Road Traffic Accidents Analysis Using Data Mining Techniques

G. Janani, N. Ramya Devi

Assistant Professor, Department of Information Technology, Sri Shakthi Institute of Engineering and Technology, Coimbatore

jananig@siet.ac.in, nramyadevi@siet.ac.in

Contribution to the state of the art

DOI: **10.7251/JIT1702084J** UDC: **656.1.08:656.08**

Abstract: Road Traffic Accidents (RTAs) are a major public concern, resulting in an estimated 1.2 million deaths and 50 million injuries worldwide each year. In the developing world, RTAs are among the leading cause of death and injury. Most of the analysis of road accident uses data mining techniques which provide productive results. The analysis of the accident locations can help in identifying certain road accident features that make a road accident to occur frequently in the locations. Association rule mining is one of the popular data mining techniques that identify the correlation in various attributes of road accident. Data analysis has the capability to identify different reasons behind road accidents. In the existing system, k-means algorithm is applied to group the accident locations into three clusters. Then the association rule mining is used to characterize the locations. Most state of the art traffic management and information systems focus on data analysis and very few have been done in the sense of classification. So, the proposed system uses classification technique to predict the severity of the accident which will bring out the factors behind road accidents that occurred and a predictive model is constructed using fuzzy logic to predict the location wise accident frequency.

Keywords: Road Traffic Accident (RTA), Data Mining, k-means, Association Rule Mining (ASM), Classification, Prediction.

Introduction

Road and traffic accidents are the major causes of fatality and disability in Coimbatore. A RTA not only causes property damage but it may lead to partial or full disability and sometimes can be fatal for human being. Increasing ratio of RTA is not a good sign for the transportation safety. The analysis of traffic accident data provides solution to identify different causes of road accidents and undertaking preventive measures. Various types of research have been done on road accident data from different countries. Significant help in this situation represents an identification of the key factors causing road traffic accidents [1]. Application of suitable data mining methods on the collected datasets representing different situations on the roads and occurred accidents can help understand the most significant factors or often repeating patterns. The success of such analysis depends strongly on the quality of the data available for the experiments. An interesting source of data in this domain is continually created by Department of Police, Coimbatore. All available data files provide detailed road safety data about the circumstances of personal injury road accidents involved and the consequential casualties. The statistics relate only to the accidents that are occurred on public roads, and are reported to the police and subsequently recorded using the RADMS (Road Accident Database Management System) accident reporting form (completed by police).

This paper consists of four main sections: the section 1 discusses about the Literature review. The proposed methodology is discussed in section 2. The section 3 describes the simulation results and

the section 4 describes the conclusion, that is summarizes extracted knowledge in respect with other relevant work.

