나이브 베이즈 빅데이터 분류기를 이용한 렌터카 교통사고 심각도 예측

Prediction of Severities of Rental Car Traffic Accidents using Naive Bayes Big Data Classifier

정 하 림* · 김 홍 회** · 박 상 민*** · 한 음**** · 김 경 현***** · 윤 일 수******

* 주저자 : 아주대학교 건설교통공학과 석사과정
** 공저자 : 일마일주식회사 수석연구원
*** 공저자 : 아주대학교 건설교통공학과 박사과정
**** 공저자 : 도로교통공단 교통과학연구원 연구원
***** 공저자 : 한국도로공사 도로교통연구원 박사후연수연구원
****** 교신저자 : 아주대학교 교통시스템공학과 부교수

Harim Jeong* · Honghoi Kim** · Sangmin Park* · Eum Han*** · Kyung Hyun Kim**** · Ilsoo Yun*

* Dept. of Transportation Eng., Ajou University
** Ilmile Corp.
*** Traffic Science Institute, Road Traffic Authority
**** Transportation Research Division, Korea Expressway Corporation Research Institute

† Corresponding author :Ilsoo Yun, ilsooyun@ajou.ac.kr

Vol.16 No.4(2017)
August, 2017
pp.01~12

ISSN 1738-0774(Print)
ISSN 2384-1729(On-line)
https://doi.org/10.12815/kits.2017.16.4.01

Received 22 May 2017
Revised 8 June 2017
Accepted 20 July 2017

요 약

교통사고는 인적요인, 차량요인, 환경요인이 복합적으로 작용하여 발생한다. 이 중 렌터카 교통사고는 운전자의 평소 익숙하지 않은 환경 등으로 인해 교통사고 발생 가능성과 심각도가 다른 교통사고와는 다를 것으로 예상된다. 이에 본 연구에서는 국내 대표 관광도시인 부산광역시, 강릉시, 제주시를 대상으로 최근 빅데이터 분석에 사용되는 기계학습 기법중 하나인 나이브 베이즈 분류기를 이용하여 렌터카 교통사고의 심각도를 예측하는 모형을 개발하였다. 또한, 기존 연구에 유의성이 검증된 변수와 수집 가능한 모든 변수를 이용하는 두 가지 모형에 대하여 모형의 예측 정확도를 비교하였다. 비교 결과 통계적 기법을 통해 유의성이 검증된 변수를 사용할 경우 모형이 더 높은 예측 정확도를 보이는 것으로 나타났다.

핵심어 : 빅데이터, 렌터카, 교통사고, 심각도, 나이브 베이즈, 기계학습

ABSTRACT

Traffic accidents are caused by a combination of human factors, vehicle factors, and environmental factors. In the case of traffic accidents where rental cars are involved, the possibility and the severity of traffic accidents are expected to be different from those of other traffic accidents due to the unfamiliar environment of the driver. In this study, we developed a model to forecast the severity of rental car accidents by using Naive Bayes classifier for Busan, Gangneung, and Jeju city. In addition, we compared the prediction accuracy performance of two models where one model uses the variables of which statistical significance were verified in a prior study and another model uses the entire available variables. As a result of the comparison, it is shown that the prediction accuracy is higher when using the variables with statistical significance.

Key words : Big data, rental car, traffic accident, severity, Naive Bayes, machine learning

Ⅰ. 서 론

1. 연구의 배경 및 목적

우리나라의 교통사고 발생 건수는 꾸준히 감소하고 있지만 여전히 높은 수준이다. 우리나라에서 발생한 인구 10만 명당 교통사고 발생 건수는 1980년에 315.2건 이었으며, 2000년 617.9건의 교통사고 발생건수를 정점으로 기록한 후 현재까지 꾸준히 감소하는 추세를 보이고 있다. 국제 비교가 가능한 최신 년도인 2014년에는 인구 10만 명당 약 443.3건의 교통사고가 발생하여 2000년 교통사고 발생 건수에 비해 약 28%가 감소하였다. 하지만 2014년 OECD 국가별 자동차 1만 대당 교통사고 발생건수 추이를 살펴보면 핀란드 12.3건, 그리스 12.4건, 노르웨이 12.4건 등으로 나타났으나, 우리나라는 93.7건의 사고가 발생하여 OECD 회원국 평균인 40.2건과 비교하여 약 2.3배 높은 것으로 나타났다. 참고로 우리나라 2016년 인구 10만 명당 교통사고 발생 건수는 435.9건이다(Korea Road Traffic Authority, 2016). 이러한 자료를 통해 볼 때 우리나라의 교통사고 발생 건수는 여전히 심각한 수준이며, 교통사고 감소를 위해 더 많은 노력이 필요함을 알 수 있다.

교통사고는 인적요인, 차량요인, 환경요인에 의해 발생한다고 알려져 있으며, 이 세 가지 요인이 복합적으 작용하여 교통사고가 발생한다. 하지만, 실제 교통사고는 인적 요인에 의한 사고가 대다수로 2011년 국내에서 발생한 사망사고 중 인적요인 98.6%, 도로요인 25.6%, 차량요인 0.2%로 대부분의 심각한 사고는 인적요인에 의해 발생한 것으로 분석된 사례가 있다(Korea Transport Institute, 2013).

특히, 렌터카의 경우 운전자 자신이 차량을 소유하고 있지 않은 경우 또는 직접 소유한 차량이 있지만 여러 가지 이유로 인해 차량을 대여하여 평소 자신의 차량과 다른 차종을 운전하게 되는 경우가 많다. 따라서 평소 익숙하지 않은 차량과 도로구조, 신호체계 등의 환경 그리고 운전자 개인의 운전경험과 실력에 의해 교통사고가 발생할 확률과 심각도 또한 일반적인 경우와 다를 것으로 예상된다(Korea Transportation Safety Authority, 2013).

기존에는 교통사고 원인 도출 및 주요 요소 구분과 같은 분석에 포아송 회귀모형과 같은 통계적 기법들이 주로 사용되어졌다. 이러한 통계적 분석을 통해 교통사고와 관련된 다양한 요인들 간의 관계를 규명하였다.. 하지만, 최근에는 빅데이터 분석 기법을 이용하여 예측력 자체를 높이기 위한 많은 시도들이 있어왔다. 따라서 교통사고 분석에도 이러한 빅데이터 분석 기법들을 적용할 필요성이 있다고 판단된다.

