기계학습 알고리즘을 이용한 흡연자 예측 연구

백종우[O], 방준일[*], 이주원[*], 김화종[**]
[*]강원대학교 IT 대학 컴퓨터정보통신학과,
[O]강원대학교 컴퓨터공학과,
[**]강원대학교 컴퓨터공학과
e-mail: {hamster_2002[O], alcatraz76[*]}@naver.com, {tkfka965[*], hjkim3[**]}@gmail.com

A Study on Smoker Prediction Using Machine Learning Algorithm

Jongwoo Baek[O], Joonil Bang[*], Joowon Lee[*], Hwajong Kim[**]
[*]Dept. of Computer and Communications Engineering, Kangwon National University,
[O]Dept. of Computer Science and Engineering, Kangwon National University,
[**]Dept. of Computer Science and Engineering, Kangwon National University

● 요 약 ●

본 논문에서는 사람에게서 나타나는 생체 특성과 흡연여부의 상관관계 분석을 위해 랜덤 포레스트와 그래디언트 부스팅 트리의 두 가지 기계학습 알고리즘을 사용하였다. 연구에 사용된 데이터는 국민건강보험공단에서 제공하고 Kaggle에서 취합하여 정리한 건강검진 정보를 사용하였다. 분류 모델의 학습에 있어 혈청 정보가 높은 관계성을 보일 것으로 예상하였으나, 실제 결과는 성별이 가장 큰 영향을 끼치는 것으로 확인되었다.

키워드: 기계학습(Machine Learning), 분류(Classification), 랜덤 포레스트(Random Forest), 그래디언트 부스팅 트리(Gradient Boosting Tree)

I. Introduction

담배는 구강암, 심장병, 폐암 등 여러 가지로 건강에 적신호를 불러올 수 있는 기호식품[1]이다. 본 논문에서는 개인별 특성에 따른 흡연자 분류 예측을 위하여 국민건강보험공단에서 제공한 데이터 중 Kaggle에서 취합한 총 27개의 특성을 가진 55692명의 데이터[2]를 이용해 랜덤 포레스트와 그래디언트 부스팅 트리 모델을 사용하여 개인별 특성에 따라 흡연 여부를 분류해내는 기계학습 모델을 구현하였다.

본 논문의 본론(II~III)에서는 흡연자 분류 모델에서 개인의 특성 중 어떤 특성의 중요도가 가장 높을 것인지 예상해보고, 두 가지 기계학습 모델에서의 중요도를 그래프 형식으로 출력해 결과를 확인하였다. 결론(IV)에서는 예상하였던 주요 특성과 모델이 학습한 특성 중요도 간 차이를 살펴보았다.

II. Backgrounds & Prediction

1. Backgrounds

1.1 Random forest

랜덤 포레스트[3]는 여러 결정 트리의 묶음으로 서로 다른 방향으로 과대적합된 트리를 다수 생성하여 결과값을 평균냄으로써 과대적합을 줄이는 앙상블 기법의 알고리즘이다.

1.2 Gradient Boosting Tree

그래디언트 부스팅 트리[4]는 앙상블 기법을 사용한다는 점에서 랜덤 포레스트와 같으나, 이진 트리의 오차를 보완하는 방식으로 순차적으로 트리를 만든다.

2. Prediction

본 논문에서 사용한 데이터셋에는 사용자의 나이(20세 이상부터 5세 단위), 키, 몸무게, 허리둘레, 좌/우 시력, 좌/우 청력, 수축기/이완기 혈압, 식전 혈당, 총 콜레스테롤, 트라글리세라이드, HDL 콜레스테롤, LDL 콜레스테롤, 혈색소, 요단백, 혈청크레아티닌, AST, ALT, 감마

Gtp(간기능), 흡연 상태, 구강검진 수검여부, 치아 우식증 유무, 치석 유무의 건강상태가 기록되어있다.

본 연구자는 개인별 특성 중 혈청 정보(AST, ALT, 혈청 크레아티닌, 헤모글로빈 등)가 흡연 여부에 높은 영향을 미칠것으로 예상하였다.

III. Training & Evaluation

데이터셋을 살펴본 결과 gender, oral, tartar 의 열이 문자열로 입력되어 있고, dental caries 열의 경우 숫자로 표기되어있으나 수치가 아닌 특성이므로 위 열들을 One-Hot 인코딩을 사용하여 전처리 하였다. 이후 전처리된 데이터셋에 랜덤 포레스트와 그래디언트 부스팅 트리를 사용해 학습하였고, 각 모델의 테스트셋 정확도는 약 82%, 80%를 기록하였다.