References Survey

Sachin Kumar. [2] Discussed about the various data mining techniques in order to cluster the data into various categories and to identify the correlation between the attributes in the dataset. Lee et al. [3] stated that statistical models were a good choice to analyze road accidents in order to identify the correlation between accident and other traffic and geometric factors. However, Chen and Jovanis [4] stated that analyzing large dimensional datasets using traditional statistical techniques may result in certain problems such as sparse data in large contingency tables and also the statistical models have their own model with specific assumptions and violation of these can lead to some erroneous results. Due to these limitations of statistical methods, data mining techniques are being used to analyze road accidents. Data mining techniques are used to extract novel, implicit and hidden information from large data. Barai [5] discussed that there are variety of applications in transportation engineering such as road roughness analysis, pavement analysis and road accident analysis which uses data mining. Various data mining techniques [6] such as ASM, classification and clustering are widely used for the analysis of road accidents. Accident cases in India are usually recorded by police officer of the region in which the accident has occurred and also the area covered by a police station is limited and they keep record of accidents that are occurred in the regions under their control. Abellan et al. [7] developed various decision trees to extract different decision rules to analyze two-lane rural highway data of Spain. It is found that bad light conditions and safety barriers badly affect the crash severity. Geurts et al. [8] used ASM technique to analyze the various circumstances that occur at high-frequency accident locations on Belgium road networks. Tesema et al. [9] used adaptive regression tree model to build a decision support system for the road accidents in Ethiopia. Kashani et al. [10] used the Classification and Regression Tree (CART) to analyze road accidents data of Iran and found that not using seat belt, improper overtaking and over speed affect the severity of accidents. Kwon et al. [11] used classification algorithm to analyze factor dependencies related to road safety. Accident severity is directly concerned with the victim involved in accidents and it only targets the type of severity and shows the circumstances that affect the injury severity of accidents. Most of the accidents are concerned with certain location characteristics which make them to occur frequently at these locations. Hence, identification of these locations where accident frequencies are high and further analyzing them is very much beneficial to identify the factors that affect the accident frequency at these locations. Depaire et al. [12] discussed that cluster analysis of road accident data can extract better information rather than analyzing data without clustering. Preeti Mulay [13] discussed that the RTA involved fatal crashes data is directly concerned with nutritional health survey data to analysis of the association of dietary habit of a motor vehicle driver's to road traffic accident by applying Association rule mining algorithms. So the previous work focuses mainly on the driver characteristics and dietary habits, where aforementioned was analysed using ASM, and this paper focused on the contribution of various road-related factors such as the role of environment, place where the accident occurred and cause of the accident in order to classify the severity of the accident.

In this paper, the data mining techniques are used to identify accident locations which are more prone to risk and further analyzing them to identify various factors that affect road accidents at those locations. Initially, the dataset is divided into k groups based on their locations using k-means clustering algorithm. Then, the association rule mining algorithm is applied on those to reveal the correlation between different attributes in the accident data and understand the characteristics of these locations. Then, the Classification algorithm (Naive Bayes) is applied to classify the severity of the accident.

Methodology

The proposed methodology consists of four phases, namely the Preprocessing, Clustering of data, Association Rule Mining, and Classification. Figure 3.1 represents the system architecture of the project.

Data Preprocessing

Data preprocessing is the initial step in data mining techniques which involves mainly transforming the raw data into an understandable format. Generally Real-world data is incomplete, inconsistent and is likely to contain many errors. Data preprocessing is a method of resolving such issues and it prepares the raw data for further processing. In this paper the Data preprocessing techniques such as Data Cleaning and Data Transformation is used.

Figure 3.1. *System Architecture*

Clustering

Clustering is an unsupervised data mining technique which is used to group the data objects into different clusters in such way that objects within a group are more similar than the objects in other clusters. K-means algorithm [14] is very popular clustering technique for numerical data. It groups the data objects into k clusters. There are various clustering algorithms existing but selection of suitable clustering algorithm depends on the type and nature of data. Our prime motive of this paper is to discriminate data into different clusters based on the accident location.

Association Rule Mining

Association rule mining is a very popular data mining technique based on market basket analysis that extracts interesting rules between various attributes in a large data set [18]. 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 is TID. Let I = {I1, I2,..., In} be 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 which are responsible for an accident occurrence. In association rule mining, various interesting measures are there to assess the quality of a rule. These interesting measures for the rule $A \rightarrow B$ are discussed as follows:

Support

The support of the rule $A \rightarrow B$ defines the percentage how often A and B occur together in a data set and can be calculated using the 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.

$$\text{Support} = \frac{P(A \cap B)}{N} \quad (1)$$

Where N is the total number of accident records.

Confidence

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

$$\text{Confidence} = \frac{P(A \cap B)}{P(A)} \quad (2)$$

Lift

Lift for the rule $A \rightarrow B$ measures the occurrence of A and B together more than expected. 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 consequent 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).

$$\text{Lift} = \frac{P(A \cap B)}{P(A) * P(B)} \quad (3)$$

Apriori Algorithm

Apriori Algorithm [14] is used to generate the frequent item-sets and the strong association rules. The input of the algorithm will be the transaction Database of Accident data and the output will be the frequent item-sets and Association rules which satisfy the minimum threshold of Lift.

