기계학습 알고리즘을 이용한 교통사고 심각도 예측 모델에 관한 연구

박세영 [1], 송영훈 [1], 김광오 [2], 한결아 [2], 조석헌*
[1] 경기대학교, [2] 조선대학교, *University of California, San Diego (UCSD)

se0park0022@gmail.com, dogs7783@gmail.com, kgo1285@gmail.com,
hka97123800@gmail.com, *justinshcho@gmail.com

Research on the Traffic Accident Severity Prediction Model Using Machine Learning Algorithms

[1]Seyoung Park, [1]Younghun Song, [2]Gwangoh Kim, [2]GyeolA Han, Seokheon Cho*
[1]Kyonggi University, [2]Chosun University, *University of California, San Diego (UCSD)

요 약

This study aims to propose a predictive model for the severity of traffic accidents based on external environmental factors, which can perform a role to reduce the number of casualties and accident rates. Data collected from traffic accidents occurring in England from 2021 to 2022 were utilized to provide a traffic accidents severity prediction model. The severity of traffic accidents included in the dataset can lead to a multi-class classification prediction model with the three labels. Moreover, since the severity of most traffic accidents is classified as slight, the dataset exhibits characteristics of imbalanced data. Four artificial intelligence algorithms, such as Adaptive Boosting, Gradient Boosting Tree, K-Nearest Neighbors, and Random Forest, were employed for predicting the severity of traffic accidents. The performance analysis of our prediction models presented that the Random Forest algorithm-based model shows the highest accuracy. However, due to the limitation of imbalanced datasets, the other performance metrics, such as Macro Recall, Macro Precision, and Macro F1-measure, for the Random Forest algorithm-based model showed lower performance compared to accuracy.

Ⅰ. 서 론

교통사고는 전 세계적으로 수많은 인명과 재산에 큰 손실을 가져오는 커다란 문제이다. 교통사고 발생의 원인은 단지 운전 중의 실수나 차량의 기술적 결함에만 국한되지 않으며 날씨나 도로 상태와 같은 다양한 외부적 요인들도 크게 작용한다. 이러한 이유로, 다양한 외부 환경에 따라 교통사고 심각성을 예측하는 것은 안전한 도로 환경을 만들고 교통사고 발생률을 줄이는데 있어 매우 중요한 요소로 동작할 수 있다. 따라서 본 연구에서는 교통사고의 심각도에 영향을 미치는 다양한 특성들을 분석하고, 인공지능 기술을 활용하여 교통사고 심각도 예측 모델을 개발하는 것을 목표로 하고 있다.

S. P. Ardakani *et al.*는 2005 년부터 2014 년에 영국에서 발생한 교통사고 데이터 [1]를 바탕으로 도로교통사고 예측 모델을 개발하기 위해 Decision Tree Classifier, Random Forest Classifier (RFC), Multinomial Logistic Regression 및 Naive Bayes Classifier 등 총 4 가지 알고리즘을 사용하였다. 교통사고 심각도 예측 부분에서, RFC 기반 모델이 약 85.58%의 높은 정확도를 보이며 가장 우수한 성능을 나타냈다 [2]. H. Lee *et al.*는 2015 년에서 2017 년 사이에 발생한 총 21,103 건의 고속도로 교통사고 데이터를 대상으로 사고 심각도 예측 모델을 개발했다. 네 가지의 Boost 알고리즘과 Random Forest 알고리즘을 적용하여, 정밀도, 재현율 및 F1-score 등을 기준으로 모델을 평가했다. 이 중 LightGBM 방법이 최상의 결과를 도출했으며, 교통사고 심각도에 가장 큰 영향을 미치는 요소가 사고에 관련된 차량의 수임을 확인하였다 [3]. 한편, A. Çelik *et al.*은 2011 년부터 2021 년까지 텍사스에서 발생한 도로 교통사고 심각도를 예측하기 위해 총 6 개의 인공지능 알고리즘을 사용하였다. 6 개의 알고리즘 중에서 Logistic Regression 알고리즘 기반 예측 모델이 82%의 재현율과 89%의 F1 score 그리고 0.881 의 ROC-AUC 값을 가져 가장 높은 성능을 보였다 [4]. 또한, M. Zheng *et al.*은 영국 리즈에서 발생한 교통사고 데이터 21,436 건을 이용하여 교통사고 심각도를 예측하였다. 심각도 예측 주요 모델로서 Traffic Accident's Severity Prediction-Convolutional Neural Network (TASP-CNN)을 선택했다 [5].

