Proceedings of Conference on Knowledge Information Technology and Systems

http://www.kkits.or.kr

Analysis of Influential Factors on Traffic Accidents Using EM Algorithm

SeungYup Baek, JangHun Kwon, Jungeun Kim*

Department of Computer Science and Engineering, Kongju National University

ABSTRACT

The traffic accident is one of major social issues that causes enormous costs, and Korea is paying an annual cost of 40 trillion won due to traffic accidents. In this paper, we analyze the factors and types of traffic accidents using large-scale traffic accident data set generated by the Seoul Metropolitan Government from 2010 to 2018. The entire dataset was clustered using the EM algorithm, and based on clustering results, major types of traffic accident were discovered. In addition, major factors of traffic accidents were analyzed through the PCA for each cluster.

KEYWORDS Traffic accident, Clustering, EM algorithm, Bayesian information criterion, PCA

1. 서 론

교통사고는 막대한 인적, 물적 비용을 초래하는 사회적 문제로 한국은 연간 40조 5천억원에 달하는 막대한 비용을 치루고 있다. 따라서 교통사고의 원인을 파악하고 궁극적으로 교통사고를 예방하는 것은 중요한 공공 문제로 인식되어오고 있다.

본 논문에서는 대규모의 교통사고 데이터 분석을 통해 교통사고의 요인과 유형을 분석하는 것을 목표로 하며 전체 구성은 다음과 같다. 제2장에서는 논문에서 사용한 데이터 셋의 특성에 대해 기술한다. 제3장에서는 EM알고리즘과 주성분 분석을 통한 교통사고 요인 분석에 대해 기술하고, 제4장에서 결론을 요약한다.

2. 데이터 셋

본 논문에서는 서울시에서 발생한 2010년부터 2018년까지의 개별사고정보 데이터를 사용하였다[4]. 데이터에 결측값이나 이상값 등이 포함된 경우는 데이터 전처리를 통해 제외하였고 사고유형 대

*Corresponding author is with the Department of Computer Science & Engineering, Kongju National University, 1223-24, Cheonan-daero, Seobuk-gu, Cheonan-si, Chungcheongnam-do, 31080, KOREA.
E-mail address: jekim@kongju.ac.kr

분류 또는 사고유형 중분류와 같이 의미상 중복인 변수들은 대표 변수 하나만을 채택하였다. 발생시간 데이터는 08시~18시는 주간, 18시~08시는 야간으로 범주화하였다. <표 1>은 본 논문에서 사용한 8개의 변수를 요약한 것으로 구체적으로는 발생시간, 요일, 사고내용, 사고유형 대분류, 가해자법규위반, 노면상태, 기상상태, 도로형태를 포함한다.

표 1. 논문에서 사용된 데이터 셋의 변수
Table 1. Variables of dataset used in this paper

발생시간						
주간				야간		
1				2		
요일						
월	화	수	목	금	토	일
1	2	3	4	5	6	7
사고내용						
경상		중상		사망		부상신고
1		2		3		4
사고유형_대분류						
차대차		차대사람		차량단독		철길건널목
1		2		3		4
가해자법규위반						
안전운전의무불이행	안전거리미확보	신호위반	교차로통행방법위반	중앙선침범	보행자보호의무위반	과속
1	2	3	4	5	6	7
노면상태						
건조	젖음/습기	서리/결빙	적설	침수	해빙	기타/불명
1	2	3	4	5	6	0
기상상태						
맑음	흐림	비	안개	눈	기타/불명	
1	2	3	4	5	0	
도로형태_대분류						
단일로		교차로		철길건널목		기타/불명
1		2		3		0

3. 데이터 분석

3.1 방법론

본 논문에서는 교통사고의 요인을 분석하기 위해 EM 알고리즘 (Expectation Maximization Algorithm)을 이용하여 데이터를 클러스터링하고 클러스터별 규칙을 찾는다. 최적의 클러스터 수를 결정하기 위해 베이즈 정보기준 (Bayesian Information Criterion)을 활용하였으며 주성분 분석 (Principal Components Analysis)을 통해 변수 간의 상관관계와 변수가 교통사고에 끼치는 영향의 정도를 분석하였다.

3.2 EM 알고리즘 결과

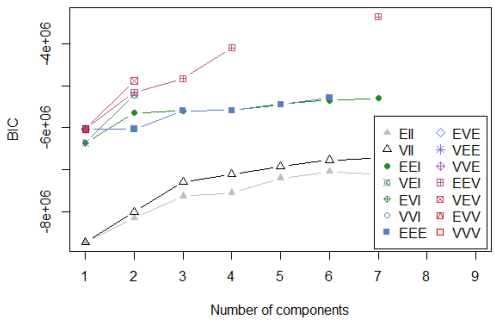

그림 1. 베이즈 정보기준 (BIC)
Figure 1. Bayes Information Criteria (BIC)

<그림 1>은 베이즈 정보 기준으로 최적의 클러스터링 개수를 판단할 때 사용되는 지표이다. 우리는 이 지표를 기준으로 최적 클러스터의 수를 7로 결정하였다. <표 2>는 각 클러스터의 변수별 평균값으로 클러스터별 교통사고 데이터의 특성을 나타낸다. 우리는 이러한 특성을 기준으로 교통사고의 몇 가지 주요 유형을 도출할 수 있었다.

