기계학습 기반 고령운전자 교통사고 사고심각도 분석 모형 개발
- 대전광역시 사례를 중심으로-

임영빈*, 김기정**
*국민대학교 행정학과
**두원공과대학교 전기자동차과
e-mail:kimkj@doowon.ac.kr

Development of a Model for Analysis of Traffic Accidents of Older Drivers based on Machine Learning
- A case study on Daejeon metropolitan city -

Youngbin Lym*, Ki-Jung Kim**
*Dept. of Public Administration, Kookmin University
**Dept. of Electric Vehicle, Doowon Technical University

요 약

최근 교통사고의 사회적 비용을 줄이는 방법을 찾기 위해 OECD에서는 교통사고 DB를 활용한 사고절감 방안 연구를 추진 중이며, OECD 회원국들은 교통사고 데이터의 공유체계를 구축하고 사고 감소방안 연구결과를 공유하고 상호비교를 통한 대처방안을 마련하고 있다. 국내에서도 한국도로교통공단에서 제공하는 교통사고분석시스템(TAAS)의 데이터를 활용한 교통사고 연구가 활발히 진행되고 있다. 하지만 TASS에서는 사고지점의 위치정보를 제공하지 않는 등의 한계가 있어 기존 연구들은 교통사고의 사고지점에 대한 공간적 정보를 제한적으로 활용하여 수행해왔다. 이러한 한계점을 극복하고자, 본 연구에서는 데이터마이닝 기법을 활용하여 사고지점의 위치정보를 포함한 지리정보 기반의 교통사고 데이터베이스를 구축하고 기계학습 기반의 교통사고 심각도 모형을 개발하였다. 이를 통해 기존 TASS 데이터만 활용한 모형과 사고지점 위치정보를 활용한 데이터를 포함한 새로운 모형을 통해 비교 분석하였으며, 결과적으로 본 연구에서 제시한 모형이 다소 높은 정확도를 보였다. 특히 사고심각도를 분류함에 있어 기존 TASS 데이터에 포함되지 않는 변수들이 중요한 인자들로 도출되었음을 확인하였다. 향후 정확도 향상을 위해 더 많은 공간정보 빅데이터를 포함하는 사고 심각도 모형 기반의 교통사고 예측 모델 개발 연구를 수행할 예정이다.

1. 서론

지난 20년간 국내 교통사고 사망자수는 2004년 6천명대에서 2021년 약 2천9백명으로 3천명 아래로 떨어져 점점 감소추세에 있다. 하지만 우리나라의 인구 10만명당 교통사고 사망자수는 2020년 5.9명으로 OECD 회원국 평균 4.7명의 1.3배 수준이며 65세 이상 고령자가 절반에 가까운 46.0% (1,258명)를 차지한 것으로 나타났다. 또한, 교통사망자를 발생시킨 운전자가 65세 이상 고령자인 경우도 26.9%(735명)로 가장 높은 비율을 차지했다. 이는 우리나라의 65세 이상 고령인구는 2021년 전체 인구의 16.5%로 지속적인 증가세 있으며, 이로 인한 고령 운전자도 증가하고 있는 상황이다. 특히 대전광역시의 경우 65세 이하 운전자의 의한 사고는 2011년에 비해 2020년에는 사망자 48%로 중상사고 28%로 감소하였지만, 65세 이상 고령운전자에 의한 사고는 2011년에 비해 2020년의 경우 사망자 44%, 중상사고 122%으로 급격하게 증가하고 있어 이에 대한 대책이 시급한 상황이라 할 수 있다.

최근 교통사고의 사회적 비용을 줄이는 방안을 강구하고자, OECD에서는 교통사고 DB를 활용한 사고절감 방안 연구를 추진 중에 있다. 이에 OECD 회원국들은 교통사고 데이터의 공유체계를 구축하고 사고 감소방안 연구결과를 공유하고 상호비교를 통한 대처방안을 마련하고 있다. 국내에서도 한국도로교통공단에서 제공하는 교통사고분석시스템(TAAS)의 데이터를 활용한 교통사고 연구가 활발히 진행되고 있다.

