Research
doi: http://dx.doi.org/10.16925/in.v13i21.1726

A Novel Framework to Use Association Rule Mining for Classification of Traffic Accident Severity

Meenu Gupta[1], Vijender Kumar-Solanki[2], Vijay Kumar-Singh[3]

[1] Research Scholar, Ansal University, Gurgaon, Haryana, India
[2] PhD in Engineering. Assistant Professor, Institute of Technology & Science, Mohan Nagar, Ghaziabad, UP, India
E-mail: spesinfo@yahoo.com
[3] Dean Academics, Ansal University, Gurgaon, Haryana, India

Received date: September 15, 2016 **Accepted date:** December 5, 2016

How to cite this article: M. Gupta, V. Kumar-Solanki & V. Kumar-Singh, "A Novel Framenwork to Use Association Rule Mining for Classification of Traffic Accident Severity", *Ingeniería Solidaria*, vol. 13, n.° 21, pp. 37-44, January 2017. doi: http://dx.doi.org/10.16925/in.v13i21.1726

Abstract. *Introduction*: Traffic accidents are an undesirable burden on society. Every year around one million deaths and more than ten million injuries are reported due to traffic accidents. Hence, traffic accidents prevention measures must be taken to overcome the accident rate. Different countries have different geographical and environmental conditions and hence the accident factors diverge in each country. Traffic accident data analysis is very useful in revealing the factors that affect the accidents in different countries. This article was written in the year 2016 in the Institute of Technology & Science, Mohan Nagar, Ghaziabad, UP, India. *Methology:* We propose a framework to utilize association rule mining (ARM) for the severity classification of traffic accidents data obtained from police records in Mujjafarnagar district, Uttarpradesh, India. *Results*: The results certainly reveal some hidden factors which can be applied to understand the factors behind road accidentality in this region. *Conclusions*: The framework enables us to find three clusters from the data set. Each cluster represents a type of accident severity, i.e. fatal, major injury and minor/no injury. The association rules exposed different factors that are associated with road accidents in each category. The information extracted provides important information which can be employed to adapt preventive measures to overcome the accident severity in Muzzafarnagar district.

Keywords: association rule mining, classification, severity analysis, traffic accident.

ISSN 1900-3102 (impreso) / ISSN 2357-6014 (en línea)

doi: http://dx.doi.org/10.16925/in.v13i21.1726

Nuevo marco para utilizar la minería de datos y reglas de asociación para la clasificación de la gravedad de accidentes de tráfico

Resumen. *Introducción*: los accidentes de tránsito son una carga indeseable para la sociedad. Cada año se reportan alrededor de un millón de muertes y más de diez millones de lesiones debido a accidentes de tráfico. Por lo tanto, se deben implementar medidas de prevención de accidentes de tráfico para superar la tasa de accidentalidad. Los países tienen diferentes condiciones geográficas y ambientales y, por ello, las variables que inciden varían en cada país. El análisis de los datos de accidentes de tráfico es muy útil para revelar los factores o variables que inciden en la accidentalidad en diferentes países. Este artículo fue escrito en el 2016 en el Instituto de Tecnología y Ciencia, Mohan Nagar, Ghaziabad, UP, India. *Metodología*: proponemos un marco para utilizar la minería de datos y reglas de asociación (ARM) para la clasificación de severidad de los datos de accidentes de tráfico obtenidos de registros policiales en el distrito de Mujjafarnagar, Uttarpradesh, India. *Resultados*: los resultados revelan ciertamente algunos factores ocultos que se pueden aplicar para entender las variables detrás de la accidentalidad de tráfico en esta región. *Conclusiones*: el marco permite establecer tres categorías en el conjunto de datos que representan el tipo de gravedad del accidente: fatal, lesiones graves, y lesiones menores o inexistentes. Las reglas de asociación expusieron diferentes factores relacionados con los accidentes de tráfico en cada categoría. Los datos extraídos proporcionan información importante que se puede emplear para adaptar las medidas preventivas para superar la gravedad de los accidentes de tráfico en el distrito de Muzzafarnagar.

Palabras clave: minería de datos, reglas de asociación, clasificación, análisis de gravedad, accidente de tráfico.