본 연구에서는 독립변수를 순서형 데이터와 명목형 데이터로 구분하며, 데이터 오류 확인을 위한 독립변수를 생성하였다. 추가로 Adaptive Boosting (AdaBoost), Gradient Boosting Tree (GBT), K-Nearest Neighbors (KNN), Random Forest (RF)와 같은 다양한 기계학습 기법을 사용하여 모델의 성능을 분석하였다.

본 논문의 구성은 다음과 같다. 제 Ⅱ 장에서는 원본 데이터셋 설명 및 데이터 전처리 과정에 대한 소개를 시작으로 교통사고 심각도 예측 분석을 위한 구축한 데이터셋을 설명한다. 제 Ⅲ 장에서는 본 논문에서 사용한 알고리즘과 성능 지표를 제시하며 교통사고 심각도 예측 모델에 대한 성능 결과를 비교하며 최종적으로, 제 Ⅳ 장에서 본 논문의 결과와 향후 연구 과제 및 기대효과에 대해 제시하며 마무리 짓고자 한다.

Ⅱ. 데이터 전처리 과정 및 분석 데이터셋 설명

2.1 원본 데이터셋 설명

본 연구에서는 Kaggle 에서 제공하는 자동차 사고 데이터셋을 활용했다 [6]. 해당 데이터셋은 2021 년 1 월 1 일부터 2022 년 12 월 31 일까지, 2 년 동안의 영국에서 발생한 교통사고 심각도를 기록한 자료이다.

2.2 데이터 전처리 과정

영국 (U.K.)의 4 개 구성국 중 잉글랜드 (England) 지역만을 고려하여 진행하였다. 주말과 주중의 차이를 통해 도로 교통사고의 심각도를 예측하는 방식을 채택했으며, 잉글랜드의 공휴일은 주말로 분류하였다. 교통사고 발생 시각 (Time)과 조명 상태 (Light Conditions)사이의 데이터 불일치를 확인할 수 있었다. 예를 들면, 교통사고 발생 시간이 새벽 2 시임에도 불구하고 당시의 조명 상태가 햇빛이 있는 낮 (daylight) 상태인 샘플들이 존재하였다. 이러한 문제를 해결하기 위해, 잉글랜드의 수도인 런던의 일출과 일몰 시간을 기준으로 오류를 확인하였다. 이 과정에서 발견된 오류와 결측값은 전체 데이터의 약 7.84%로 다소 작은 비율이기 때문에 분석 시 오류와 결측값을 제외하였다. 교통사고 발생 시간 데이터는 분 단위가 아닌 시간 단위만을 고려하여 분석의 복잡성을 줄였다. 마지막으로, 날씨와 바람 상태 (Weather Conditions)를 나타내는 변수는 날씨 상태 (Weather)와 바람 강도 (Wind)를 나타내는 각각 새로운 변수로 분할하였다. 예컨대, 날씨와 바람 상태 변수가 'fine + high winds'인 경우에 'fine'은 날씨 상태 변수 값으로 'high winds'은 바람 강도 변수 값으로 분리하였다.

데이터를 두 가지 유형으로 구분했다. 첫 번째는 순서가 있거나 등급을 나타낼 수 있는 순서형 데이터이며, 두 번째는 범주형으로 나타낼 수 있지만 순서나 등급이 없는 명목형 데이터이다.