Classification

Classification is the process of finding a derived model which describes the data classes. The main purpose is to be able to use the model to predict the class of objects whose labels are unknown. The derived model is based on the analysis set of training data.

Naive Bayes Classifier

Naive Bayes classifier [14] uses the probabilistic method to predict a class for every instance of data set. The input of the algorithm is Test data and the output will be the predicted severity level. The specific working process of the Naive Bayes is as follows:

Let T be the training sample set. Each sample has category labels. Sample set has a total of m classes: C1, C2,...Cm. Each sample is represented by an n-dimensional vector System designX={x1, x2, ..., xn}, and each vector describes n attributes A1, A2,...,An. Different ways in calculating the probability of the class are explained below.

1. Given a simple X, the classifier will predict that X belongs to the highest posterior probability of class. If and only if $P(C_i|X)>P(C_j|X)$, $1<=i, j<=m$, X is predicted to belong to class C_i. According to the Bayes' theorem, the probability is calculated as in equation (4).

$$P(C\_i/X)=(P(X/C\_i)*P(C\_i))/(P(X) \quad (4)$$

Because P(X) is the same for all classes, it only need to find the largest $P(X|C_i)P(C_i)$. The prior probability of class C_i can be calculated. $P(C_i)=s_i/s$, s_i is the number of training samples of class C_i, and s is the total number of training samples. If the prior probability of class C_i is unknown, it is usually assumed that the probability of these classes are equal, then P(C1)= P(C2)= ...= P(Cm), therefore the problem is transformed into how to get maximum $P(X| C_i)$.

2. If the data set has many attributes, the workload of calculating $P(X| C_i)$ is very high. In order to reduce the computational overhead of $P(X|C_i)$, simple assumptions are used that under certain condition attribute characteristic value is independent of each other. $P(X| C_i)$ is calculated as in equation (5)

$$P\left(\frac{X}{C_i}\right) = \prod_{k=1}^{n} P\left(\frac{x_k}{C_i}\right) \quad (5)$$

3. Probability $P(x_1|C_i)$,$P(x_2|C_i)$, ..., $P(x_n|C_i)$ can be calculated from the training set. Here x_k refer to the attribute A_k of sample X.

4. For each class, calculating $P(X|C_i)P(C_i)$. If and only if $P(X|C_i)P(C_i)$ is maximum, the classifier prediction sample X belongs to class C_i. Bayes' theorem is used for classification as the past information about a parameter can be incorporated and form a prior distribution for future analysis.

Performance Evaluation

Classification performance is evaluated in terms of three commonly used metrics: accuracy, recall and precision as defined in equation (6) – (8). Table 3.1 is a confusion matrix whose entries are given as a function of two typical classes in severity classification.

- Accuracy is the percentage of test set samples that are correctly classified by the model.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP} \quad (6)$$

- Precision is the fraction of retrieved instances which are relevant.

$$\text{Precision} = \frac{\text{Number of TP}}{\text{Number of TP+FP}} \quad (7)$$

- Recall is the fraction of relevant instances which are retrieved.

$$\text{Recall} = \frac{\text{Number of TP}}{\text{Number of TP+FN}} \quad (8)$$

Where
TP= True Positive FP= False Positive
TN= True Negative FN= False Negative

Table 3.1. *Confusion Matrix*

	Predicted Slight	Predicted Severe
Actual Slight	TP	TN
Actual Severe	FP	FN

Simulation Results

The proposed methodology is implemented using the Java language and executed in NetBeans IDE.

Data Set

In Coimbatore, all accident related details are collected and maintained by the Department of Police using software called Road Accident Database Management System (RADMS) and the information is stored at the central server which is located at one particular place in TamilNadu. Hence, these data provide information about accidents that have happened in the road network of entire city. The data for this study is obtained from the Commissioner office, Coimbatore. The data consists of 570 accident details for 3 years from 2013 to 2015 in Coimbatore. After pre-processing, 542 accident records were considered for this study. A description about the data set is provided in the table 4.1. Sample dataset is provided in the table 4.2.