Novo referencial para utilizar a mineração de dados e regras de associação para a classificação da gravidade de acidentes de trânsito

Resumo. *Introdução*: os acidentes de trânsito são uma carga não desejável para a sociedade. A cada ano, são relatados ao redor de um milhão de mortes e mais de dez milhões de lesões devido a acidentes de trânsito. Portanto, devem-se implantar medidas de prevenção desses acidentes para superar a taxa de acidentabilidade. Os países têm diferentes condições geográficas e ambientais e, por isso, as variáveis que incidem variam em cada país. A análise dos dados de acidentes de trânsito é muito útil para revelar os fatores ou variáveis que incidem na acidentabilidade em diferentes países. Este artigo fio escrito em 2016 no Instituto de Tecnologia e Ciência, Mohan Nagar, Ghaziabad, UP, Índia. *Metodologia*: propomos um referencial para utilizar a mineração de dados e regras de associação (arm) para a classificação de severidade dos dados de acidentes de trânsito obtidos de registros policiais no distrito de Mujjafarnagar, Uttarpradesh, Índia. *Resultados*: os resultados revelam certamente alguns fatores ocultos que podem ser aplicados para entender as variáveis por trás da acidentabilidade de tráfego nessa região. *Conclusões*: o referencial permite estabelecer três categorias no conjunto de dados que representam o tipo de gravidade do acidente: fatal, lesões graves e lesões menores ou inexistentes. As regras de associação expuseram diferentes fatores relacionados com os acidentes de trânsito em cada categoria. Os dados extraídos proporcionam informação importante que pode ser empregada para adaptar as medidas preventivas para superar a gravidade dos acidentes de trânsito no distrito de Muzzafarnagar.

Palavras-chave: mineração de dados, regras de associação, classificação, análise de gravidade, acidente de trânsito.

1. Introduction

The fatalities, injuries and different types of property damage due to traffic accidents provide a tremendously bad impact on socio-economic development. Traffic accidents are one of the leading causes of deaths and injuries across the world [1]. The World Health Organization (WHO) stated in its global safety report that 1.2 million deaths and more than 4 million severe injuries took place each year worldwide, which is a big concern [2]. One of the primary world interests is to adopt some preventive measures so that accidents can be avoided [3]. Although, it is rather impossible to fully prevent accidents, some preventive measures can be taken. Data analysis can certainly help us to identify factors that affect the severity of road accidents [4].

Data mining [5] is a set of techniques that can be used to analyze data and extract meaningful information [6]. Different data mining techniques [7] such as classification, clustering, association rule mining and anomaly detection have been proved to provide productive results in several domains. Also, various traffic safety studies use different data mining techniques and reveal some hidden factors. In this paper, we are going to use data mining techniques to analyze traffic accident data obtained from police records. The organization of the paper is as follows: next section will discuss about related work in traffic safety domain, section three will discuss a proposed framework and section four will discuss the experimental results to finally conclude in the last section.

2. Literature Review

There are several studies in the traffic safety domain that analyze traffic accident data from different countries. It is rather difficult to discuss the whole literature; we will discuss some of the most important studies that analyze traffic accident data in this section of the paper.

Initially, regression based techniques were widely used for traffic accident analysis, [8] used logistic regression to analyze accident data. In this study authors found that residential sites and shopping sites are more hazardous for traffic accidents than rural sites. Another study [9] used neural network to analyze the road accident data. They found that non-signalized intersections are more dangerous to pedestrian hit accidents during night time. Also, [10] analyzed traffic accident frequency data using tree-based models and negative binomial regression models. They found that tree-based models performed better on road accident data.

Some studies claim that heterogeneity can be a big problem in analyzing traffic accident data as some accident factors remain hidden in the presence of heterogeneity. Ona et al. [11] used latent class clustering (LCC) to remove the heterogeneity from traffic accident data and their results prove that clustering prior to analysis of data reduces the level of heterogeneity and provides better outcomes. Further, [12] used k-modes clustering and claimed that k-modes clustering works faster than LCC and provides good results. They also claim that clustering prior to heterogeneity removes the heterogeneity from the data. In another study, [13] compared both LCC and k-modes clustering techniques over a new road accident data set. They found that both techniques are good in order to perform the clustering except that k-modes perform faster than LCC. Some studies [14-16] show that traffic accident frequencies are also location dependent and some locations are more prone to accidents while others are not.

