J. Mod. Transport. (2016) 24(1):62–72
DOI 10.1007/s40534-016-0095-5

www.springer.com/40534

A data mining approach to characterize road accident locations

Sachin Kumar[1] · **Durga Toshniwal**[2]

Received: 20 August 2015 / Revised: 5 January 2016 / Accepted: 6 January 2016 / Published online: 11 February 2016

Abstract Data mining has been proven as a reliable technique to analyze road accidents and provide productive results. Most of the road accident data analysis use data mining techniques, focusing on identifying factors that affect the severity of an accident. However, any damage resulting from road accidents is always unacceptable in terms of health, property damage and other economic factors. Sometimes, it is found that road accident occurrences are more frequent at certain specific locations. The analysis of these locations can help in identifying certain road accident features that make a road accident to occur frequently in these locations. Association rule mining is one of the popular data mining techniques that identify the correlation in various attributes of road accident. In this paper, we first applied *k*-means algorithm to group the accident locations into three categories, high-frequency, moderate-frequency and low-frequency accident locations. *k*-means algorithm takes accident frequency count as a parameter to cluster the locations. Then we used association rule mining to characterize these locations. The rules revealed different factors associated with road accidents at different locations with varying accident frequencies. The association rules for high-frequency accident location disclosed that intersections on highways are more dangerous for every type of accidents. High-frequency accident locations mostly involved two-wheeler accidents at hilly regions. In moderate-frequency accident locations, colonies near local roads and intersection on highway roads are found dangerous for pedestrian hit accidents. Low-frequency accident locations are scattered throughout the district and the most of the accidents at these locations were not critical. Although the data set was limited to some selected attributes, our approach extracted some useful hidden information from the data which can be utilized to take some preventive efforts in these locations.

Keywords Road accidents · Accident analysis · Data mining · *k*-Means · Association rule mining

1 Introduction

Road and traffic accidents (RTA) are one of the important problems in India. MORTH [1] mentioned in its report that every year there are 0.4 million accidents reported in India, which makes India a country with large accident rate. This report shows that there is a negative trend of accidents from 2012 to 2013; however, as accidents are unpredictable and can occur in any type of situation, there is no guarantee that this trend will sustain in future also. Therefore, the identification of different geographical locations where most of the accidents have occurred and determining the various characteristics related to road accidents at these locations will help to understand the different circumstances of accident occurrence. Kannov and Janson [2] stated that systematic relationship between accident frequency and other variables such as geometry of road, road side

✉ Sachin Kumar
sachinagnihotri16@gmail.com

Durga Toshniwal
durgatoshniwal@gmail.com

[1] Centre for Transportation Systems (CTRANS), Indian Institute of Technology Roorkee, Roorkee, Uttarakhand 247667, India

[2] Computer Science & Engineering Department, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand 247667, India

features, traffic information and vehicle information can help to develop effective accident prevention measures.

Lee et al. [3] indicated that statistical models were a good choice in the past to analyze road accidents to identify the correlation between accident and other traffic and geometric factors. However, Chen and Jovanis [4] determined that analyzing large dimensional datasets using traditional statistical techniques may result in certain problems such as sparse data in large contingency tables. Also, statistical models have their own model specific assumptions and violation of these can lead to some erroneous results. Due to these limitations of statistical techniques, data mining techniques are being used to analyze road accidents. Data mining is a set of techniques to extract novel, implicit and hidden information from large data. Barai [5] discussed that there are various applications of data mining in transportation engineering such as road roughness analysis, pavement analysis and road accident analysis. Various data mining techniques [6] such as association rule mining, 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. Also, the area covered by a police station is limited and they keep record of accidents that have occurred in their regions only. Ponnaluri [7] discussed that the report prepared by police only contains the basic information that are not much useful for the research purpose. He suggests that data collection method used by police needs a lot of improvement. However, Indian researchers used these data and analyzed it for some highway portions using statistical methods [8, 9]. Data mining can be described as a novel technique to extract hidden and previously unknown information from the large amount of data. Several data mining techniques such as clustering, classification and association rule mining are widely used in the road accident analysis by researchers of other countries. Geurts et al. [10] used association rule mining technique to understand the various circumstances that occur at high-frequency accident locations on Belgium road networks. Tesema et al. [11] used adaptive regression tree model to build a decision support system for the road accidents in Ethiopia. Abellan et al. [12] developed various decision trees to extract different decision rules for different trees to analyze two-lane rural highway data of Spain. They found that bad light conditions and safety barriers badly affect the crash severity. Depaire et al. [13] used clustering technique to analyze road accident data of Belgium and suggest that cluster-based analysis of road accident data can extract better information rather analyzing data without clustering. Kashani et al. [14] used classification and regression tree (CART) to analyze road accidents data of Iran and found that not using seat belt, improper overtaking and speeding badly affect the severity of accidents. Kwon et al. [15] used Naïve Bayes and decision tree classification algorithm to analyze factor dependencies related to road safety. Severity of accident is directly concerned with the victim involved in accidents, and its analysis only targets the type of severity and shows the circumstances that affects the injury severity of accidents. Sometime accidents are also concerned with certain locations characteristics, which makes 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.

