머신러닝 기반 졸음 검출을 위한 심장변이도의 특성 선택 기법 연구

김송연○, 홍승혁
수원대학교 데이터과학부
olik72@naver.com, shongdr@gmail.com

A Study on the Feature Selection about Heart Rate Variability for Sleepiness Detection Based on Machine Learning

Songyeon Kim○, Seunghyeok Hong
Division of Data Science, The University of Suwon, South Korea

요 약

매년 많은 수의 교통사고들이 졸음운전으로 인해 발생한다. 졸음운전으로 인한 사망률은 19%에 육박하며 교통사고 사망률이 대체로 11%인 것을 감안할 때 다른 원인보다 압도적으로 높은 수치라고 볼 수 있다. 이에 졸음운전을 사전에 파악하고 예방할 수 있는 시스템의 필요성이 제기되고 있다. 운전자의 생체리듬 신호를 파악하는 다양한 방법이 있지만 본 연구에서는 운전자에게 가장 적은 제약조건으로 신호를 추정할 수 있는 심전도 신호를 사용하여 졸음운전을 구별하는 모델을 구현하기 위한 연구를 진행한다. 또한, 모델의 정확도를 높일 수 있는 다양한 방법을 연구해 본다.

1. 서 론

한국도로공사 통계에 따르면 졸음운전은 지난 10년 동안 고속도로 교통사고 원인별 사망률 1위로 뽑혔다. 사망률은 약 19%에 육박하며 이는 일반 교통사고 치사율인 11%인 것을 감안할 때 압도적으로 높은 수치로 볼 수 있다. 이처럼 치명적인 사고를 초래하는 졸음운전을 방지하기 위해 운전자의 생체리듬을 파악하여 졸음운전을 방지하는 시스템 출현의 필요성이 제기되고 있다.

졸음운전 여부를 판별하는데 사용할 데이터를 얻기 위해 지능 시스템을 활용하였다. 지능 시스템을 활용한 대표적인 방식으로는 뇌파(EEG, Electroencephalogram), 심전도(ECG, Electrocardiogram) 등이 있다. 두피에 따라 전극을 배치하고 뉴런과 뇌 사이의 전기적 신호를 기록하는 뇌파는 수면 상태를 파악하는데 용이하지만, 센서를 부착하는 과정이 어렵고, 많은 센서를 부착해야 한다는 단점이 존재한다. [1] 따라서 본 연구에서는 적은 센서의 제약으로 운전자 심장의 전기적 신호를 기록하는 심전도(ECG, Electrocardiogram)를 이용하여 기록한 심박변이도(HRV, Heart Rate Variability) 분석을 통해 운전자의 졸음운전 여부를 구별해주는 모델을 구현해보고자 한다.

2. 연구 방법

2.1 실험 설계

본 실험에 참여한 피험자 6명(25~32세 남성)은 심리적, 신경학적 장애가 없는 사람을 대상으로 선정되었다. 졸음에 영향을 미칠 수 있는 화학물질의 섭취를 통제하여 실험을 진행하였으며 배고픔으로 인한 졸음 억제를 피하기 위해 참여자 모두 식사 후 40~70분 정도의 시간이 경과된 후 실험을 시작하였다. 실험자 대부분이 실험 전날 무리한 운동은 하지 않았다고 답하였다. 모든 참여자들은 실험에 참여하기 전, 서울대학교 병원의 기관 윤리 심의위원회 (IRB No. C-1509-074-704)를 통해 실험에 참여할 것을 동의하였다. 또한 참여자들이 시험을 시작하기에 앞서 시뮬레이터 작업에 익숙해지기 위해 시작 전 5분 동안 운전하는 시간을 가졌다.

실험에 사용된 ECG신호는 MP150데이터수집시스템(BIOPAC Systems Inc, USA)를 사용하여 1kHz의 샘플링 속도에서 측정되었다. ECG 신호 측정에는 3개의 Ag/AgCI 전극(Lead II)과 가슴에 있는 무선 심전도 송신기 (BN-RSPEC-T, BIOPAC 시스템)를 사용하였다. MP150의 3개의 아날로그 입력은 스티어링 휠의 최상단 위치에 고정된 3축 가속도계의 스티어링 모션을 기록했다.