Table 4.1. *Data set Description*

Attribute Name	Type
Date & Time of Accident	Nominal
Road name	Nominal
Road No	Nominal
Municipality	Nominal
Fatal	Nominal
Grievous	Nominal
Injury	Nominal
Property Damage	Nominal
Nature of Accident	Nominal
Reason as in FIR	Nominal
Place	Nominal
Lightning	Binary

Table 4.2. *Sample Data set*

Road name	Road no	Municipality	fatal	Grievous	Nature
1	1	1	1	0	1
2	1	2	0	1	2
2	1	2	1	1	3
3	2	1	0	1	1

Categorization of accident locations

K-means clustering technique was applied on the accident data to get three clusters based on the accident locations. The clusters are renamed as the Area under city, Area beyond city limit and Area under Highways. There are 52 locations where accidents happened in Coimbatore.

Association rule mining

Apriori algorithm is used to generate the rules. To find the strong association rules minimum support of 5% is set. Association rules provide the correlation between the different attributes when an accident happens. Based on the lift value the interesting rules have been chosen in this paper. The rules for various clusters are discussed below:

Association rules for Cluster 1

The association rules of Cluster 1 shows that most of the accidents that happened in these locations are mainly due to over speed and careless driving. These locations are highly sensitive to Hit and Run. Most of the accidents happened here led to injuries & some led to property damage. If the nature of the accident is RTA then the area comes under the road type CH. Strong rules with high lift value show that the accidents are happening mostly near the junction areas and due to poor lightning and road surface.

Association Rules for Cluster 2

The association rules of Cluster 2 show that most of the accident happened in these locations are mainly due to self accident. Most of the accidents happened in these locations are due to negligence and some are due to the intersection road feature. The accidents that happened in these areas belong to the road type SH. When compared to Cluster 1, the fatal and injury levels are less in this cluster. Most of the vehicles involved in the accidents have crossed the minimum speed limit.

Association Rules for Cluster 3

The association rules of Cluster 3 show that most of the accidents happened in these locations are mainly due to rash driving. Most of the accidents' nature is RTA. Most fatal accidents happen due to rash driving and few self accidents also happened. The rules suggest that highways are more prone to accidents. When compared to the other areas, Cluster 3 areas are more prone to severe fatal accidents. Most of the accidents happened in the highways areas.

The association rules for the various clusters show the factors behind the accident and they reveal the correlation between different attributes. Some of the rules in all clusters are similar to each other. Similar rules such as: if the nature of the accident is RTA and the FIR is Rash, and few other rules are also similar, we come to the conclusion that rash driving leads to fatalities and injuries. If the road lightening is poor, then accident is likely to occur in those locations.

Classification

Naïve Bayes algorithm is used to classify the severity of accidents. The severity of an accident is directly concerned with the victims involved in the accident. Based on the affected victims, the severity level of an accident is classified. To train, the Model 70% of data is taken and to test the model 30% of data is used. Based on the attribute such as Fatal, Grievous, Injury and Damage the class label is created. The class label represents the severity level of the accident happened. Class 0 represents the low severity and class 1 represents the high severity. Naïve Bayes performs well in terms of accuracy when compared with other classification algorithms such as Decision tree J48, Random forest. The outcome of this phase is the severity level of the accidents occurred. In Coimbatore, 40% of accident happened belong to the severity level high and 60% of accident happened belong to the severity level slight. To measure the performance of the classifier, the classification accuracy is computed from the test set.

Prediction

In the proposed system, the Prediction model using fuzzy logic is built in order to predict the probability of accident occurrence in Coimbatore. Fuzzy rule based systems are an extension of classical rule based systems. Fuzzy rules are linguistic IF-THEN constructions that have the general form "IF A, THEN B" where A and B are propositions containing linguistic variables. In effect, the use of linguistic variables and fuzzy IF-THEN rules exploits the tolerance for imprecision and uncertainty. In this respect, fuzzy logic mimics the crucial ability of the human mind to summarize data and focus on decision-relevant information.