하지만 TASS에서는 사고지점의 위치정보를 제공하지 않는 등의 한계가 있어 기존 연구들은 교통사고의 사고지점에 대한 공간적 정보를 제한적으로 활용하여 수행해왔다. 이러한 한계점을 극복하고자, 본 연구에서는 데이터마이닝 기법을 활용하여 대전광역시의 교통사고 지점의 위치정보를 포함한 지리정보 기반의 교통사고 데이터베이스를 구축하고 기계학습 기반의 교통사고 심각도 모형을 개발하였다. 이를 통해 기존 TASS 데이터만 활용한 모형과 사고지점 위치정보를 활용한 데이터를 포함한 새로운 모형을 통해 비교 분석하였으며, 결과적으로 본 연구에서 제시한 모형이 다소 높은 정확도

를 보였다. 특히 사고심각도를 분류함에 있어 기존 TASS 데이터에 포함되지 않는 변수들이 중요한 인자들로 도출되었음을 확인하였다.

2. 연구 방법 및 범위

교통사고분석시스템(TAAS)를 활용하여 2007~2021년 사이에 발생한 대전광역시 내 모든 교통사고 총 98,205건의 원시데이터 중 65세 이상 고령 운전자에 의해 발생된 8,274건의 데이터를 수집 하였다. 교통사고분석시스템의 제공 자료의 경우 사고발생 지점은 제시하고 있으나, 구체적인 좌표 정보(사고발생 지점)는 기본 데이터베이스에서 제공하지 않는 한계점이 존재한다. 따라서 본 논문에서는 웹크롤링 기법을 활용해 대전광역시에서 발생한 개별 교통사고의 좌표정보를 수집하고 QGIS의 공간정보와 결합하고,. 이를 통해 지능형 교통체계 표준 노드와 링크 등의 공간정보를 포함하는 확장된 교통사고 데이터 베이스를 구축하였다. 이러한 데이터베이스를 바탕으로 기계학습을 위한 전처리(Preprocessing) 및 Feature Engineering을 수행하고 표 1과 같이 2 개의 종속변수 모형을 구축하였다. 또한 본 논문에서 활용한 설명변수는 TAAS에서 제공하는 사고요일 및 사고유형 등 12개 변수와 공간정보를 통해 추가 확보한 도로등급 등 8개의 변수를 활용하였다.

표 1. 사고심각도 분석을 위한 데이터

사고심각도 분석을 위한 데이터(N=7,407)		
구분	**세부내용**	**비고**
종속변수		
Severity A	▪ 사고심각도: 경상(Minor; 부상신고+경상), 중상(Severe; 중상+사망)	범주형
Severity B	▪ 사고심각도: 경상(Minor; 부상신고+경상), 중상(Severe), 사망(Fatality)	범주형
설명(독립)변수		
CrashDay (사고요일)	▪ 월요일, 화요일, 수요일, 목요일, 금요일, 토요일, 일요일	TASS 원시데이터
CrashType (사고유형)	▪ 차량단독, 차대사람, 차대차-추돌, 차대차-정면충돌, 차대차-측면충돌, 차대차-기타	
Violation (법규위반)	▪ 신호위반, 안전운전불이행, 중앙선침범, 보행자보호의무위반, 기타	
RoadSurface (노면상태)	▪ 건조, 젖음	
Climate(기상상태)	▪ 맑음, 흐림, 비	
RoadType (도로형태)	▪ 교차로, 단일로, 기타	
VehicleType (가해운전자 차종)	▪ 승용차, 승합차, 트럭	
Gender (성별)	▪ 남성, 여성	
NewAge (연령)	▪ Young_Old, Mid_Old, Old_Old	
Year (사고년도)	▪ 사고년도(2007~2021)	
Season (계절)	▪ 봄, 여름, 가을, 겨울	
DayTime (사고발생시간)	▪ 1~4, 5~8, 9~12, 13~16, 17~20, 21~24(4시간 단위)	
RoadRank (도로등급)	▪ 지방도, 특별·광역시도, 일반국도, 고속국도	공간정보 포함 한 추가 데이터
Speed (최고제한속도)	▪ 최고제한속도	
Lanes (차로수)	▪ 차로 수	
Rest_Veh (통행제한차량)	▪ 이륜차 제한, 모두 통행가능	
Rest_H (통과제한높이)	▪ 통과제한 높이	
Length (도로연장)	▪ 도로연장	
Latitude	▪ 위도	
Longitude	▪ 경도	