2.2 심박변이도(HRV)

시간에 따른 심장 박동의 주기적인 변화가 심박변이도이다. 심박변이도 분석 방법으로는 시간 영역 분석과 주파수 분석이 존재한다. 본 실험에서는 5분 동안 관측된 데이터를 시

간 영역 분석과 주파수 분석을 모두 사용하여 심박변이도를 관찰하였다. 실험자로부터 얻은 심전도(ECG) 데이터에서 심전도 신호 R-Peak 점을 검출하였고 R-Peak 점들의 간격을 의미하는 R-R interval 데이터를 얻었다. 검출된 R-R interval 데이터의 이상치를 제거하고 얻어진 결과를 심박수의 가변성 분석을 제공하는 HRV analysis 모듈에 대입하여 심장 박동과 졸음운전과의 상관관계를 유추하는데 유의미한 변수를 도출하였다. (그림 1)

Time domain	mean_nni, sdnn, sdsd, nni_50, pnni_50, nni_20, pnni_20, rmssd, median_nni, range_nni, cvsd, cvnni, mean_hr, max_hr, min_hr, std_hr
Frequency domain	lf, hf, lf_hf_ratio, lfnu, hfnu, total_power, vlf

그림 1. HRV분석으로 도출된 독립 변수

2.3 반응시간(Reaction Time, RT)

반응시간은 그림 2 를 통해 알 수 있듯이 운전자가 상황을 인지하고 반응하기까지의 과정에 걸리는 시간을 의미한다. 통계상 운전자 표본의 90%에서 2.5 초 이내의 반응을 보인다고 알려져 있다. (American Association of State Highway and Transportation Officials Standard) 위 기준에 따라 반응시간이 2.5 초를 초과하는 경우 졸음으로 정의하였다. 또한 반응시간이 2.5 초를 초과하는 경우를 제외한 반응 시간의 중앙값인 1 초를 기준으로 졸음운전을 하지 않는 경우를 의미하는 깸을 부여하였고, 이 기준을 종속 변수 target 으로 생성하였다. 정확도의 신뢰성을 높이기 위하여 label 의 균형을 맞추는 작업을 진행하였다. 졸음과 깸을 의미하는 데이터 셋 중 개수가 적은 경우를 기준으로 균형을 맞추는 과정을 통해 평균적으로 약 46 개의 데이터 셋이 사용되었고 표준편차는 약 15 로 나타났다. 이후 심박변이도 분석을 통해 얻어진 독립 변수 데이터와 반응시간을 기준으로 얻어진 종속 변수 데이터를 5 겹 교차검증(5-fold cross-validation)을 진행하여 모델을 학습시켰다. 6 명 운전자 데이터를 학습 시켜 총 4 회의 모델 평균을 도출한 결과 약 80.11%의 정확도를 보였다. (그림 3)

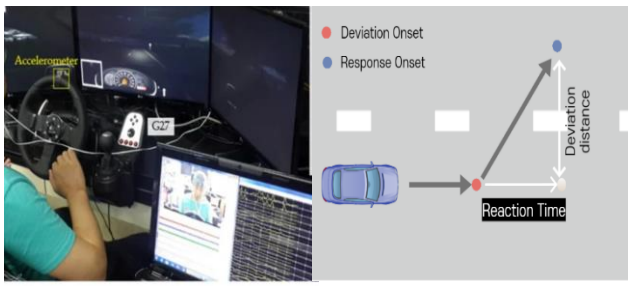

그림 2. 실험환경 및 반응시간

2.4 머신러닝 기법에 따른 정확도

본 연구에서는 앙상블 기법을 활용한 머신러닝 기법을 선택하여 학습을 진행하였다. 앙상블 기법이란 하나의 높은 정확도를 보이는 모델 사용하는 것보다 낮은 정확도를 가지는 여러 모델을 조합하는 것이 더 높은 성능을 보일 수 있다는 이론으로 크게 배깅(bagging)과 부스팅(boosting)이 있다.