두 모델의 특성 중요도를 시각화 한 결과, 각 특성들이 모델 학습에 영향을 준 정도가 상이하였다.

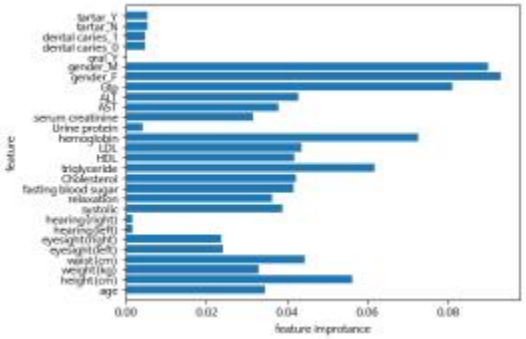

Fig. 1.. 랜덤 포레스트의 특성 중요도

랜덤 포레스트의 경우, 최초 예상하였던 혈청 정보는 특성 중요도에 있어 최저치를 보이지는 않았으나 가장 학습에 영향을 많이 준 특성은 성별로 나타났다. 감마 Gtp 및 헤모글로빈 수치 등이 뒤를 이었다. 따라서 최초 예상(혈청 정보)과 실험 결과(성별)이 다른 것을 확인할 수 있었다.

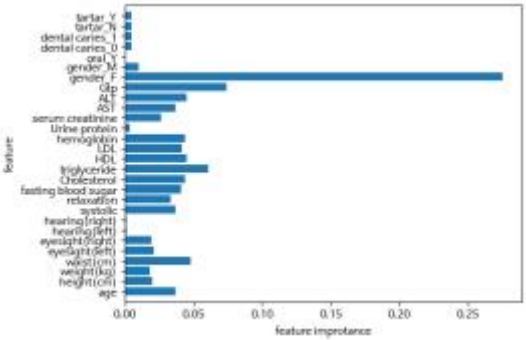

Fig. 2. 그래디언트 부스팅 트리의 특성 중요도

반면, 그래디언트 부스팅의 경우 남성 성별에 대한 중요도가 랜덤 포레스트에 비해 현저히 낮아지고, 여성 성별에 대한 중요도가 높아지는 것을 확인할 수 있었다. 그 외의 나머지 특성들의 중요도는 크게 변화가 없었으나 랜덤 포레스트에서 높았던 헤모글로빈 수치 특성의 중요도보다 허리둘레, triglyceride(중성지방) 등의 특성 중요도가 더 높아지는 것을 확인할 수 있었다.

IV. Conclusions

본 연구자는 흡연여부를 예측하는 분류 모델에서 혈청 정보(AST, ALT, 혈청 크레아티닌, 헤모글로빈 등)가 가장 특성 중요도가 높은 정보일 것으로 예상했으나, 실제 실험 결과 성별 특성이 가장 큰 영향을 미치는 것을 확인하였고, 분류 모델 간 학습에 영향받는 특성과 그 정도가 상이함을 확인하였다. 예시로, 그래디언트 부스팅 트리의 경우 랜덤 포레스트에 비하여 여성 성별 데이터의 중요도가 특히 높은 것을 볼 수 있었다.

다만, 본 연구에서는 분류 모델의 학습에 특성 중요도가 저조한 여러 특성들을 혼합하여 학습하였으므로 모델의 분류 정확도가 약 80% 수준으로 낮고, 모델에 따라 하나 또는 수 개의 특성에 특성 중요도가 집중되는 모습을 확인할 수 있다. 따라서 일차적인 학습 후, 특성 중요도가 저조한 특성들을 제외하여 학습함으로써 정확도 향상 및 특성 중요도의 쏠림 현상 해소가 가능한 추가 연구가 필요할 것으로 보인다.

ACKNOWLEDGEMENT

본 과제(결과물)는 2023년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과입니다. (2022RIS-005)

REFERENCES

[1] "흡연으로 인한 건강피해." 전주시보건소. 2023년06월26일 접속, https://health.jeonju.go.kr/index.jeonju?menuCd=DOM_000000105002001003

[2] Body signal of smoking https://www.kaggle.com/datasets/kukuroo3/body-signal-of-smoking

[3] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001).

[4] Andreas Müller, Sarah Guido 『Introduction to Machine Learning with Python: A Guide for Data Scientists』, Haesun Park, Seoul:Hanbit Media Inc, PP.115-125, 2017.