A fuzzy rule based system consists of four major modules: fuzzification, inference engine, knowledge base and defuzzification module [18]. The fuzzification module transforms the crisp input(s) into fuzzy values. These values are then processed in fuzzy domain by inference engine based on the knowledge base supplied by the domain expert(s). The knowledge base is composed of the Rule Base (RB), which characterizes the control goals and control policy of the domain expert by a set of linguistic control rules, and of the Data Base (DB), containing the term sets and the membership functions defining their semantics. Finally, the processed output is transformed from fuzzy domain to crisp domain by defuzzification module.

The structure of a rule base can be stated as follows:

R_i : if X_i is A_{i1} ... X_n is Ain then Y is B_j

Where A_{in} and B_j are fuzzy sets defined on the input and output domains respectively. $X_1 ... X_n$ and Y are input and output linguistic variables, respectively, and A_{i1} ... A_{in} and B_j linguistic labels, each one of them having associated a fuzzy set defining its meaning.

Figure 4.1 represents the yearly distribution of accidents happened in Coimbatore. In the year 2013, 207 accidents occurred, in the year 2014, 229 accidents occurred and in the year 2015, 218 accidents occurred.

Figure 4.2 represents the monthly distribution of accidents happened in the various clusters.

Figure 4.3 represents the rate of accidents that occurred in the various locations of Coimbatore.

Figure 4.4, Figure 4.5 and Figure 4.6 represent the comparison of performance metrics of different classification algorithms.

Figure 4.7 and Figure 4.8 represent the comparison of the Prediction results and Location wise prediction results respectively.