이에 본 연구에서는 국내 대표 관광도시인 부산, 강릉, 제주시를 대상으로 2011년부터 2013년 동안의 렌터카 교통사고자료를 이용하여 렌터카 교통사고의 심각도를 예측하는 모형을 개발하였고, 이를 실제 사회에 활용할 방안을 제시하고자 한다. 또한, 기존 연구에 유의성이 검증된 변수와 수집 가능한 모든 변수를 이용하는 두 가지 모형에 대하여 모형의 예측 정확도를 비교하였다. 렌터카 교통사고 심각도 예측 모형 개발을 위해 최근 빅데이터 분석에 사용되는 머신러닝 기법을 사용하였고 분류 기법 중 하나인 나이브 베이즈 분류 기법을 사용하였다.

2. 연구의 범위 및 절차

1) 연구의 범위

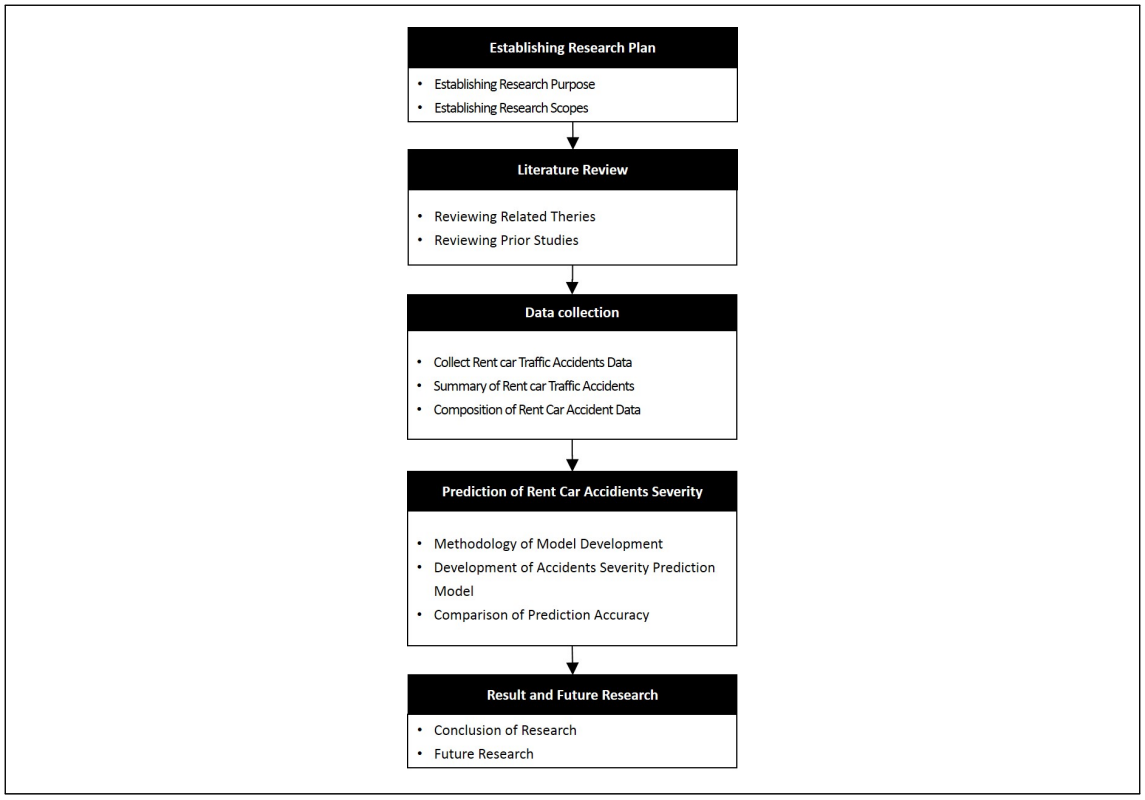
본 연구의 시간적 범위는 수집 가능한 교통사고 자료의 시간적 한계에 따라 2011년부터 2013년으로 설정하였다. 공간적 범위는 앞서 말한 바와 같이 렌터카 운전자에 작용하는 차량, 도로 등의 익숙하지 않은 환경의 영향이 잘 나타나는 관광지를 대상으로 하였고, 그 중 국내 대표 관광도시인 부산광역시, 강릉시, 제주시

를 공간적 범위로 선정하였다. 내용적 범위로는 렌터카 교통사고 자료를 이용한 렌터카 교통사고 심각도 예측으로 설정하였다.

2) 연구의 방법 및 절차

렌터카 교통사고 심각도 예측 모형 개발을 위해 <Fig. 1>과 같은 절차에 따라 연구를 수행하였다. 우선, 연구의 배경 및 목적, 범위 및 절차와 같은 연구의 계획을 설정하였다. 다음으로 관련 이론과 연구에 대한 고찰을 통해 기계학습 분류 기법 중 간단한 계산 과정을 요구함에도 불구하고 좋은 성능을 보이는 나이브 베이즈 분류 기법을 선정하였다. 다음으로 렌터카 교통사고 데이터에 나이브 베이즈 분류 기법을 적용하여 렌터카 교통사고 심각도 예측 모형을 개발하였다. 본 연구에서는 사용한 데이터인 렌터카 교통사고 자료는 운전자 특성, 기하구조 특성, 사고 특성으로 이루어져 있으며, 이러한 데이터의 특성 상 이상치를 정의하기 어려워 이상치를 제거하는 전처리과정을 생략하였다. 그리고 모형 개발에 사용한 입력 변수의 차이에 따른 분류 성능의 비교를 위해 전체 변수를 사용한 경우와 기존 통계적 분석 방법을 통해 유의성이 나타난 변수를 사용한 경우의 총 2가지 모형을 개발하였다. 또한 검증 데이터를 이용하여 각 모형의 예측 정확도를 비교하였고 이를 통해 두 모형의 성능을 비교하였다.