Bagging(Bag) 기법은 앙상블 기법의 하나로 복원 랜덤 샘플링을 통해 데이터 셋을 생성하는 부트스트랩(Bootstrap) 방식을 이용하며 이를 여러 모델에 넣어 나온 결과의 평균을 이용하여 모델의 학습을 진행한다. Random Forest(Ran) 기법은 배깅을 활용한 대표적인 예시이다. 배깅보다 많은 무작위성을 부여하여 높은 정확도와 일반화 성능을 가진다. [2]

앙상블 기법의 또 다른 방법인 부스팅(boosting)은 모델을 진행함에 있어 약한 성능을 보이는 모델에 가중치를 부가하여 모델의 성능을 개선하며 학습을 진행한다는 부분에서 배깅(bagging)과 차이가 존재한다. AdaBoost(Ada)는 부스팅을 활용한 기법으로 다양한 학습 알고리즘에 가중치를 부여해 가며 학습한다는 특징이 존재한다. 여러 모델을 혼합하여 사용할 수 있다는 장점이 존재하지만, 과적합 (overfitting)이 자주 발생한다는 단점이 있다. [3] 이 외에도 부스팅을 활용한 기법으로 Gradient Boosting Classifier(GBC), Xgboost(XGB) [4]가 있다.

다양한 기법을 이용해 학습한 결과 머신러닝 기법에 따라 모델의 정확도 차이가 발생하며 알맞은 머신러닝 기법을 선택하는 것은 모델의 정확도를 상승시키는 하나의 방법임을 알게 되었다. (그림 3)

2.5 변수 선택에 따른 정확도 비교

모든 변수를 사용하여 모델을 학습시키는 것은 시간적, 메모리적 측면에서 매우 비효율적이다. 그렇기 때문에 효율적인 모델 학습을 위해 졸음운전을 판별하는데 기여도가 낮은 변수는 제거한 후 분석을 진행하는 것은 모델의 정확도와 효율성을 높이는 또 다른 방법이 될 수 있다. 따라서 본 연구에서는 심박변이도의 분석을 통해 얻은 23가지의 독립 변수 (그림 1) 중에서 졸음운전을 판별하는데 일정 수준의 기여도가 존재한다고 판단되는 변수들을 추출하여 모델의 정확도를 높이는 실험을 진행하였다. 변수를 추출하는 방법은 대표적으로 2가지 방법이 존재한다. Sklearn 라이브러리에서 제공되는 selectkbest 함수와 feature_importances 함수가 대표적인 예시이다. 두 패키지 모두 변수를 선택한다는 공통점이 있지만, 변수를 선택하는 방법에는 차이가 존재한다. Selectkbest 함수는 데이터의 F통계량 분석을 통해 얻은 p-value 값의 크기 순서대로 독립 변수를 선택하며 이 방법은 각 독립변수의 특성을 독립적으로 평가한다는 특징이 존재

한다. 이와 다르게 feature_importances 함수는 트리 기반의 앙상블 모델을 이용할 때 높은 성능을 보이는 변수 선택 방법으로 특정 feature가 트리를 분할하는데 기여한 기여도에 따라 feature에 중요도를 부여하고 중요도의 크기 순서대로 변수를 선택한다. 분류 트리는 데이터의 불순도(Impurity)를 최소화하는 방향으로 트리를 분할하며 학습을 진행한다. 이때 불순도는 해당 범주에 서로 다른 데이터가 섞인 정도를 의미하기 때문에 트리 분할의 기여도는 해당 feature가 모델의 불순도를 최소화 하는데 기여한 정도라고 볼 수 있다.

본 연구에서는 6개의 데이터를 각각 변수를 선택하는 함수에 대입하여 독립변수의 빈도 순위를 얻은후 각각의 데이터별 유의미한 변수로 선택된 독립 변수의 평균을 구하고 빈도수를 기준으로 정렬하였다. 최소 3개의 데이터에서 유의미한 변수로 선택된 변수 중 상위 변수를 임의로 16개, 9개 선택하여 모델을 학습을 진행한 결과 그림 3과 같은 정확도를 얻었다.

3. 연구 결과

그림 3을 통해 모든 독립변수를 선택하여 모델을 학습 시켜 얻은 정확도 (Accuracy)는 80.11%에 비해 feature_importance함수를 사용하여 변수를 선택한 경우 82.69%, 81.87%로 정확도가 상승한 반면 selectkbest함수를 이용한 경우 79.61%, 69.55%로 정확도가 하락한 것을 볼 수 있다. 이를 통해 의사결정 트리를 기반으로 하는 머신러닝 기법에서는 데이터의 불순도를 최소화 할 수 있는 변수를 선택하여 모델을 학습시키는 것이 더 효율적인 변수 선택 방법임을 확인했다.