Abellan et al. [17] developed various decision trees to extract different decision rules for different trees to analyze two-lane rural highway data in Spain. They found that bad light conditions and safety barriers badly affect the crash severity. Tesema et al., [18] used an adaptive regression tree model to build a decision support system for the road accidents in Ethiopia. Kashani et al. [19] used classification and a regression tree (CART) to analyze road accidents data from Iran and found that not using seat belt, improper overtaking and speeding badly affect the severity of accidents. Depaire et al. [20] used a clustering technique to analyze road accident data form Belgium and suggest that cluster-based analysis of road accident data can extract better information rather than analyzing data without clustering. Kwon et al. [21] used Naïve Bayes and the decision tree classification algorithm to analyze factor dependencies related to road safety. The severity of accidents is directly concerned with the victim involved in accidents, and its analysis only targets this type of severity and shows the circumstances that affect the injury severity of accidents.

All the above studies show that data mining techniques perform better than traditional techniques to analyze traffic accident data. Hence, in this

study we are using data mining techniques such as association rule mining (ARM) [22] to analyze traffic accident data. We also propose a framework to demonstrate how we can use ARM as a classification technique. Further, we perform severity analysis of traffic accident data and reveal some factors that affect different levels of severity.

3. Proposed Framework for Classification using ARM

The proposed framework to use ARM for the classification of traffic accident data is illustrated in Figure 1.

The different phases of framework are discussed as follows.

3.1 Traffic Accident Data

Traffic accident data for this study was obtained from police records of Muzzafarnagar district of Uttarpradesh, India. All police records were in document form containing 1,470 traffic accident records for the year 2015. The attributes selected and their possible values for the analysis are mentioned in table 1.

3.2 Data Preprocessing

Data preprocessing is the most important part in data mining process. The initial data obtained was in raw form (i.e., recorded on paper document). Afterwards, different possible values were extracted from the physical data. Also, data transformation was applied to convert some of the attributes into a suitable form for analysis. Different values that were assigned to different attributes after preprocessing are mentioned in Table 1.

3.3 Cluster Formation

Cluster formation is the most important step of this framework. ARM is a popular data mining technique based on market basket analysis. ARM extract different attribute values that occurre frequently in certain accidents and that may reveal different factors associated with road accidents. But it cannot specifically tell to which kind of severity does certain factor belong. Hence, it cannot be used for classification purposes because classification needs a class value to be predicted. As we have three different severity levels in our traffic accident data, we can segment our data into three different clusters based on accident severity, namely: fatal,

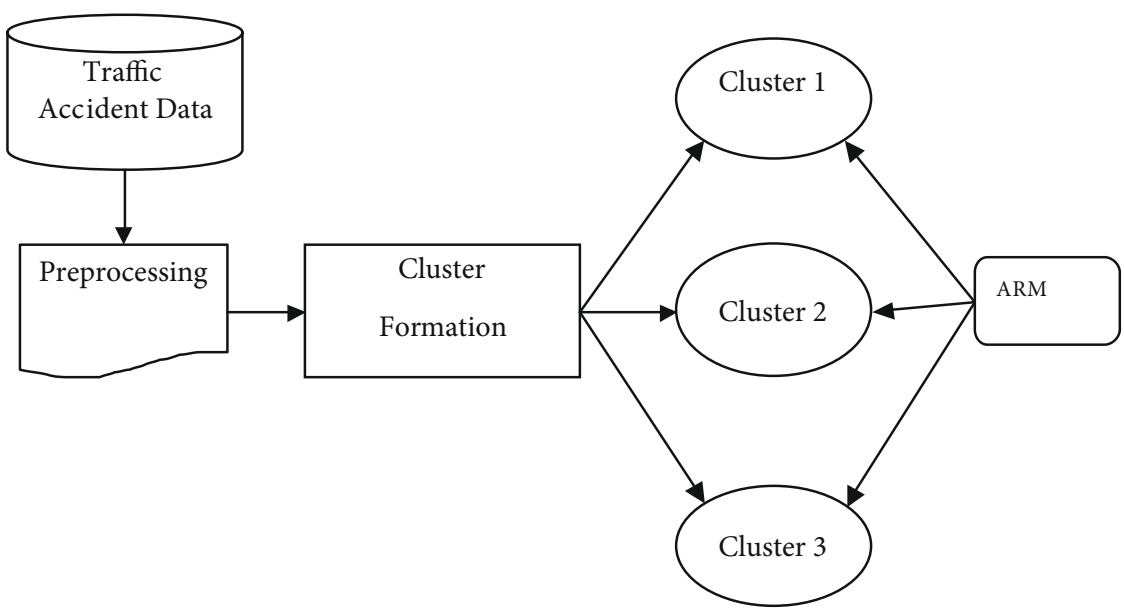

Figure 1. Proposed framework for severity analysis of traffic accidents
Source: Compiled by the authors