본 연구는 기존 포아송 회귀모형과 같은 통계적 분석 기법과 빅데이터 분석 기법의 성능을 비교하는 논문은 아니다. 즉, 기존에 다른 분야에서 활용되고 있는 빅데이터 분석 기법인 나이브 베이즈 분류 기법이 교통사고 예측에 적용될 수 있는 지에 대한 검증에 초점을 두고 있다는 점을 다시 한 번 밝히고자 한다.



〈Fig. 1〉 Research Process

Ⅱ. 관련 이론 및 연구 고찰

1. 기계학습

머신러닝(machine learning)으로 흔히 알려진 기계학습은 인공지능(artificial intelligence)의 한 분야이며, 패턴 인식과 컴퓨터학습 관련 연구로부터 진화하였다. 기계학습은 데이터를 기반으로 컴퓨터가 스스로 학습을 하고 예측을 하며, 성능을 향상시키는 시스템과 이를 위한 알고리즘을 구축하는 기술이다. 최근 구글(Google)에서 주최한 구글 딥마인드(deepmind) 바둑 대국에서 이세돌을 승리한 알파고(AlphaGo)를 학습시킨 딥러닝(deep learning) 또한 기계학습의 한 종류이다.

이러한 기계학습은 크게 지도학습(supervised learning)과 비지도학습(unsupervised learning)으로 나뉜다. 이 두 가지의 가장 큰 차이점은 학습시키는 학습에 주어진 데이터에 분류 항목 표시나 목적 변수의 유무 차이이다. 따라서 무엇을 예측하고 분류할 지를 알려주는 지도학습을 통해 각 데이터들을 분류하는 것이 가능하며, 대표적인 방법으로 나이브 베이즈(Naive Bayes), K-최근접 이웃(K-Nearest Neighbors, KNN), 지지 벡터 머신(Support Vector Machines, SVM) 등이 있다. 학습 데이터에 분류 항목이나 목적 변수가 없는 비지도학습을 통해서는 유사한 데이터들을 군집으로 모으는 것이 가능하며, 대표적으로 K-평균(K-Means), 디비스캔(DBSCAN), 기대 극대화(Expectation Maximization) 등이 있다.

적절한 기계학습 알고리즘을 선정하기 위해서는 먼저 목적에 대한 정의와 데이터에 대한 이해가 필요하다. 본 연구의 목적과 같이 어떠한 교통사고 심각도 예측과 같은 목적과 이를 위해 사용되는 데이터에 목적 변수의 존재에 따라 지도학습과 비지도학습 중 어떤 것을 사용할지 결정할 수 있다. 또한 데이터에 적합한 알고리즘을 선택하기 위해서는 각 알고리즘이 작동하는 과정을 이해해야 하며, 여러 알고리즘을 적용하여 성능을 비교하는 시행착오의 과정을 통해 가장 적절한 알고리즘을 선택할 수 있다(Peter, 2013).

2. 나이브 베이즈 분류기

나이브 베이즈 분류기(Naive Bayes Classifier)는 머신러닝의 지도학습을 사용한 가장 간단한 기법 중 하나이다. 나이브 베이즈 분류기의 장점으로는 모형이 단순하며, 계산 과정이 간단함에도 분류 성능이 우수하다는 점과 독립성 가정으로 인해 독립 변수의 차원이 늘어날 경우 늘어난 독립 변수의 수에 비해 기하급수적으로 많은 데이터를 필요로 하는 차원의 저주(curse of dimensionality) 문제를 완화 할 수 있다는 점이 있다. 단점으로는 단순히 학습에 사용하는 데이터를 이용하여 확률 값을 계산하여 그 크기를 비교하기 때문에 학습에 사용되는 데이터에 따라 분류 성능이 차이가 크게 난다는 특징이 있다.

나이브 베이즈 분류기는 사용되는 데이터의 모든 특성 값은 서로 독립임을 가정하며, 분류를 위해 베이즈의 정리(Bayes's Theorem)를 기본적으로 사용한다. 분류를 위해 입력한 데이터의 특성인 $x_1,\cdots,x_n$로 구성된 집합을 X, 각 데이터의 분류를 나타낸 값을 C_k로 가정할 경우, 새로운 정보인 X의 특성을 가진 데이터가 분류 C_k에 속할 확률은 베이즈의 정리를 이용하여 아래 식(1)과 같이 표현 가능하다. 또한, 베이즈의 정리는 이전의 경험과 현재의 자료를 바탕으로 어떤 새로운 사건의 확률을 추론하는 것이라고도 볼 수 있다(Park et al., 2013).

$$p(C_k|X) = \frac{p(X|C_k)p(C_k)}{p(X)} \tag{1}$$

새로운 정보가 각 분류에 속할 확률은 베이즈의 정리에 따라 식(1)의 우변과 같이 표현가능하며, 이를 계산하기 위한 과정에서 만약 데이터의 모든 특성 값에 대하여 독립성 가정을 하지 않을 경우 식(2)의 우변과 같이 각각의 특성이 다른 특성에 미치는 영향을 고려해야하는 복잡한 연산을 수행해야만 한다. 하지만 독립성 가정을 통해 식(3)과 같이 각 특성의 확률에 대한 곱으로 표현가능하며, 간단하게 계산할 수 있다(Park et al., 2013).

$$\begin{aligned} p(C_k, x_1, \cdots, x_n) &= p(C_k)p(x_1, \cdots, x_n|C_k) \\ &\quad p(C_k)p(x_1|C_k)p(x_2, \cdots, x_n|C_k, x_1) \\ &\quad p(C_k)p(x_1|C_k)p(x_2|C_k, x_1)p(x_3, \cdots, x_n|C_k, x_1, x_2) \\ &\quad p(C_k)p(x_1|C_k)p(x_2|C_k, x_1)\cdots p(x_n|C_k, x_1, x_2, x_3, \cdots, x_{n-1}) \end{aligned} \tag{2}$$

$$\begin{aligned} p(C_k|x_1, \cdots, x_n) &\propto p(C_k, \cdots, x_n) \\ &\quad p(C_k)p(x_1|C_k)p(x_2|C_k)p(x_3|C_k)\cdots \\ &\quad p(C_k)\prod_{i=1}^{n} p(x_i|C_k) \end{aligned} \tag{3}$$