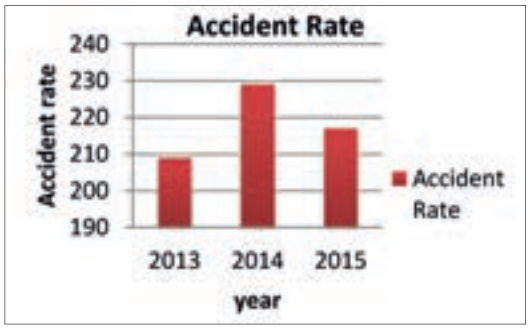

Figure 4.1. *Yearly Distribution of Accident rate*

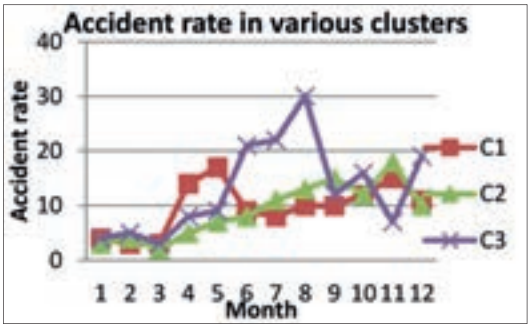

Figure 4.2. *Month wise Accident rate in various clusters*

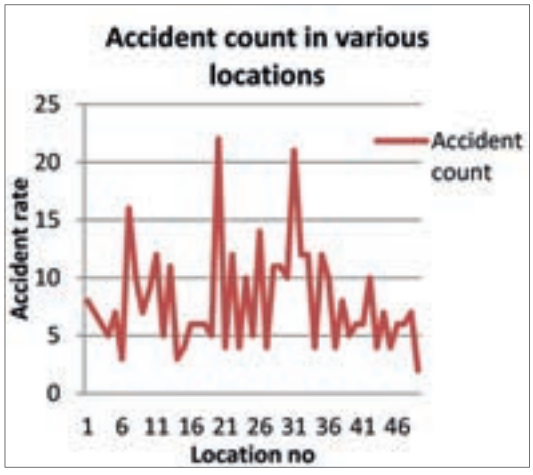

Figure 4.3. *Location wise accident rate*

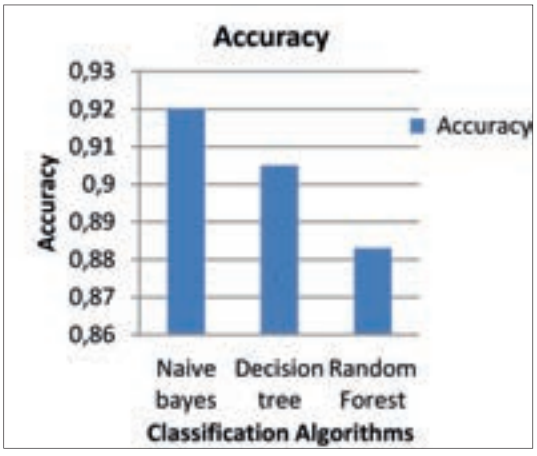

Figure 4.4. *Performance Evaluation of Classifier in terms of accuracy*

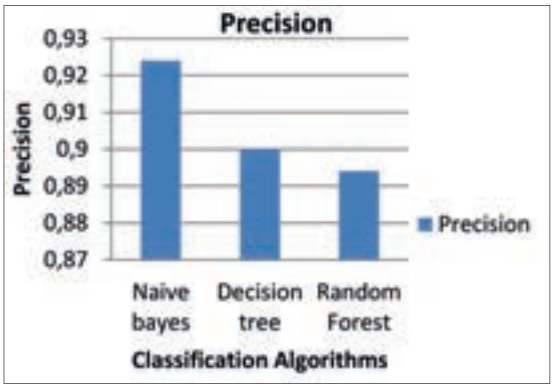

Figure 4.5. *Performance Evaluation of Classifier in terms of Precision*

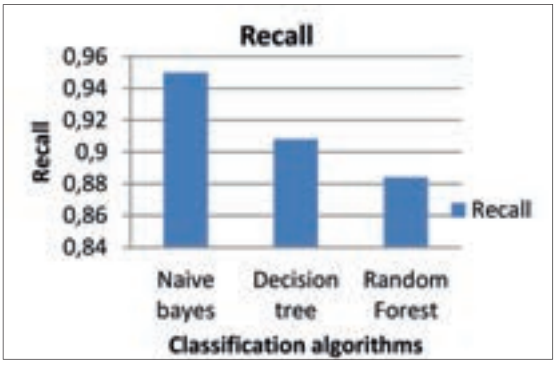

Figure 4.6. *Performance Evaluation of Classifier in terms of Recall*

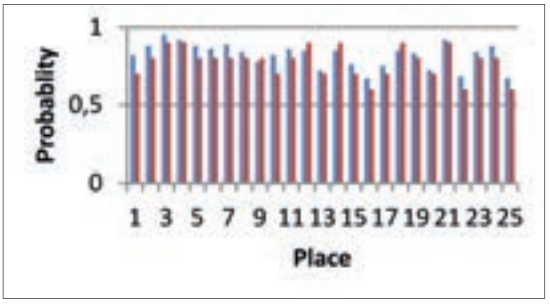

Figure 4.7. *Comparison of prediction results*

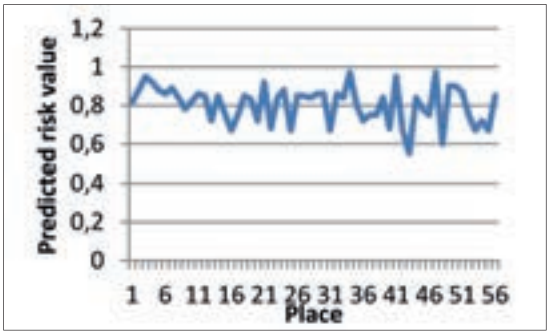

Figure 4.8. *Location wise prediction results*

Conclusion And Future Work

In this paper, traffic accident data of Coimbatore is collected and cleaned in order to use it to test the predictive model. The endeavour of this paper is to spot the factors behind an accident and severity

of accidents. The assessment of the Classification model showed that Naive Bayes algorithm outperforms with an accuracy of 92.45 % when compared with other algorithms. In contrast with the previously published work of authors, which focused on driver characteristics and dietary habits, this paper focused on the contribution of various road-related factors such as the role of environment, place where the accident occurred and cause of the accident that have impact on the accident severity. The results of this study could be used by the respective authorities to promote road safety and create awareness about risk factors. Thus, this work could have tremendous impact on the well-being of Coimbatore civilians and a predictive model is constructed in order to predict the probability of accident occurrence which helps the Coimbatore civilians to have awareness about the accident prone zones in advance.