In this paper, we are making use of data mining techniques to identify high-frequency accident locations and further analyzing them to identify various factors that affect road accidents at those locations. We first divide the accident locations into k groups based on their accident frequency counts using k-means clustering algorithm. Then association rule mining algorithm is applied on these to reveal the correlation between different attributes in the accident data and understand the characteristics of these locations. Hence, our main emphasis will be the interpretation of the outcomes.

2 Methodology

2.1 k-Means clustering

Clustering is an unsupervised data mining technique whose main task is to group the data objects into different clusters such that objects within a group are more similar than the objects in other clusters. k-means algorithm [16] 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 type and nature of data. Our prime motive of this paper is to discriminate the accident location based on their frequency count. We have two choices to do this: First, we can decide a threshold level for each of the category of accident locations and group them in some categories. The problem with this approach is that it is very difficult to identify the number of categories of accident locations and decide a threshold level for each category. The other way is to use k-means algorithm which can divide the accident locations into different groups. The number of groups can be identified using some cluster selection criteria such as gap statistic.

Initially, we have frequency counts of 87 locations with 7,027 road accidents. In order to divide the location into

different groups we used k-means algorithm. A brief formal description of k-means algorithm is given below [18]:

Input: $\{D, k\}$ // D→ Data set consists of n data objects, k→ Number of clusters
Output: k clusters
Method:

1. Choose the k objects at random from D, as initial cluster centers.
2. **repeat**
3. Assign each data object to the cluster to which its distance is closest
4. Update the cluster means, i.e., calculate the mean value of the objects for each cluster.
5. **Until** no data object changes its cluster membership or any other convergence criteria is met.

2.2 Number of cluster selection

The major problem with a clustering algorithm is to identify the number of clusters to be made. The weakness of k-means clustering is that the user has to provide the value for k. An inappropriate value for k may lead to wrong clustering results. In this paper, we have used the gap statistics [17] in order to find the value of k that can be supplied to divide the accident locations into different groups based on their frequency counts. Gap statistics can be used with any type of clustering technique, but they have been scarcely used to determine the number of clusters in road accident analysis.

Consider a data set D_{ij}, $i = 1,2,\ldots,m$, $j = 1,2,\ldots,n$, consisting of m data objects with values of n attributes. Assuming d_{xy} is the squared Euclidean distance between objects X and Y given by $d_{xy} = \sum (X_j - Y_j)^2$. If the data set has been clustered into k clusters, c_1, $c_2,\ldots,c_k$, where c_i indicates the ith cluster, then $n_i = |c_i|$.

Let $D_i = \sum d_{xy}$, (where $x, y \in c_i$) is the sum of pair-wise distances for all points in cluster i and W_k is the collective within cluster sum of squares around the cluster means and is given by Eq. (1). $\text{Gap}_n(k)$ can be defined as the difference between expected and observed values of $\log(W_k)$ and given in Eq. (2). K can be taken for the value maximizing $\text{Gap}_n(k)$.

$$W_k = \sum_{i=1}^{k} \left(\frac{1}{2n_i}\right) D_i, \tag{1}$$

$$\text{Gap}_n(K) = E_n^*\{\log(W_k)\} - \log(W_k), \tag{2}$$

where E_n^* denotes the expectation under a sample size n from the reference distribution.

2.3 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 $T \in D$. Let $I = \{I_1, I_2,\ldots, I_n\}$ 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 a rule $A \rightarrow B$ are discussed as follows:

2.3.1 *Support (S_p)*

The support of a rule $A \rightarrow B$ defines the percentage how often A and B occur together in a data set and can be calculated using Eq. (3). 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.

2.3.2 *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 Eq. (4). Higher the confidence values of a rule $A \rightarrow B$, 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.

2.3.3 *Lift (L_t)*

Lift for a rule $A \rightarrow B$ measures the occurrence of A and B together 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 Eq. (5).

2.3.4 *Leverage (L_v)*

Leverage for a rule $A \rightarrow B$ measures the difference of A and B appearing together in the data set and the expectation if A and B are statistically dependent [19]. It can be calculated using Eq. (6). The values for leverage range

from [−0.25 to +0.25]. A leverage value 0 indicates that the variables are statistically independent. It will increase towards +1 if the variables occur more often together and will decrease towards −1 if one of the variables alone occurs more often.