나이브 베이즈 분류기는 아래 식(4)와 같이 각 데이터에 대하여 속할 수 있는 모든 분류에 대한 확률을 계산하여, 가장 높은 확률을 가지는 분류를 해당 데이터가 속할 분류로 예측하게 된다(Park et al., 2013).

$$\begin{aligned} C_{MAX} &= \arg max \ p(C_k|X) \\ &= \arg max \frac{p(X|C_k)p(C_k)}{P(X)} \quad \cdots\cdots\cdots (4) \\ &= \arg max \ p(X|C_k)p(C_k) \end{aligned} \tag{4}$$

이러한 나이브 베이즈 분류기는 기존 통계모형들이 중요시 하는 교통사고와 관련된 각 요인(elements) 또는 요인들 간의 관계 또는 중요도를 밝히는 데는 기여하는 바가 적으나, 그러한 관계들을 바탕으로 예측을 정확도를 높이는 데는 기존 통계모형들 보다 우수할 수 있을 것으로 판단된다.

3. 관련 연구

1) 교통사고 심각도 관련 연구

Choi et al.(2004)는 신호교차로 교통사고심각도 예측을 위해 인공신경망을 사용한 모형을 개발하였다. 모형의 검증을 위해 다중회귀모형을 개발하여 비교하였으며 그 결과 인공신경망을 사용한 사고심각도 예측모형이 더 뛰어난 예측력을 보였다.

Choi et al.(2011)는 초보 운전자가 보행자-차량 교통사고 발생 시 교통사고 심각도에 미치는 영향을 이항 로지스틱 회귀분석을 적용하여 분석하였다. 독립변수에는 교통사고 심각도에 영향을 미칠 것으로 판단

되는 성별, 연령, 법규위반 횟수 등을 설정하였고, 이를 운전자 운전 경력에 따라 분류하였다. 분석 결과, 운전경력에 관계없이 연령, 법규위반 횟수, 교통사고 위치가 교통사고 심각도 증가에 영향을 주는 것으로 나타났다.

Ko et al.(2016)는 국내 관광도시인 부산, 강릉, 제주시의 2011년부터 2013년까지의 렌터카 교통사고의 심각도를 분석하였다. 또한, 렌터카 교통사고와 함께 승용차 교통사고를 함께 분석하여 교통사고에 영향을 주는 요인을 비교하였다. 분석에는 포아송회귀모형과 음이항회귀모형을 이용하였고 모형 분석결과 모형의 로그우도함수값, AIC, BIC 등을 기준으로 음이항회귀모형이 더 적합한 것으로 나타났다. 렌터카 교통사고 심각도에 영향을 미치는 요인으로는 사고발생지역 거주 여부, 사고유형, 법규위반 등이 도출되었다. 사고발생지역 거주 여부의 경우 운전자의 거주지역과 사고발생지역이 같은 경우의 계수가 -0.0859로 비거주지역과 비교하여 사고심각도가 낮고, 사고유형의 경우 차대차와 차대사람 일 때의 계수가 각각 -0.2882와 -0.5384로 차량 단독 사고일 때보다 심각도가 낮은 것으로 분석되었다. 또한, 정면충돌, 진행 중 추돌, 주정차 중 추돌이 기타사고에 비해 심각도가 높으며 특히 정면충돌의 계수가 0.4294로 가장 높아 사고 심각도에 큰 영향을 주는 것으로 분석되었다.

Park et al.(2008)는 도시부 4지 신호교차로에서 발생한 교통사고 이력자료와 교차로 현장 조사자료를 이용하여 사고예측 모형 및 사고심각도 모형을 개발하였다. 사고심각도 모형 개발을 위해 순서형 프로빗 모형과 순서형 로짓모형의 2가지 모형을 개발하였으며 모형의 설명력과 적합성을 비교한 결과 사고심각도 모형에 순서형 프로빗 모형이 적합한 것으로 나타났다. 또한 교차로 사고심각도에 영향을 주는 변수로는 부도로 중차량 비율, 주도로 차량속도 제약시설, 부도로 차로당 평균폭으로 나타났다.

Won et al.(2009)은 2004년부터 2006년까지 서해안 고속도로에서 발생한 구간별 교통사고자료를 이용하여 사고 심각도 예측모형을 개발하였다. 개발된 모형을 이용하여 사고심각도에 영향을 주는 주요요인을 분석하였고 과속, 차량결함, 사고 유형, 교통량, 곡선반경 등에 따라 사고심각도가 달라지는 것으로 나타났다. 또한, 분석된 주요요인들을 통해 서해안 고속도로의 특정구간에서 주로 교통사고가 발생하며, 해당 구간에 구간단속 카메라, 차내·외 경고정보제공 등의 교통정보 및 시설을 제공하는 방안을 제시하였다.

2) 나이브 베이즈 분류기

나이브 베이즈 분류기의 경우 교통분야에서 사용된 사례가 없으며, 대부분 스팸문서 처리와 같이 문서를 분류하는 것을 목적으로 많이 사용되고 있는 것으로 나타났다.

Kim et al.(2000)는 나이브 베이즈 분류기를 이용하여 문서의 주제를 자동으로 구분하는 분류기를 개발하였다. 이를 위해 유즈넷 뉴스그룹의 문서들을 이용하였고 20개 분류에 대하여 각 분류당 1,000개씩 총 20,000개의 문서를 이용하였다. 뉴스그룹 분류 성능의 평가를 위해 전체 데이터의 30%를 실험데이터로 사용하였고 예측 정확도는 77.77%로 나타났으며, 학습에 사용되는 특성의 수가 20,000개 일 경우에 가장 높은 예측 정확도를 보였다.

Park et al.(2013)는 나이브 베이즈 분류기와 의사결정나무를 이용하여 유방암 진단 데이터를 분석하였고 두 가지 방법론의 성능을 비교하였다. 분석을 위해 UCI(University of California, Irvine)에서 제공하는 유방암 진단 관련 데이터 699개를 사용하였다. 분석 결과 나이브 베이즈 분류기가 96.0%로 의사결정나무에 비해 약 1.4%정도 높은 성능을 보였다.