표 1과 같은 데이터 구축을 통해 본 논문에서는 선형(Linear) 모형, KNN(K-nearest neighbor), Random Forest, SVM(Support vector machine), GBM(Gradient boosting method)의 5개 알고리즘 적용하여 교통사고 분석을 수행하였다. 이러한 기계학습 알고리즘을 활용하기 위해 먼저 Training 과 Testing set을 7(5187 records):3(2222 records)의 비율로 구분하는 데이터 분할(Split)을 수행하였으며, 이 때 분할은 Random sampling을 활용하였다. 분석은 R의 기계학습용 Library인 'Caret'을 활용하여 동일한 기준에서 분석이 이루어질 수 있도록 설정하고 특히 Training set의 분석과정에서 교차검증을 위해 traincontrol을 10-fold repeated cross validation으로 설정하였다.

3. 분석 결과

3.1 종속변수 3 Classes 모델

본 연구에서 고려하는 종속변수인 사고심각도는 경상

(Minor), 중상(Severe), 사망(Fatality)의 세 범주로 구성하였다. 이는 TAAS 데이터에는 부상신고, 경상, 중상, 사망의 4개 범주로 제시되어 있으나 부상신고의 관측치가 65세이상 고령 운전자 사고의 1.8%를 차지하고 있어, Minority class에 해당되어 부상신고와 경상을 합쳐 경상으로 재분류하였다.

본 연구에서는 사고심각도(3개 클래스)를 분류하기 위해 다항로짓(Multinomial logit) 모형, KNN(K-nearest neighbor), 의사결정나무(Classification tree), Random Forest, ,GBM(Gradient boosting method), XGB(Extreme gradient boosting)의 6개 알고리즘을 고려하여 수행하였다.

분석결과 Ensemble 기반의 학습기법인 GBM과 XGB의 결과가 다른 알고리즘에 비해 분류성능이 다소 높은 것으로 파악 되었다. TAAS 원시데이터와 공간정보를 포함하여 확장한 데이터 셋과 비교했을 때, 정확도에서 유의미한 차이성을 보여주지 못하고 있다. 이는 모든 알고리즘이 minor class인 Fatality를 정확하게 분류하는 데 실패하고 있기 때문에 종속변수의 Class를 2개로 재분류하여 분석을 수행하고 이를 비교해 보는 것을 고려할 필요성이 도출되었다.

표 2. 모형별 분류성능 비교

Original dataset			New dataset	
Model	Train data	Test data	Train data	Test data
	Metric: Accuracy		Metric: Accuracy	
Multinomial Logit	0.73291	0.73144	0.73183	0.72785
KNN	0.71818	0.71975	0.70266	0.70355
Classification Tree	0.72566	0.7265	0.72928	0.7283
RF	0.7238	0.7211	0.72408	0.72155
GBM	0.73283	0.72785	0.73275	0.73234
XGB	0.73284	0.73099	0.73495	0.72785

3.2 종속변수 2 Classes 모델

본 연구에서 고려하는 종속변수인 사고심각도는 부상신고와 경상을 합쳐 경상으로, 중상과 사망을 합쳐 중상으로 재분류하여 두 범주로 구성하여 수행하였다. 본 연구 역시 6개의 기계학습 알고리즘을 적용하여 수행하였으며, 3개 클래스를 분석했을 때와 유사하게 Ensemble 기반의 학습기법인 GBM과 XGB의 결과가 타 알고리즘에 비해 분류성능이 다소 우수한 것으로 파악되었다. 한편, TAAS 원시데이터와 New dataset을 비교했을 때 분류성능은 큰 차이를 보여주지 못하고 있는 것을 알 수 있다. 이는 기존 3 classes 기반의 분석 사례에서 나타난 바와 같이 Minority class인 Fatality에 대한 분류의 정확도가 매우 낮기 때문인 것으로 판단된다.