Table 1. Description of relevant attributes for analysis

Serial No.	Attribute Name	Assigned code	Different values
1.	Road type	ROT	Highway=1, Local=2
2.	Number of lanes	NOL	1 lane=1, 2 lanes=2, 4 lanes=4
3.	Road feature	ROF	Intersection=1, Curve=2, Straight=3
4.	Surrounding area	SUA	Residential=1, Agricultural=2, Industrial=3, Commercial=4
5.	Light condition	LIG	Day_light=1, road_light=2, no_light=3
6.	Accident severity	SEV	Fatal=1, major= 2, No_injury/minor=3
7.	Age of victim	AGE	0-18=1, 18-40=2, 40+=3
8.	Time of accident	TIM	0-6 hours=1, 6-12 hours=2, 12-18 hours=3, 18-24 hours=4
9.	Day	DAY	Weekday=1, Weekend=2
10.	Month	MON	Jan=1, Feb=2, …., Dec=12

Source: Compiled by the authors

major injury and minor/no injury. Additionally, we performed ARM on each cluster and extracted association rules. These association rules belong only to the cluster in which they appeared. Hence, we can tell that an accident factor appeared in a determined cluster and affects that severity level. In this way, we may classify our data using ARM.

3.4 Association Rule Mining

ARM is a popular data mining technique based on market basket analysis. Association rule mining produces a set of rules that define the underlying patterns in the data set. Given a data set D of n transactions where each transaction $T \in D$. Let $I = \{I_1, I_2, .., I_n\}$ is a set of items. An item set A will occur in T if and only if $A \subseteq T$. $A \rightarrow B$ is an association rule, provided that $A \subset I$, $B \subset I$ and $A \cap B = \emptyset$. In case of road accident data, an association rule can identify the various attribute values responsible for an accident occurrence.

In association rule mining, various interesting measures [16] exist to assess the quality of a rule (i.e. support, confidence and lift).

Support (S_p): The support of a rule $A \rightarrow B$ defines the percentage of how often A and B occur together in a data set and can be calculated using Equation 1. Support is also known as frequency constraint. A set of items satisfying certain support threshold is known as frequent item set. These frequent item sets are further used to generate association rules based on other measures.

Confidence (C_f): Confidence of a rule $A \rightarrow B$ defines the ratio of the occurrence of A and B together to the occurrence of A only and can be calculated using Equation 2. The higher the confidence values of a rule $A \rightarrow B$, the higher the chances of occurrence of B with the occurrence of A. Sometimes, only confidence values are not sufficient to evaluate the descriptive interest of a rule.

Lift (L_t): Lift for a rule $A \rightarrow B$ measures the expected occurrence of A and B together. In other words, lift is the ratio of the confidence and the expected confidence of a rule. Expected confidence can be defined as the occurrence of A and B together with the occurrence of B. A lift value ranges from 0 to ∞. Lift values greater than 1 make a rule potentially useful for predicting the appearence in future data sets. Lift determines how far from independence are A and B. Lift measures co-occurrence only and is also symmetric with respect to A and B. Lift can be calculated using Equation 3.

The formula to calculate these values for a rule $A \rightarrow B$ is given as Equations 1-3 show.

$$Support\left(S_p\right)=\frac{P\left(A\cap B\right)}{N} \qquad \textit{Equation no. 1}$$

$$Confidence\left(C_f\right)=\frac{P\left(A\cap B\right)}{P\left(A\right)} \qquad \textit{Equation no. 2}$$

$$Lift\left(L_f\right)=\frac{P\left(A\cap B\right)}{P\left(A\right)\times P\left(B\right)} \qquad \textit{Equation no. 3}$$

4. Results and Discussion

4.1 Preliminary Analysis

Initially, the preprocessed data set was grouped into three different clusters based on the severity level. As there were three severity levels in the data set, three clusters were formed. The description of the clusters is shown in Table 2.

Table 2. Cluster Description

Cluster Id	Cluster name	Number of accidents
C1	Fatal accidents	340
C2	Major injury accidents	658
C3	Minor/No injury accidents	472

Source: Compiled by the authors

Further, we applied ARM technique on these clusters. In order to use ARM, we applied the Wekatool. To generate strong association rules, different support values were used. We used support values of 10 %, 20 % and 30 %. The number of rules generated on different support values for each cluster are shown in Table 3.