Kang et al.(2016)는 특허 문서 분류를 위해 연구 분야와 관련성이 떨어지는 자료를 제거하는 과정을 기계학습을 통해 자동화 하고자 하였다. 이를 위해 나이브 베이즈, KNN, SVM 방법론을 이용하였고 방법론의 성

능을 비교하였다. 연구에 사용한 700여 건의 자료를 수집하여 분류한 결과 나이브 베이즈, KNN, SVM 순으로 분류 성능이 높게 나타났다.

Ⅲ. 자료 수집 및 구성

1. 렌터카 교통사고 자료 수집

본 연구에서는 2011년부터 2013년까지 국내 대표 관광도시인 부산광역시, 강릉시, 제주시에서 수집된 렌터카가 초래한 인사 교통사고 자료를 이용하였다. 분석기간인 2011년부터 2013년 동안 발생한 렌터카 교통사고는 총 1,858건으로 <Table 1>과 같다. 이 중 사고심각도별 발생한 사고는 사망사고 36건, 중상사고 718건, 경상사고 1,036건, 부상신고 68건이 발생한 것으로 나타났다. 지역별 발생한 사고는 부산시가 749건, 강릉시가 144건 그리고 제주시가 965건의 교통사고가 발생하여 분석 대상인 3개 도시 중 가장 많은 교통사고가 발생한 것으로 나타났다.

〈Table 1〉 Summary of Rent Car Traffic Accidents

Classification	Killed	Serious injured	Slightly injured	Minor injured	Total
Busan	7	309	411	22	749
Gangneung	4	40	95	5	144
Jeju	25	369	530	41	965
Total	36	718	1,036	68	1,858

2. 렌터카 교통사고 자료의 구성

렌터카 교통사고는 아래 <Table 2>와 같이 총 14개의 항목으로 구성되어 있다. 이 중 교통사고의 피해 정도에 따라 사망, 중상, 경상, 부상으로 구분한 사고심각도를 종속변수로 사용하였다. 이를 제외한 나머지 13개의 변수를 크게 동일거주지 거주 유무, 성별, 연령 등과 같은 운전자 인적사항과 주야, 교통사고 유형, 법규위반 유형과 같은 사고 특성 그리고 도로선형, 구배, 기상 상태, 노면 상태 등과 같은 도로 기하구조적 특성으로 구분할 수 있으며 렌터카 교통사고의 심각도를 예측하기 위한 독립변수로 사용하였다.

〈Table 2〉 Composition of Rent Car Accident Data

Variables			Variables Properties (Counts)
Dependant variable		Accident severity	Killed(36), Serious injured(718), Slightly injured(1,036), Minor injured(68)
Independantv ariable	Personal characteristic	Residence	Same residence(867), Other residence(991)
		Gender	Male(1,503), Female(352), Unknown(3)
		Age	10s(460), 20s(109), 30s(741), 40s(323), 50s(173), 60s or Older(49), Unknown(3)
		No. of years after getting a driving license	Less than 5yr(294), Less than 10yr(664), Less than 15yr(394), More than 15yr(423), Unlicensed(18), Unknown(65)

〈Table 2〉 Composition of Rent Car Accident Data (continued)

Variables			Variables Properties (Counts)
Independant variable	Accident characteristic	Time of day	Day(1,031), Night(827)
		Type I Accidents	Car-to-Person(170), Car-to-Car(388), Single Car only(1,300)
		Type II Accidents	Collision with structures(539), While passing on roadside(32), Running off the road other(15), Running off the road rolling down(43), While passing on sidewalk(129), Turning Over(201), Head-on collision(619), Collision with parked vehicle(165), Rear-end collision when parking or stopping(0), Rear-end collision when moving(0), While passing on road(0), Crossing collision(0), While crossing(0), Other(0)
		Type of Violation	Speed limit violation(31), Improper driving at intersection(7), Violation of pedestrian protection(140), Carelessness of pedestrian(82), Improper turning(9), Failure to slow down and temporarily stop(10), Violation of traffic signal(186), Driving too close to vehicle ahead(175), Infringement of safe driving(998), Violation of overtaking prohibition(0), Improper overtaking(0), Intrusion of median strip(109), Impending travelling for Straight or right turn(111), Failure to yield of course(0), Failure to drive on proper traffic passageway(0), Failure to yield of priority(0), Other(0)
	Geometric characteristic	Road Alignment	Straight(1,684), Left curve(75), Right curve(83), Other(16)
		Grade	Flat area(1,577), Ascent(166), Descent(99), Other(16)
		Weather Condition	Clear(1,518), Cloud(114), Rain(173), Snow(30), Fog(5), Unknown(99)
		Road Condition	Dry(1,556), Wet(228), Snow/Slush(13), Icy(17), Other(44)
		Signal operation	Signal operation(493), Non signal(1,282), Flashing(75), Lights out(8), Breakdown(0)

Ⅳ. 렌터카 교통사고 심각도 예측

1. 모형 개발 방법론

1) R 프로그램

R 프로그램은 뉴질랜드 오클랜드 대학에서 통계 및 데이터 분석을 위해 개발된 프로그래밍 언어이다. R의 장점으로는 첫째, 무료로 배포되고 있어 별도의 요금을 지불하고 구입하지 않아도 된다. 둘째, R에는 기본적으로 제공되는 기능에 패키지를 설치하여 원하는 기능을 추가할 수 있다. 이러한 패키지는 함수, 데이터, 컴파일된 코드 등을 모아놓은 것을 뜻하며, 인터넷 등을 통해 무료로 이용가능하다. 본 연구에서 사용한 패키지인 klaR은 기계학습의 지도학습에 속하는 나이브 베이즈 분류기와 지지 벡터 머신 분석 기능이 포함되어 있다(The R Foundation, 2017).