표 2. 모형별 분류성능 비교

Original dataset			New dataset	
Model	Train data	Test data	Train data	Test data
	Metric: Accuracy		Metric: Accuracy	
Binomial Logit	0.73761	0.7274	0.73939	0.7256
KNN	0.71794	0.71345	0.70363	0.69951
Classification Tree	0.73442	0.7283	0.74132	0.7256
RF	0.72515	0.72605	0.73075	0.713
GBM	0.73785	0.73009	0.74078	0.72335
XGB	0.73905	0.7229	0.74147	0.72425

3.3 주요 인자 분석

분류에 영향을 미치는 중요한 인자 파악을 위해 본 연구에서는 XGB를 최적알고리즘으로 선정하여 표 4와 같이 Feature importance matrix를 도출하였다.

표 4. Variable importance matrix

No	Variables	Gain	Cover	Freq	Importance
1	**차대사람사고**	0.2995	0.1279	0.12	0.2995
2	사고발생연도	0.2284	0.1462	0.14	0.2284
3	**차대차-추돌**	0.0790	0.0849	0.08	0.0790
4	**도로연장**	0.0745	0.1157	0.12	0.0745
5	**신호위반**	0.0589	0.0597	0.06	0.0589
6	차량단독사고	0.0549	0.0596	0.06	0.0549
7	**차대차-정면충돌**	0.0313	0.0390	0.04	0.0313
8	가해운전자차종-승용차	0.0230	0.0393	0.04	0.0230
9	도로유형-기타	0.0209	0.0383	0.04	0.0209
10	**최고제한속도**	0.0183	0.0375	0.04	0.0183
11	**차로 수**	0.0169	0.0201	0.02	0.0169

표 4 결과에 따르면 차대 사람 사고, 차대차 추돌사고 및 신호위반, 최고제한속도, 차로 수 등이 사고심각도를 분류함에 있어 중요한 인자라고 제시한 것으로써 특히 공간정보를 추가함으로써 제시된 도로연장과 최고제한 속도 그리고 차로

수 변수가 주요 인자가 도출됨으로써 기존 TASS 데이터를 기반으로 하는 분석 연구의 한계점을 알 수 있었다. 하지만 이러한 인자들이 사고심각도에 어떠한 영향 즉 사고심각도를 높이는지 아니면 낮추는지에 대한 여부를 판단하지 못한다는 단점이 존재한다는 측면에서 기계학습 기반의 사고심각도 연구의 한계라 할 수 있다.

4. 결론

2030년 초고령사회로 진입하는 국내의 상황을 고려할 때, 고령운전자 증가와 고령운전자로 인한 교통사고 위험 해소는 시급히 해결해야 할 현안으로 대두되고 있다. 이에 국내에도 한국도로교통공사 교통사고분석시스템(TAAS)의 데이터를 활용하여 교통사고 저감대책 마련을 위한 기반은 이미 구축되었다고 할 수 있다. 하지만, 교통사고 심각도에 따라 효율적인 대처방안을 마련할 수 있는 빅데이터와 AI 기술을 활용한 교통사고 예측모델 개발이 필요하고. 또한, 이러한 교통사고 예측모델에는 지역특성이 고려된 다양한 교통연계 데이터 및 사회인구학적 데이터 환경변수 등을 적용하는 연구가 필수라 할 수 있다. 본 연구에서는 데이터마이닝 기법을 활용하여 대전광역시의 교통사고 지점의 위치정보를 포함한 지리정보 기반의 교통사고 데이터베이스를 구축하고 기계학습 기반의 교통사고 심각도 모형을 개발하였다. 이를 통해 기존 TASS 데이터만 활용한 모형과 사고지점 위치정보를 활용한 데이터를 포함한 새로운 모형을 통해 비교 분석하였으며, 결과적으로 본 연구에서 제시한 모형이 다소 높은 정확도를 보였다. 특히 사고심각도를 분류함에 있어 기존 TASS 데이터에 포함되지 않는 변수들이 중요한 인자들로 도출되었음을 확인하였다.

참고문헌

[1] 권철우, 장현호(2021). XGBoost를 활용한 이륜자동차 교통사고 심각도 비교분석. 한국ITS 학회 논문지

[2] 장재민, 최재성, 김태형(2016). 주행환경이 고령운전자의 교통사고 심각성에 미치는 영향 분석. 교통연구

[3] 이지원, 김태형(2019). 서울시 고령운전자 교통사고 특성 분석. 국토연구

[4] 보행자 교통사고 특성 및 감소방안(2020), 기본연구보고서, 대전세종연구원