gene Expression), 인지장애(Cognitive Impairment), 차원 축소(Dimension Reduction)

I. Introduction

알츠하이머성 치매는 뇌세포의 퇴화로 인지기능이 저하되면서 일상생활의 장애를 초래하는 만성질환이다. 완전한 인지능력 저하가 나타나기 전 경도 인지장애의 단계를 거친다. 이 단계는 기억력과 인지기능 저하가 나타나지만 일상생활 하는 데 지장이 없는 상태이며 정도에 따라 조기경도 인지장애와 후기경도 인지장애로 나눌 수 있다[1]. 알츠하이머와 같은 퇴행성 뇌질환은 근본적인 치료제 개발이 어려워 조기 진단을 통해 환자들이 약물, 혹은 생활방식의 변화로 질병을 예방해야 한다. 이를 위해서는 치매 이전 단계인 경도인지장애에서의 식별이 중요하다. 그러나 경미한 인지 능력 변화로 인해 임상적인 검사로는 인지장애 단계를 식별하기 어렵기 때문에 생체 데이터를 이용한 진단법이 필요하다. 마이크로 어레이를 통해 얻을 수 있는 유전자 발현 정보는 환자의 생리학적 상태를 반영하는 지표로 질병의 표현형과 유전자형 사이의 관계를 통해 진단과 치료법에 대한 정보를 제공한다. 그러나 유전자 발현 측정 시 수만, 수십 만개의 유전자의 발현량을 한꺼번에 측정하는 반면 얻을 수 있는 샘플의 수는 수십 명, 수백 명으로 적다.

이러한 유전자 발현 데이터는 고차원성 저표본 크기 문제의 대표적인 예시이다. 차원에 비해서 표본이 적을 경우 각 차원에 대한 의미를 모델이 충분히 학습하지 못하며 이는 훈련 데이터 셋에만 과도하게 최적화 되어 새로운 데이터에서 성능이 떨어지는 모델 과적합 문제를 일으킨다[2]. 이를 해결하기 위해서는 매우 많은 차원으로 구성된 다차원 데이터에서 새로운 차원의 데이터로의 차원 축소가 선행되어야 한다. 이러한 차원 축소는 단순히 데이터를 줄이는 것이 아닌 데이터를 잘 설명할 수 있는 요소를 추출하는 과정으로, 더욱이 질병이

나 생물학적 경로에 대해서는 소수의 유전자만이 관련되어 있기 때문에 질병을 설명할 수 있는 유전자 식별이 필요하다[3].

본 논문에서는 혈액 유전자 발현을 통한 인지장애 예측을 위해 차원 축소와 신경망을 적용한다. 유전자 발현 데이터에서의 유전자 선택은 모델의 과적합을 방지하고, 관련 없는 유전자 정보에 의한 노이즈를 제거하여 분류 정확도를 향상시킨다. 또한 질병 관련 유전자를 식별하여 향후 치료법과 진단 마커 탐색에 기여한다[4].

II. The Proposed Scheme

1. 사용 데이터 및 전처리

사용된 유전자 발현 데이터는 Alzheimer's Disease Neuroimaging Initiative(ADNI)[5] 에 공개되어 있으며 정상 인지능력(Cognitive Normal, CN) 260명 , 조기 경도인지장애 215명 (Early Mild Cognitive impairment, EMCI) 후기 경도 인지장애 226명 (Late Mild Cognitive impairment, LMCI)의 말초혈액에서 얻은 4만 2천개의 유전자 발현 정보가 담겨있다. 전처리는 z-score 정규화와 MinMaxScaler를 이용하였다.

2. 차원 축소 방법

차원 축소 중 변수(feature) 선택은 분류에 기여하는 중요한 변수만을 선택하는 과정이다. 기계학습 방법을 통한 변수 선택법은 전체 변수 중에 각 기계학습 모델 결정에 영향을 주는 변수만을 선택한다. 논문에서는 Genetic Algorithm(GA)와, 기계학습 모델을 이용한 Lasso 제약, Support Vector Machine(SVM), Decision Tree(DT), Random Forest(RF)를 이용해 차원 축소를 진행하였다.

3. 신경망 모델

1D CNN을 사용하였으며 구조는 그림 1과 같다[6].

Fig. 1. The architecture of 1D CNN

III. Result

변수 선택 방법 별 선택된 변수 개수와 신경망을 통한 분류 정확도는 다음 표 1와 같다. 표 2는 가장 높은 성능을 보인 Lasso 를 이용한 추출 변수 개수를 조절했을 때의 분류 정확도이다.

Table 1. Accuracy according to dimension reduction method and number of selected features

Method	Lasso	SVM	GA	DT	RF
Numbers	1972	151	182	102	8927
Acc.(%)	97	63	61	30	29

Table 2. The Accuracy for different Feature numbers using Lasso dimension reduction method

Numbers	200	400	600	800
Acc.(%)	81	82	91	96

IV. Conclusions

Lasso를 적용하여 축소된 차원에서 신경망을 통해 정상과 조기 경도 인지장애, 후기 경도 인지장애 환자의 3 단계를 97%의 정확도로 분류하였으며, 선택 개수를 줄였을 때도 좋은 분류 능을 보였다. 이번 연구로 혈액 유전자 발현 데이터에서 경도 인지장애 단계를 높은 성능으로 예측할 수 있었으며 이는 인지능력 저하에 따른 발현 유전자 식별과 치료제 개발 및 알츠하이머 조기진단에 기여할 수 있다.

ACKNOWLEDGEMENT

이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.2020-0- 01343, 인공지능융합연구센터지원 (한양대학교 ERICA))과, 과학기술일자리진흥원의 지원(2022년도, '지역산업연계 대학 Open-Lab육성지원사업') 과2022년도 정부(산업통상자원부) 재원으로 한국산업기술진흥원의 지원을 받아 수행된 연구임 (P0008691, 2022년 산업혁신인재성장지원사업)

REFERENCES

[1] Bi, Xia-an, et al. "Analysis of progression toward Alzheimer's disease based on evolutionary weighted random support vector machine cluster." Frontiers in Neuroscience 12 (2018): 716.

[2] Wei, Tingyang, et al. "Multiclass classification on high dimension and low sample size data using genetic programming." IEEE Transactions on Emerging Topics in Computing (2020).

[3] Mahendran, Nivedhitha, et al. "Improving the classification of alzheimer's disease using hybrid gene selection pipeline and deep learning." Frontiers in Genetics 12 (2021).

[4] Lee, Taesic, and Hyunju Lee. "Prediction of Alzheimer's disease using blood gene expression data." Scientific reports 10.1 (2020): 1-13.

[5] "Microarray Gene Expression Profile Data." ADNI, n.d., https://adni.loni.usc.edu/. accessed 28 Jun 2022.

[6] Mostavi, Milad, et al. "Convolutional neural network models for cancer type prediction based on gene expression." BMC medical genomics 13.5 (2020): 1-13.