다음은 순서형 데이터들에 대한 분석을 위한 처리 과정이다. 차량 유형 (Vehicle Type)에 대해, 승용차와 택시를 가장 낮은 무게 등급인 가벼운 차량을 의미하는 'light = 1'으로 버스와 미니버스는 중간 무게 등급인 'middle = 2'으로 그리고 3.5 톤 이상의 화물차는 높은 무게 등급인 'high = 3'로 분류하였다. 바람 강도 (Wind)는 강한 바람과 강한 바람이 없음을 각각 'high winds = 1'과 'no high winds = 0'으로 설정하였다. 사고 발생 장소의 도심 여부 (Urban or Rural area) 변수는 도시와 농촌을 각각 'urban = 0'과 'rural = 1'으로 각각 분류하였다. 사고 발생 노면 상태 (Road Surface Conditions)는 다음과 같이 설정하였다: 건조를 'dry = 1'으로, 3cm 이상 물에 잠김을 'flood over 3cm deep = 2'로, 눈을 'snow = 3'으로, 젖거나 습함을 'wet or damp = 4'로, 서리나 얼음을 'frost or ice = 5'로 설정하였다. 사고 당시 조명 상태 (Light Conditions)는 낮, 어둡지만 길 조명 동작 그리고 조명이 없는 어둠 상태 각각을 'daylight = 1' 'darkness-lights lit = 2' 'darkness = 3'로 분류하였다.

날씨와 도로 유형에 대해서는 명목형 변수로 처리하고, 교통사고 심각도에 미치는 영향을 분석하기 위하여 두 변수에 대해서는 원-핫 인코딩 (One-Hot Encoding) 방법을 적용했다. 즉, 날씨 변수는 0 또는 1 의 값 만을 가지는 맑음 (Fine), 비 (Raining), 눈 (Snowing), 안개 (Fog)의 새로운 변수들로 나누었다. 또한, 도로 유형은 일방 통행 (One Way Street), 단방향 도로 (Single Carriageway), 양방향 도로 (Dual Carriageway), 회전 교차로 (Roundabout)로 나누었다. 예를 들어 비가 오는 날 단방향 도로에서 사고가 발생했다고 가정할 경우, 비와 단방향 도로 변수들에 1 을 할당하지만 위의 나머지 변수들은 0 으로 설정하는 것이다.

2.3 교통사고 심각도 예측 분석을 위한 데이터셋

원본 데이터셋으로부터 다양한 전처리 과정을 거쳐 생성되고 교통사고 심각도 예측 모델을 위해 사용될 데이터셋을 표 1 에 나타내었다. 교통사고 심각도 (Accident Severity)변수는 종속변수이다. 그 외의 20 개의 독립 변수들이 존재하고 있다. 데이터 샘플의 총 개수는 224,046 개이다.

표 1. 전처리 이후 데이터셋

Dependent Variable
Accident Severity ({slight, serious, fatal})
Independent Variable
- Day Type ({weekend, weekday})
- Light Conditions ({daylight, darkness-lights lit, darkness})
- Number of Casualties ({1, …, 42})
- Number of Vehicles ({1, …, 32})
- Road Surface Conditions ({1, …, 5})
- Speed Limit ({20, 30, …, 70})
- Urban or Rural ({0, 1})
- Vehicle Weight ({1, 2, 3})
- Wind ({0,1})
- Month ({1, …, 12})
- Hour ({0, …, 23})
- One Way Street ({0, 1})
- Single Carriageway ({0, 1})
- Dual Carriageway ({0, 1})
- Roundabout ({0, 1})
- Slip Road ({0, 1})
- Fine ({0, 1})
- Raining ({0, 1})
- Snowing ({0, 1})
- Fog ({0, 1})

그림 1 은 종속변수인 교통사고 심각도의 구성 비율을 보여준다. 교통사고 심각도는 3 개의 라벨 (label)로 구성되어 있다. 보통 (slight), 심각 (serious), 사망 연루 (fatal), 총 224,046 개의 데이터 샘플 중에서 보통, 심각, 사망 연루 샘플의 개수가 각각 192,344 개, 29,121 개 그리고 2,581 개이고 비율은 각각 약 86%, 13% 그리고 1%로 구성되어 있다. 본 연구에서 고려하는 데이터셋이 심각한 불균형 데이터셋 (Unbalanced Dataset)임을 확인할 수 있다.