Table 3. Rules generated for each cluster on different support values

Support value (%)	Number of rules generated		
	C1	C2	C3
10	15	20	24
20	5	14	18
30	1	3	2

Source: Compiled by the authors

Support values indicate in how many cases certain attribute values frequently occurred together. Hence, more support value extracts stronger rules. It can be seen from Table 3 that for 30 % support value the number of rules generated for each cluster is too small. Hence, we decided to fix the threshold value for support parameter to 15 %, so that at least enough number of rules could be generated that reveal some interesting facts about traffic accidents at different severity levels.

4.2 Association Rule Generation

The Wekatool is used to generate association rules for each of the severity levels. For this purpose, we used the Apriori algorithm to extract strong rules for each of the clusters. Some important and relevant rules for each of the severity levels of road accidents are mentioned in Table 4.

Table 4. Some relevant rules generated for different severity levels

Rule No.	Rule body	C_f	L_f	Severity level
	ROT=1 and SUA=2 → TIM=1	0.98	2.65	1
	ROT=1 and SUA=3 → TIM=4	0.84	2.22	1
	AGE=2 and LIG=3 → SUA=2	0.79	1.95	1
	AGE=1 and LIG=1 → SUA=1	0.95	1.98	3
	SUA=1 and ROF=1 → AGE=3	0.92	2.45	2
	LIG=3 and SUA=2 → TIM=4 and ROF=1	0.86	1.78	2
	SUA=4 and ROF=1 → AGE=2 and LIG=1	0.78	1.65	3
	LIG=2 and SUA=2 → TIM=4	0.73	1.88	2
	ROT=1 and ROF=2 → SUA=2	0.69	1.26	2
	AGE=1 and SUA=4 → TIM=2	0.71	1.65	2

Source: Compiled by the authors

The information revealed from the association rules for different severity levels is discussed below.

4.2.1 Association rules for C1

C1 represents all the traffic accidents that lead to fatality of victims injured in accidents. For this cluster, few rules were obtained, which reveal that highways at night in Muzzafarnagar district are mainly involved in fatal accidents. Also, most of the people that died in these accidents had 18 to 40 years of age. Industrial areas and agricultural land areas are found to be more dangerous in early 00:00 to 06:00 hours and late 18:00 to 00:00. Most of the accidents in these timings were reported fatal.

4.2.2 Association rules for C2

This cluster represents all traffic accidents in which a victim received major or severe injuries. Most of the rules in this cluster are similar to cluster C1. Some other factors were identified such as people with ages over 40 were mainly involved in major injury accidents. Also, it is found that intersections near residential areas are mainly involved in major injury accidents. Major injury accidents have occurred in all light conditions, but the frequency of accidents is high in road light condition. It is found that road accidents during 06:00 to 12:00 mostly involved victims of age under 18 years and the area associated with these accidents was either residential or commercial.

4.2.3 Association rules for C3

Cluster C3 represents the accidents that received minor/no injury. Although these type of accidents are not usually taken seriously by people. Sometimes, the involvements in such accidents overcome the fear of people towards major accidents because in these accidents the injury level is almost negligible. So it is important to see the factors responsible for these categories. The association rules revealed that these accidents are scattered all over the places where major or fatal accidents took place. But the commercial areas such as markets are the places where the frequency of the low injury accidents is higher compared to other locations. Association rules show that in market areas these accidents mostly occurred during day light conditions, whereas in residential areas it usually occurred in no light conditions or road light conditions.

We have therefore identified different road accident factors in the form of association rules that affect severity of road accidents. To achieve this, a framework was proposed that enabled us to use ARM for severity classification of road accident data.

5. Conclusions

In this paper, we have proposed a framework to use ARM as a classification technique for severity analysis of road accident data. The data were collected from Muzzafarnagar police station records. The framework enables us to find three clusters from the data set. Each cluster represents a type of accident severity (i.e. fatal, major injury and minor/no injury). The association rules exposed different factors that are associated with road accidents in each of the categories. The information extracted provides important information which can be utilized to adopt some preventive measures to overcome the accident severity in Muzzafarnagar district.