표 2. 클러스터 변수 분포
Table 2. Cluster variable distribution

클러스터	1	2	3	4	5	6	7
발생시간	1.9764	1.9717	1.9677	1.9576	1.9384	1.9732	1.9521
요일	1.5074	3.4882	2.1115	5.5259	4.0386	5.0245	6.4250
사고내용	1.4323	1.3519	1.5414	1.5356	1.9074	1.3041	1.4410
사고유형_대분류	1.3402	1.2925	1.2888	1.2532	1.5719	1.2648	1.3292
가해자법규위반	1.0773	1.0957	3.9919	3.9336	1.1957	1.0926	1.0845
노면상태	1.0174	0.9978	1.0901	1.0812	1.7100	0.9967	1.0172
기상상태	1.0500	1.0225	1.1861	1.1709	2.2593	1.0187	1.0514
도로형태_대분류	1.2729	1.2767	1.6392	1.6488	1.2881	1.2752	1.2697

1. 도로의 상태가 건조하고 맑은 날씨 야간에 안전운전 의무 불이행 차대차사고가 단일로에서 일어났을 때 경상이 많았다. (클러스터 1, 2, 6, 7)
2. 도로의 상태가 건조하고 맑은 날씨 야간에 교차로 통행 방법 위반 차대차사고가 교차로에서 일어났을 때 경상, 중상의 사고가 일어났다. (클러스터 3, 4)
3. 도로의 상태가 습하고 흐림, 비오는 날씨 야간에 안전운전 의무 불이행 차대차사고가 단일로에서 일어났을 때 중상의 사고가 일어났다. (클러스터 5)

3.3 주성분 분석 결과

클러스터별 교통사고의 주요 요인이 무엇인지 도출하기 위해 주성분 분석을 실행하였다. 주성분 분석의 핵심은 데이터의 차원을 손실 없이 최소화하여 데이터를 압축하는 것으로 이 과정에서 주성분에 가장 영향을 끼치는 변수를 찾을 수 있다는 장점이 있다.

하지만 본 논문의 실험에서는 제1주성분의 기여율이 전체 데이터를 대표할 만큼의 결과가 나오지 못했고, 제4주성분까지 고려한 결과 클러스터 5에서는 요일, 기상상태, 노면상태가 교통사고에 가장 많은 영향을 미친 변수라는 결과를 얻었다. 하지만 나머지 클러스터에서는 어떤 변수도 교통사고와 연관 있는 주요 요인으로 특정하기 어려웠다.

4. 결 론

본 논문에서는 2010년부터 2018년도까지 서울시에서 발생한 교통사고 데이터를 사용하였고 최종적으로 8개의 변수를 선정하여 연구를 진행하였다. EM알고리즘을 사용하여 분석해본 결과 3가지 규칙들을 찾아낼 수 있었다. 하지만 주성분 분석을 통해 주요 요인을 찾아내는 실험에서는 만족스러운 결과를 얻지 못했다. 이는 교통사고에 영향을 미치는 요인이 매우 다양할 수 있기 때문이라 생각되며 더 폭넓고 구체적인 정보를 포괄하는 데이터셋의 구축과 공개가 필요하다고 생각된다.

References

[1] H. Y. Kim, Feature Selection and Visualization Based on Expectation Maximization and Principal Component Analysis for Traffic Accident Point Prediction Vol. 14, No. 4, pp. 13-23, 2019.

[2] J. T. OH , I. S. Yun, , J. W. Hwang , and E. Han, A Comparative Study On Accident Prediction Model Using Nonlinear Regression And Artificial Neural Network, Structural

Equation for Rural 4-Legged Intersection, pp. 266-279, 2014.
[3] Model-based clustering and Gaussian mixture model in R https://en.proft.me/2017/02/1/model-based-clustering-r/
[4] Traffic Accidents in Seoul https://www.data.go.kr/index.do
[5] Data Analytics with R, Oh Sejong, Hanbit Academy, 2020.

기댓값 최대화 알고리즘을 이용한 교통사고 주요 요인 분석

백승엽, 권장훈, 김정은
공주대학교 컴퓨터공학부

요 약

본 논문에서는 2010년부터 2018년도까지 서울시에서 발생한 교통사고 데이터를 EM 알고리즘을 이용하여 클러스터링하고 그 결과를 기반으로 몇 가지 주요 사고 유형을 발견하였으며 클러스터별 주성분 분석을 통해 교통사고의 주요 요인을 분석하였다.

감사의 글

이 논문은 2020년 국립대학 육성사업(교육부)의 재원으로 공주대학교의 지원을 받아 수행된 연구임.