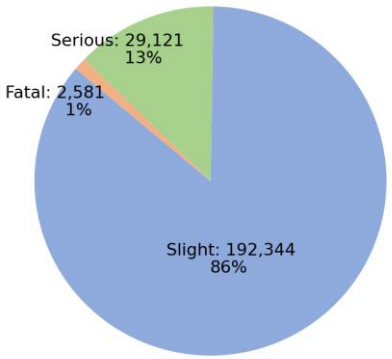

그림 1. 종속 변수 (Accident Severity) 구성 비율

Ⅲ. 교통사고 심각도 예측 모델 분석

3.1 알고리즘 및 성능 평가 지표

교통사고 심각도를 분류 예측하기 위해서 Adaptive Boosting (AdaBoost), Gradient Boosting Tree (GBT), K-Nearest Neighbors (KNN) 그리고 Random Forest (RF) 등 총 4 개의 기계학습 알고리즘을 사용하고자 한다. 또한, 평가 지표로는 Accuracy, Macro recall, Macro Precision 및 Macro F1-measure 을 고려하였다. 고려하고 있는 교통사고 심각도 분류 예측 모델은 다중 클래스 분류로서 각 평가 지표의 클래스마다 개별적으로 성능을 측정하여 Macro 평균을 통한 성능을 평가하고자 한다.

3.2 교통사고 심각도 예측 모델 분석 결과

그림 2 는 본 연구에서 제시하는 교통사고 심각도 분류 예측 모델들에 대한 정확도를 보여주고 있다. RF 알고리즘 기반 예측 모델이 0.858 의 정확도를 보여주며 4 개의 모델들 중에서 교통사고 심각도를 예측하는데 가장 효과적임을 발견하였다. 이와는 달리, AdaBoost 알고리즘 기반 예측 모델의 정확도는 0.822 로 4 개의 모델들 중에서 가장 좋지 않은 정확도를 보인다. 0.822 의 정확도는 절대적으로 평가했을 때에는 높은 정확도라고 할 수 있다. 하지만, 본 연구에서 고려하는 데이터셋은 불균형 데이터이기 때문에 정확도보다는 불균형 데이터 분석에 적합한 다른 성능 지표들을 확인하는 것이 더욱 의미가 있다.

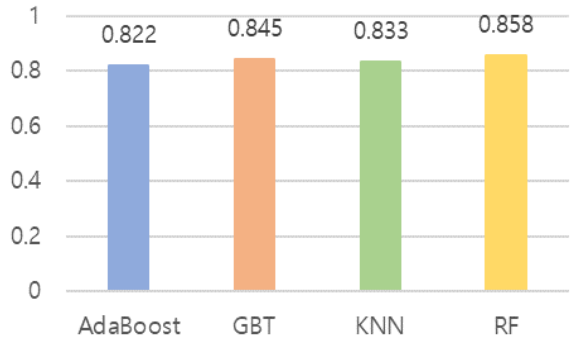

그림 2. 예측 분류 모델 별 정확도 결과

표 2 는 교통사고 심각도 분류 예측 모델들에 대한 Macro Recall, Macro Precision 및 Macro F1-measure 등의 성능 결과를 보여주고 있다. 그림 2 의 교통사고 심각도 분류 예측 모델에 대한 정확도 결과와는 달리, 나머지 3 개의 성능 지표 값이 낮게 나옴을 확인할 수 있다. Macro Recall 값과 Macro F1-measure 값을 기준으로 AdaBoost 알고리즘을 사용하는 모델이 가장 좋은 성능을 보이지만, Macro Recall 값과 Macro F1-measure 값이 각각 0.348 과 0.347 으로 교통사고 심각도 분류 예측 모델로 권장할만한 모델들이 아니다. 이렇게 Macro Recall 값과 Macro F1-measure 가 낮은 이유는 실제로 발생한 교통사고 심각도가 사망 연루(fatal)인 경우에 있어서 예측이 틀리기 때문이다.