References:

[1] Abellan J, López G, and De Oña J, "Analysis of traffic accident severity using Decision Rules via Decision Trees", Expert Systems with Applications, 40, 6047–6054, 2013.

[2] Addi, Ait-Mlouk et al. "An approach based on association rules mining to improve road safety in Morocco", International Conference on Information Technology for Organizations Development (IT4OD), 2016.

[3] Barai S, "Data mining application in transportation engineering". Transport 18:216–223, 2003.

[4] Beshah, Tibebe and Shawndra Hill. "Mining Road Traffic Accident Data to Improve Safety: Role of Road-Related Factors on Accident Severity in Ethiopia", AAAI Spring Symposium: Artificial Intelligence for Development, 2010.

[5] Chen W, Jovanis P, "Method for identifying factors contributing to driver-injury severity in traffic crashes". Transp Res Rec, 2000.

[6] Depaire B, Wets G, Vanhoof K, "Traffic accident segmentation by means of latent class clustering", Accid Anal Prev 40:1257–1266, 2008.

[7] František Babi, Karin Zuskáová, "Descriptive and Predictive Mining on Road Accidents Data", IEEE 14th International Symposium on Applied Machine Intelligence and Informatics, January 21-23, 2016.

[8] Geurts K, Wets G, Brijs T, Vanhoof K, "Profiling of high frequency accident locations by use of association rules". Transp Res Rec, 2003.

[9] Han J, Kamber M "Data mining: concepts and techniques", Morgan Kaufmann Publishers, Burlington, 2001.

[10] Kashani T, Mohaymany AS, Rajbari A, "A data mining approach to identify key factors of traffic injury severity". Promet-Traffic Transp 23:11–17, 2011.

[11] Kwon OH, Rhee W, Yoon Y, "Application of classification algorithms for analysis of road safety risk factor dependencies", Accid Anal Prev 75:1–15, 2015.

[12] Lee C, Saccomanno F, Hellinga B, "Analysis of crash precursors on instrumented freeways". Transp Res Rec, 2002.

[13] Matsatsinis N, "A fuzzy decision aiding method for the assessment of corporate bankruptcy," Fuzzy economic review, vol. 8, 2003.

[14] Preeti Mulay and Selam Mulat, "What You Eat Matters Road Safety: A Data Mining Approach", Indian Journal of Science and Technology, Vol 9(15), 2016.

[15] Sachin Kumar, Durga Toshniwal, "A data mining approach to characterize road accident locations" Springer Journal Vol. 24(1):62-72, 2016.

[16] Tan PN, Steinbach M, Kumar V "Introduction to data mining". Pearson Addison-Wesley, Boston, 2006.

[17] Tesema TB, Abraham A, Grosan C, "Rule mining and classification of road accidents using adaptive regression trees". Int J Simul 6:80–94, 2005.

Submitted: September 10, 2017.
Accepted: November 30, 2017.

ABOUT THE AUTHORS

Janani G holds a M. Tech in IT by the Anna University and is a assistant professor for the Department of IT. Her main area of interest is the study of data mining and analytics. She has presented papers at conferences and published papers in various journals. She has taught Grid and Cloud Computing and Problem solving and Python Programming.

Ramya Devi N holds a M. Tech in IT by the Anna University and is a assistant professor for the Department of IT. Her main area of interest is the study of data mining and networking. She has presented papers at conferences and published papers in various journals. She has taught Web Programming and ComputerArchitecture.