References

[1] S. Kumar & D. Toshniwal, "A novel framework to analyze road accident time series data", *Journal of Big Data*, vol. 3, no. 8, pp.1-11, 2016. doi:10.1186/s40537-016-0044-5

[2] World Health Organization (WHO), *Global status report on road safety 2013. Supporting a decade of action,* Luxembourg: WHO, 2013. Available: http://www.who.int/violence_injury_prevention/road_safety_status/2013/en/

[3] M. Karlaftis & A. Tarko, "Heterogeneity considerations in accident modeling", *Accid. Anal. Prev.*, vol. 30, no. 4, pp. 425-433, 1998.

[4] S. Kumar & D. Toshniwal, "Analysis of hourly road accident counts using hierarchical clustering and cophenetic correlation coefficient (CPCC)", *Journal of Big Data*, vol. 3, no. 13, pp. 1-11, 2016.

[5] P. N. Tan, M. Steinbach & V.Kumar, *Introduction to data mining*, Boston: Pearson Addison-Wesley, 2006, p. 769.

[6] S. Kumar & D. Toshniwal, "Analysing road accident data using association rule mining", in *International conference on computing communication and security (ICCCS-2015)*, Kanyakumari, India, Nov. 2-3, 2015.

[7] J. Han & M. Kamber, *Data mining: concepts and techniques*, United States: Morgan Kaufmann Publishers, 2001.

[8] P. J. Ossenbruggen et al.,"Roadway safety in rural and small urbanized areas", *Accidents Analysis and Prevention*, vol. 33, no. 4, pp. 485-498, 2001.

[9] L. Mussone et al., "An analysis of urban collisions using an artificial intelligence model", *Accident Analysis and Prevention,* vol. 31, pp. 705-718, 1999.

[10] L. Chang & W. Chen, "Data mining of tree based models to analyze freeway accident frequency", *Journal of Safety Research*, vol. 36, pp. 365- 375, 2005.

[11] J. D. Oña, G. López, R. Mujalli & F. J. Calvo, "Analysis of traffic accidents on rural highways using latent class clustering and bayesian networks", *Accid Anal Prev*, vol. 51, pp. 1-10, 2013.

[12] S. Kumar & D. Toshniwal, "A data mining framework to analyze road accident data", *Journal of Big Data*, vol. 2, no. 1, pp. 1-18, 2015. doi:10.1186/s40537-015-0035-y

[13] S. Kumar, D. Toshniwal & M. Parida, "A comparative analysis of heterogeneity in road accident data using data mining techniques", *Evolving Systems*. doi: 10.1007/s12530-016-9165-5

[14] K. Geurts, G. Wets, T. Brijs & K. Vanhoof, "Profiling of high frequency accident locations by use of association rules". *Transportation Research Record Journal of the Transportation Research Board*, vol. 1840, 2003. doi:10.3141/1840-14

[15] L. Thakali, T. Kwon & L. Fu, "Identification of crash hotspots using kernel density estimation and Kriging methods: a comparison", *J. Mod. Transp.*, vol. 23, no. 3, pp. 93-106, 2015.

[16] S. Kumar & D. Toshniwal, "A data mining approach to characterize road accident locations", *J. Mod. Transp.*, vol. 24, no. 1, pp. 62-72, 2016.

[17] J. Abellan, G. López & J. Ona,"Analysis of Traffic Accident Severity using Decision Rules via Decision Trees", *Expert System with Applications*, vol. 40, no. 15, pp. 6047-6054, 2013. doi:10.1016/j.eswa.2013.05.027

[18] T. B. Tesema, A. Abraham & C. Grosan, "Rule mining and classification of road accidents using adaptive regression trees", *Int J Simulation*, vol. 6, pp. 80-94, 2005.

[19] T. Kashani, A. S. Mohaymany & A. Rajbari, "A Data Mining Approach to Identify Key Factors of Traffic Injury Severity", *Promet-Traffic & Transportation*, vol. 23, pp. 11-17, 2011.

[20] B. Depaire, G. Wets & K. Vanhoof, "Traffic Accident Segmentation by means of Latent Class Clustering", *Accident Analysis and Prevention*, vol. 40, pp. 1257-1266, 2008.

[21] O. H. Kwon, W. Rhee & Y. Yoon, "Application of Classification Algorithms for Analysis of Road Safety Risk Factor Dependencies", *Accident Analysis and Prevention*, vol. 75, pp. 1-15, 2015.

[22] R. Agrawal & R. Srikant,"Fast Algorithms for Mining Association Rules in Large Databases", in *Proceedings of the 20th International Conference on Very Large Data Bases*, Santiago de Chile, Chile, Sept. 12-15, 1994, pp. 487-499.