표 2. 교통사고 심각도 예측 분류 모델 성능 결과

	Ada Boost	GBT	KNN	RF
Macro Recall	0.348	0.341	0.34	0.334
Macro Precision	0.364	0.371	0.353	0.41
Macro F1-measure	0.347	0.33	0.332	0.309

본 연구에서 제시하는 교통사고 심각도 분류 예측 모델들이 전반적으로 정확도는 높고 상대적으로 Macro Recall 과 Macro F1-measure 가 낮기 때문에, 사용하는 데이터셋에 포함된 특성을 변화시키거나 다양한 샘플링 (Sampling) 기법을 활용하여 불균형 데이터에 대해 균형성을 맞추는 방안들을 고려해야 할 필요성이 있다.

Ⅳ. 결 론

본 연구에서는 잉글랜드 (England)에서 발생한 교통사고 데이터를 기반으로 교통사고 심각도 분류 예측 모델을 개발하여 그 성능을 비교 분석하였다. 다양한 종속변수들을 수정하고 월별 일출 및 일몰 시간에 따른 데이터들의 오류 검출을 통해 예측 모델에 대해 정확한 분석을 진행하였다. Adaptive Boosting (AdaBoost), Gradient Boosting Tree (GBT), K-Nearest Neighbors (KNN) 그리고 Random Forest (RF) 등 총 4 개의 기계학습 알고리즘을 기반으로 한 예측 모델의 성능을 평가했다. 그 결과, RF 알고리즘을 사용한 모델이 0.858 의 가장 높은 정확도를 나타냈다. 하지만, 다른 모델 성능 지표들인 Macro Recall 값과 Macro F1-measure 값이 교통사고 심각도 분류 예측 모델로 사용하기엔 4 개의 모델들 모두 전반적으로 너무나 낮았다. 따라서 이처럼 Macro Recall 값과 Macro F1-measure 값이 향상된 예측 모델을 개발해야 할 필요성이 있다. 향후 연구에서는 이러한 문제를 해결하기 위해서 데이터의 균형을 맞춘 새로운 데이터셋을 구성하고 다양한 특성들을 포함 또는 수정하여 교통사고 심각도 예측 모델의 성능을 높이는 모델을 개발하고자 한다. 궁극적으로 이러한 모델 개발을 통해 안전한 도로 환경을 마련하여 교통사고로 인한 부상 및 사망률을 줄이는 데 기여하고자 한다.

ACKNOWLEDGMENT

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the National Program for Excellence in SW supervised by the IITP (Institute of Information & communications Technology Planning & Evaluation) (No.2021-0-01393).

참 고 문 헌

[1] Dave Fisher-Hickey, “1.6 million UK Traffic Accidents,” Kaggle, 2017, Available :https://www.kaggle.com/daveianhickey/2000-16-traffic-flow-england-sco tland -wales.

[2] Saeid Pourroostaei Ardakani, Xiangning Liang, Kal Tenna Mengistu, Richard Sugianto So, Xuhui Wei, Baojie He, and Ali Cheshmehzangi, "Road Car Accident Prediction Using a Machine-Learning-Enabled Data Analysis," Sustainability 2023 vol. 15, no. 7, Mar. 2023

[3] Hyun-Mi Lee, Gyo-Seok Jeon, and Jeong-Ah Jang, " Predicting of the Severity of Car Traffic Accidents on a

Highway Using Light Gradient Boosting Model," The Journal of the Korea Institute of Electronic Communication Sciences, vol. 15, no. 6, pp. 1123-1130, Dec. 2020.

[4] Ali Çelik and Onur Sevli, "Predicting Traffic Accident Severity Using Machine Learning Techniques," Turkish Journal of Nature and Science, vol. 11, no. 3, pp. 79 - 83, Sep. 2022.

[5] Ming Zheng, Tong Li, Rui Zhu, J. Chen, Zifei Ma, Mingjing Tang, Zhongqiang Cui, and Zhan Wang, "Traffic Accident's Severity Prediction: A Deep-Learning Approach-Based CNN Network," IEEE Access, vol. 7, pp. 39897 - 39910, Mar. 2019.

[6] Saher Muhamed, "Car Accident Dataset," Kaggle, 2024, Available:https://www.kaggle.com/datasets/nextmillionaire/car-accident-dataset.