또한, 독립변수 선택 개수에 따라 정확도의 차이가 발생하는 것을 확인하였다. 독립변수를 제거하여 모델을 학습시키는 방법은 항상 정확도의 상승을 가져올 것이라 예상했던 것과 다르게 일정 수준 이상의 변수를 줄이면 종속 변수를 예측하는 데 필요한 중요 정보가 삭제되어 모델의 정확도를 하락시킬 수 있다는 사실을 확인하였다. 이를 통해 특정 변수를 선택하여 모델을 학습시키는 과정은 모델 학습에 필요한 메모리와 시간을 크게 감소시키고 모델을 간소화하여 해석하는데 간편함을 제공하지만 동시에, 과도한 변수 선택은 중요 정보의 삭제로 이어질 수 있다는 것을 알게 되었다. 따라서 변수 제거 과정을 수행하기에 앞서 독립변수 간의 관계를 파악하여 제거할 변수와 제거하지 말아야 할 변수를 구별해 내는 과정의 필요성을 알게 되었다.

뇌파 신호의 활용을 통해 얻은 생체 신호 데이터는 운전자의 졸음운전을 구별하는데 높은 정확도를 가지지만 뇌파 신호를 측정하기 위해 부착하는 센서들은 운전자에게 큰 불편함을 준다. 따라서 본 연구에서는 운전자에게 센서 부착의 부담을 최소화하여 얻을 수 있는 ECG 신호를 사용하여 운전자의 불편함은 최소화하여 졸음운전을 분석하는 모델을 구현해보았다. 추후, 심박수를 측정할 수 있는 스마트워치와 같은 휴대용 전자기기에 졸음운전을 구별해내는 시스템을 결합하여 상용화한다면 일상생활에서 졸음운전으로 인해 발생하는 교통사고를 예방할 수 있을 것이라 기대된다.

모델	원본	Importances_16	Selectkbest_16	Importances_9	Selectkbest_9
Ada	80.82%	82.21%	75.83%	83.31%	66.11%
Xgb	77.77%	84.03%	79.16%	82.37%	69.72%
Ran	78.61%	84.89%	79.44%	83.92%	73.61%
Gbc	84.44%	80.44%	83.05%	79.92%	66.94%
Bag	78.38%	81.83%	80.55%	79.83%	71.39%
모델평균	80.11%	82.69%	79.61%	81.87%	69.55%

그림 3. 변수 선택에 따른 정확도

Importances_16	nni_50, total_power, mean_hr, mean_nni, median_nni, vlf, hf, lf, sdsd, rmssd, pnni_50, pnni_20, sdnn, min_hr, nni_20, lfnu
Selectkbest_16	vlf, lf_hf_ratio, lfnu, pnni_20, hfnu, mean_hr, mean_nni, median_nni, sdnn, nni_20, rmssd, sdsd, nni_50, pnni_50, total_power, std_hr
Importances_9	median_nni, nni_20, mean_hr, mean_nni, Total_power, pnni_50, pnni_20, nni_50, vlf
Selectkbest_9	pnni_20, sdsd, rmssd, cvsd, median_nni, total_power, nni_50, min_hr, nni_20

그림 4. 모델 학습에 선택된 독립 변수

4. Acknowledgements

본 논문은 2020학년도 수원대학교 학술진흥연구비 지원에 의한 논문임

5. 참고문헌

[1] Chun-Hsiang Chuang, Chih-Sheng Huang, Li-Wei Ko, Chin-Teng. “An EEG-based perceptual function integration network for application to drowsy driving” Knowl. Based Syst. 2015.

[2] P. Thanh Noi, M. Kappas, Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery, Sensors (Basel). 2017.

[3] Rojas, R. AdaBoost and the super bowl of classifiers a tutorial introduction to adaptive boosting. Freie University, Berlin, Tech. Rep.2009.

[4] Chen, Tianqi; Guestrin, Carlos. "XGBoost: A Scalable Tree Boosting System". In Krishnapuram, Balaji; Shah, Mohak; Smola, Alexander J.; Aggarwal, Charu C.; Shen, Dou; Rastogi, Rajeev (eds.). Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, ACM. pp. 785–794. August 13-17, 2016.