2) 성능 비교 방법 설정

렌터카 교통사고 데이터에 나이브 베이즈 분류기를 적용하기 위하여, 앞서 말한 데이터 분석에 널리 사용되는 프로그래밍 언어인 R을 사용하였고 R에서 제공하는 기계학습 라이브러리인 KlaR을 사용하였다. 또한,

아래 <Table 3>과 같이 나이브 베이즈 분류기에 사용한 독립변수에 교통사고 내용을 제외한 13개의 모든 변수를 사용한 경우와 Ko et al.(2016)의 연구를 통해 렌터카 교통사고 심각도에 영향을 주는 동일 거주지 여부, 교통사고 유형 I, 교통사고 유형 II, 법규위반, 도로구배의 5개의 변수만을 사용한 경우의 총 2가지 경우에 대하여 예측 정확도 비교를 하였다. 여기서 동일 거주지 여부란 렌트카 교통사고를 발생시킨 운전자가 해당 지역 거주자인 지 여부를 말한다. 교통사고 유형 I은 차대사람, 차대차, 차량단독 해당 여부를 말한다. 교통사고 유형 II는 공작물 충돌, 길가장자리 구역 통행 중, 도로 이탈 기타, 도로 이탈 추락, 보도 통행 중, 전도 전복, 정면충돌, 주정차차량 충돌, 주정차중 추돌, 진행중 추돌, 차도 통행 중, 측면직각 충돌, 횡단 중, 기타로 구성된다. 법규위반은 과속, 교차로 통행방법 위반, 보행자보호 의무 위반, 보행자 과실, 부당한 회전, 서행 및 일시정지 위반, 신호 위반, 안전거리 미확보, 안전운전 의무 불이행, 앞지르기금지 위반, 앞지르기방법 위반, 중앙선 침범, 직진 및 우회전차량의 통행 방해, 진로양보 의무 불이행, 차로 위반(진로변경 위반), 통행우선순위위반, 기타(운전자 법규위반)로 구성된다. 도로구배는 평지, 내리막, 오르막, 기타구역으로 구분된다 (Ko et al., 2016).

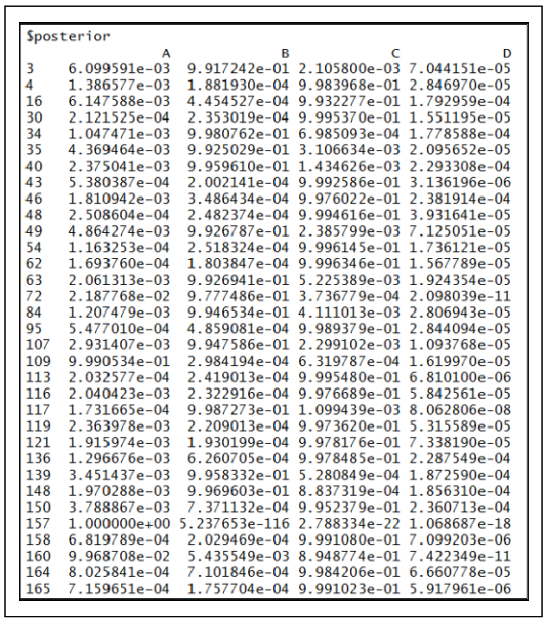
〈Table 3〉 Comparison of Variable Used for Model Development

Using All Variable		Using Significant Variable	
Residence	Gender	Residence	Type I accidents
Age	No. of years after getting a driving license	Type II accidents	Type of violation
Time of Day	Type I accidents	Grade	
Type II accidents	Type of violation		
Road Alignment	Grade		
Weather Condition	Road Condition		
Signal operation			

모형의 성능을 평가하기 위하여 전체 1,858개의 데이터 중 80.0%를 임의로 학습 데이터로 사용하였고, 나머지 20.0%를 평가 데이터로 지정하여 예측 정확도를 평가하였다. 이러한 과정을 각각 5회씩 반복하여 그 값을 산술평균한 값을 예측 정확도로 사용하였다.

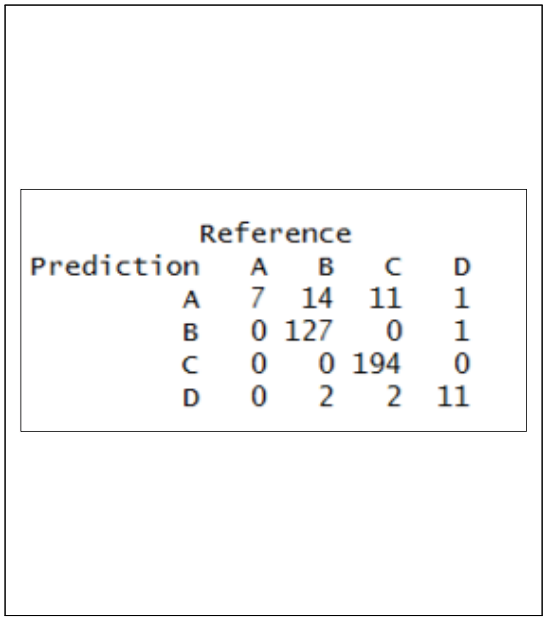
2. 교통사고 심각도 예측 모형 개발

앞의 과정에 따라 나이브 베이즈 분류기를 이용하여 렌터카 교통사고 심각도 예측 모형을 개발하였다. 개발된 모형은 아래 <Fig 2>과 같이 각 데이터별 4가지 사고심각도 분류에 속할 확률을 산출하게 되며, 이 중에 가장 큰 확률 값을 가지는 분류를 예측 값으로 선정하게 된다. 다음으로 <Fig 3>과 같이 실제 데이터의 값과 예측 값을 비교하여 혼동행렬(confusion matrix)로 표현 된다. 모형을 통해 기존 통계적 기법이 어떤 변수가 사고 심각도에 영향을 주는지에 분석이 가능한 반면 기계학습이라는 방법의 특성 상 변수에 대한 분석이 아닌 해당 데이터가 총 4가지 분류 중 어디에 속할 것인지에 대한 예측 값을 얻을 수 있다.



$posterior	A	B	C	D
3	6.099591e-03	9.917242e-01	2.105800e-03	7.044151e-05
4	1.386577e-03	1.881930e-04	9.983968e-01	2.846970e-05
16	6.147588e-03	4.454527e-04	9.932277e-01	1.792959e-04
30	2.121525e-04	2.353019e-04	9.995370e-01	1.551195e-05
34	1.047471e-03	9.980762e-01	6.985093e-04	1.778588e-04
35	4.369464e-03	9.925029e-01	3.106634e-03	2.095652e-05
40	2.375041e-03	9.959610e-01	1.434626e-03	2.293308e-04
43	5.380387e-04	2.002141e-04	9.992586e-01	3.136196e-06
46	1.810942e-03	3.486434e-04	9.976022e-01	2.381914e-04
48	2.508604e-04	2.482374e-04	9.994616e-01	3.931641e-05
49	4.864274e-03	9.926787e-01	2.385799e-03	7.125051e-05
54	1.163253e-04	2.518324e-04	9.996145e-01	1.736121e-05
62	1.693760e-04	1.803847e-04	9.996346e-01	1.567789e-05
63	2.061313e-03	9.926941e-01	5.225389e-03	1.924354e-05
72	2.187768e-02	9.777486e-01	3.736779e-04	2.098039e-11
84	1.207479e-03	9.946534e-01	4.111013e-03	2.806943e-05
95	5.477010e-04	4.859081e-04	9.989379e-01	2.844094e-05
107	2.931407e-03	9.947586e-01	2.299102e-03	1.093768e-05
109	9.990534e-01	2.984194e-04	6.319787e-04	1.619970e-05
113	2.032577e-04	2.419013e-04	9.995480e-01	6.810100e-06
116	2.040423e-03	2.322916e-04	9.976689e-01	5.842561e-05
117	1.731665e-04	9.987273e-01	1.099439e-03	8.062806e-08
119	2.363978e-03	2.209013e-04	9.973620e-01	5.315589e-05
121	1.915974e-03	1.930199e-04	9.978176e-01	7.338190e-05
136	1.296676e-03	6.260705e-04	9.978485e-01	2.287549e-04
139	3.451437e-03	9.958332e-01	5.280849e-04	1.872590e-04
148	1.970288e-03	9.969603e-01	8.837319e-04	1.856310e-04
150	3.788867e-03	7.371132e-04	9.952379e-01	2.360713e-04
157	1.000000e+00	5.237653e-116	2.788334e-22	1.068687e-18
158	6.819789e-04	2.029469e-04	9.991080e-01	7.099203e-06
160	9.968708e-02	5.435549e-03	8.948774e-01	7.422349e-11
164	8.025841e-04	7.101846e-04	9.984206e-01	6.660778e-05
165	7.159651e-04	1.757704e-04	9.991023e-01	5.917961e-06

〈Fig 2〉 Result of Probability Calculation



Prediction \ Reference	A	B	C	D
A	7	14	11	1
B	0	127	0	1
C	0	0	194	0
D	0	2	2	11

〈Fig 3〉 Confusion Matrix of Result

3. 예측 정확도 비교 결과

모형의 성능을 평가하기 위해 두 가지 모형에 대하여 앞서 말한 바와 같이 검증 데이터를 사용하여 5회씩 반복하여 산술평균하였고, 그 결과는 아래 <Table 4>와 같다. 전체 변수를 사용했을 경우 88.1%의 예측 정확도를 보였고, 기존 연구를 통해 교통사고 심각도에 유의한 영향을 주는 변수를 사용 한 경우는 94.1%의 예측 정확도를 보였다. 예측 정확도의 표준편차는 유의 변수를 사용한 경우가 더 낮은 것으로 나타났다. 이는 단순히 모든 변수를 사용할 때보다 1차적으로 통계적 기법을 통해 사고에 영향을 주는 변수만을 사용할 경우 더 우수한 예측성능을 가지는 것을 의미한다.

〈Table 4〉 Prediction Accuracy of Rent Car Accident Severity

Classification	Using All Variable	Using Significant Variable
Prediction Accuracy	88.1%	94.1%
Standard Deviation	7.5%	1.7%

이러한 나이브 베이즈 분류기는 기존 통계모형들이 중요시 하는 교통사고와 관련된 각 요인(elements) 또는 요인들 간의 관계 또는 중요도를 밝히는 데는 기여하는 바가 적다. 따라서 교통사고에 영향을 미치는 요인(elements)들의 상호 중요도 등을 확인하기 위해서는 나이브 베이즈 분류기가 아닌 다른 방법론을 고려해야 할 것으로 판단된다.

Ⅴ. 결론 및 향후 연구과제

1. 결 론

본 연구에서는 국내 대표 관광도시인 부산, 강릉, 제주시를 대상으로 2011년부터 2013년 동안의 렌터카 교통사고 자료와 나이브 베이즈 분류기를 이용하여 교통사고 심각도를 예측 모형을 개발하였다.

우선, 연구의 범위와 내용을 설정한 후, 관련 이론과 연구에 대한 고찰을 하였다. 다음으로 렌터카 교통사고 데이터에 나이브 베이즈 분류 기법을 적용하여 교통사고 심각도 예측 모형을 개발하였고 입력 변수의 차이에 따른 분류 성능의 비교를 위해 전체 변수를 사용한 경우와 유의 변수를 사용한 경우의 총 두 가지 모형을 개발하였고, 검증 데이터를 이용하여 각 모형의 예측 성능을 비교하였다.

나이브 베이즈 분류기를 이용하여 렌터카 교통사고 심각도를 예측한 결과 전체 변수를 사용했을 경우에도 88.1%의 어느 정도의 신뢰성을 보이지만, 통계적 기법을 통해 유의하다고 분석된 변수만을 사용할 경우 94.1%로 더 정밀하게 분류한 것으로 나타났다. 따라서 머신러닝 기법을 이용하여 분석할 경우에도 기존 통계적 분석 방법을 함께 이용할 경우 좀 더 나은 결과를 도출할 수 있을 것으로 예상된다.

본 연구를 통해 렌터카 교통사고 데이터를 이용하여 렌터카 교통사고의 심각도를 분류하였다. 이를 이용하여 렌터카 대여업체, 보험사 등과 같은 관련 업종에서 운전자 인적정보를 활용하여 사고발생 시의 심각도를 예측하여 보험료와 차량 대여료 조정에 참고할 수 있을 것으로 예상된다.

2. 향후 연구과제

본 연구에서는 렌터카 교통사고 데이터를 머신 러닝 기법중 하나인 나이브 베이즈 분류기를 통해 렌터카 교통사고 심각도 분류 모형을 개발하였고 교통사고 심각도 분류가 가능함을 확인하였다. 본 연구를 토대로 발전된 결과를 얻기 위해서 다음과 같은 연구가 필요하다.

첫 번째는 연구의 시간적 공간적 범위에 대한 확대가 필요하다. 본 연구에서는 데이터의 한계로 인해 2011년에서 2013년까지 3년 동안의 데이터를 사용하였다. 향후 연구에서는 교통사고의 시계열적 특성 및 차이점을 고려하기 위해 시간적 범위의 확대가 필요하며, 공간적 범위의 확대를 통해 렌터카에 대한 공간적인 대표성을 보완하는 것이 필요하다.

두 번째는 나이브 베이즈 분류기의 정확도를 좀 더 높이는 방안에 대한 연구가 필요하다. 본 연구의 결과인 94.1%로 대부분의 경우에 대하여 정확한 판단을 하고 있다고 볼 수 있지만, 실제 상황에 적용할 경우 제대로 분류하지 못한 케이스에 속하여 잘못된 판단으로 인한 불이익을 겪을 수 있다. 따라서 실생활에 적용하기 위해서는 분류 성능의 향상이 필요하다.

세 번째는 렌터카의 특성을 보다 잘 나타내기 위한 데이터 및 변수의 추가가 필요하다. 렌터카를 대여한 사람이 기존에 운전하던 차량과 대여한 차량의 차이에 대한 정보가 있다면 새로운 결과가 도출될 것으로 예상되며 이러한 렌터카의 특성을 보다 잘 반영할 수 있는 데이터를 추가한다면 사고의 유형과 연계하여 더 의미 있는 결과를 도출할 수 있을 것으로 예상된다.

네 번째는 모형 구축 시 사용할 변수 선정에 대한 연구가 필요하다. 본 연구에서는 기존 통계적 분석방법의 결과를 통해 선정된 변수를 사용하여 분류 성능이 향상된 것을 확인하였고 변수 선정에 따라 성능의 변화가 일어날 수 있는 가능성을 보였다. 하지만 이러한 방법이 모든 경우에 해당하는지에 대한 검증이 부족하며, 이와 관련한 추가적인 연구가 필요하다.

마지막으로 추가적인 머신러닝 기법을 이용한 연구가 필요하다. 본 연구에서는 머신러닝 기법 중 간단한

방법 중 하나인 나이브 베이즈 분류기를 이용하여 예측 정확도만을 분석하였다. 하지만, 추가적인 연구를 통해 렌터카 교통사고 심각도 분류 및 예측에 좀 더 나은 성능을 보이는 방법론과 데이터의 각 요소들이 교통사고심각도에 미치는 영향을 분석가능한 방법론을 이용하여 기존 통계적 분석방법의 결과와 비교하는 것도 의미있을 것으로 판단된다.

ACKNOWLEDGEMENTS

본 연구는 국토교통과학기술진흥원 교통물류연구사업(17LRP-B117133-02)과 2015년도 정부(교육과학기술부)의 재원으로 한국연구재단(NRF)의 기초연구사업(2015R1A1A1A05028008)의 지원을 받아 수행되었습니다. 본 논문은 한국ITS학회의 2017년도 춘계학술대회에 발표되었던 논문을 수정·보완하여 작성하였습니다.

REFERENCES

Choi J. W., Kim S. H., Cho J. H. and Kim W. C.(2004), “A Study to predict the Traffic Accident Severity Level Applying Neural Network at the Signalized Intersections,” *Journal of Korean Society of Transportation*, vol. 22, no. 3, pp.127-135.

Choi S., Park J. H. and Oh C.(2011), “Factors Affecting Injury Severity in Pedestrian-Vehicle Crash by Novice Driver,” *Journal of Korean Society of Transportation*, vol. 29, no. 4, pp.43-51.

Kang J. H., Kim J. C., Lee J. H., Park S. S. and Jang D. S.(2016), “A Comparative Study on Patent Document Classification Algorithms,” *Proceedings of KIIS Spring Conference*, vol 26, no. 1, pp.9-10.

Kim J. S. and Shin Y. K.(2000), “An Automatic Document Classification with Bayesian Learning,” *Journal of the Korean Data & Information Science Society*, vol. 11, no. 1, pp.19-30.

Ko H. G., Yun I., Kim K. H., Song H. I. and Heo, T. Y.(2016), “A study on Analysis Severities of Rental Car Traffic Accidents : Case of Major Sightseeing Cities Including Busan, Gyeongju and Jeju Island,” *Journal of the Korean Data Analysis Society*, vol. 18, no. 2, pp.755-769.

Korea Road Traffic Authority(2016), Comparison of Traffic Accident of OECD Member States.

Korea Transport Institution(2013), A Study on the Strategies for ‘Vision Zero’ Goal of Traffic Fatalities in Korea.

Korean Transportation Safety Authority(2013), The twenties cause the half of the entire deadly traffic accidents involving rent cars, analysis of status of deadly rent car traffic accidents during recent 5 years.

Won M. S., Lee G. R., Oh C. and Kang K. W.(2009), “A Study on the Application of Accident Severity Prediction Model,” vol. 27, no. 4, pp.167-173.

Park J. T., Lee S. B., Kim J. W. and Lee D. M.(2008), “Development of a Traffic Accident Prediction Model for Urban Signalized Intersections,” *Journal of Korean Society of Transportation*, vol. 26, no. 4, pp.99-110.

Park N. Y., Kim J. I. and Jung Y. G.(2013), “Breast Cancer Diagnosis using Naive Bayes Analysis Techniques,” *Journal of Service Research and Studies*, vol. 3, no. 1, pp.87-93.

Peter H.(2013), “Machine Learning in Action,” JPub(Paju, Korea), pp.11-13.

The R Foundation, https://www.r-project.org/, 2017.05.16.