2.3.5 Conviction (C_v)

Conviction is another measure that undertakes some of the weaknesses of confidence and lift [20]. Conviction of a rule $A \rightarrow B$ compares the probability that A occurs without B if they are dependent with the actual frequency of the appearance of A without B. Conviction is not symmetric i.e. conviction $(A \rightarrow B) \neq$ conviction $(B \rightarrow A)$. Conviction is rather inspired in the logical definition of implication and attempts to calculate the degree of implication of any rule. The value for conviction ranges within $[0.5, \infty]$. The values which are distant from 1 indicate interesting rules. In conviction, the supports of both antecedent and consequent are taken into account. It can be calculated using Eq. (7):

$$S_p = \frac{P(A \cap B)}{N}, \tag{3}$$

where N is the total number of accident records.

$$C_f = \frac{P(A \cap B)}{P(A)}, \tag{4}$$

$$L_t = \frac{P(A \cap B)}{P(A) \times P(B)}, \tag{5}$$

$$L_v = P(A \cap B) - P(A) \times P(B), \tag{6}$$

$$C_v = \frac{P(A) \times P(B)}{P(A \cap \bar{B})}, \tag{7}$$

For a better understanding of the above concept, consider the following short example in Table 1, from road accident domain in which the set of items is $I = \{$Fog, High Traffic, Speed > 100, Low Traffic, Fatal Accident$\}$. In Table 1, 1 shows the presence of the item and 0 indicates the absence of the item.

An example rule {Fog, Speed > 100} → {Fatal Accident} specifies that if Fog and Speed > 100 occur together, Fatal Accident will also take place. To select the interesting rules, minimum threshold value of support is provided.

In the above example, the support and confidence value for the item set {Fog, Speed > 100, Fatal Accident} can be computed using Eqs. (3) and (4) as follows:

Table 1 Example data set

Accident id	Fog	Speed > 100	High traffic	Low traffic	Fatal accident
1	1	1	0	1	1
2	1	1	1	0	0
3	0	0	1	0	0
4	0	0	1	0	0
5	1	1	1	0	1

$$S_{\mathrm{p}} = P(\mathrm{Fog} \cap \mathrm{Speed} > 100 \cap \mathrm{Fatal\ Accident}) / \mathrm{Total\ Number\ of\ Accidents} = 2/5 = 0.4.$$

A support value of 0.4 indicates that in 40 % of accident records, Fog, Speed > 80 and Fatal Accident take place together.

$$C_f = \frac{P(\mathrm{Fog} \cap \mathrm{Speed} > 100 \cap \mathrm{Fatal\ Accident})}{P(\mathrm{Fog} \cap\ > 100)} = \frac{(2/5)}{(3/5)} = 0.66.$$

The confidence value of 0.66 means that in 66 % of accident cases when Fog and Speed > 100 occur together, then Fatal Accident also occurs. This value indicates that there are 66 % chances of fatal accident if the weather is foggy and vehicle speed is greater than 100.

The lift of the above rule can be calculated using Eq. 5 as follows:

$$L_f(\{\mathrm{Fog, Speed} > 100\} \rightarrow \{\mathrm{Fatal\ Accident}\}) = \frac{P(\mathrm{Fog} \cap \mathrm{Speed} > 100 \cap \mathrm{Fatal\ Accident})}{P(\mathrm{Fog}) \times P(\mathrm{Speed} > 100) \times P(\mathrm{Fatal\ Accident})} = \frac{0.4}{0.144} = 2.77,$$

The above lift value indicates that fog, speed more than 100 and fatal accidents are strongly correlated with each other and this rule can be used to predict future accident with good accuracy.

The leverage for the above rule can be calculated using Eq. (6) as follows:

$$\mathrm{Leverage}\ (\{\mathrm{Fog, Speed} > 100\} \rightarrow \{\mathrm{Fatal\ Accident}\}) = P(\mathrm{Fog} \cap \mathrm{Speed} > 100 \cap \mathrm{Fatal\ Accident}) - P(\mathrm{Fog}) \times P(\mathrm{Speed} > 100) \times P(\mathrm{Fatal\ Accident}) = 0.4 - 0.144 = 0.256.$$

The conviction for the above rule can be calculated using Eq. (7) as follows:

$$C_v(\{\text{Fog, Speed} > 100\} \rightarrow \{\text{Fatal Accident}\}) = \frac{P(\text{Fog} \cap \text{Speed} > 100) \times P(\text{Fatal Accident})}{P(\text{Fog} \cap \text{Speed} > 100 \cap \text{Fatal Accident})} = \frac{4}{1} = 4.0.$$

2.4 Data set

In India, 108 is an emergency ambulance service which provides help to the accident victims. This emergency service is being operated in several states of India. Uttarakhand is an Indian state where this service is running. This service is operated and handled by Emergency Management Research Institute (GVK-EMRI) [21], which provides emergency service to the accident victims and also keeps track of record of every accident served. Also, this information is stored at their central server located at one particular place in the state. Hence, these data provide information about accidents that have happened in the road network of entire city, district and state. The data for this study have been obtained from GVK-EMRI, Dehradun. The data set consists of 15,574 road accidents for 6 years period from 2009 to 2014, in Dehradun District of Uttarakhand State. After preprocessing, 9,640 accident records have been considered for this research.

The attributes of the data are mostly categorical in nature. A brief description of the data is given in Table 2 as follows:

3 Results and discussion

3.1 Categorization of accident location

In order to categorize accident locations on the basis of their accident frequency count. We extracted the frequency count of accident locations from the data. In total there were around 158 locations found with 9,640 road accidents. In order to provide better analysis, we excluded those locations which have frequency count of less than 20 over 6 years of duration, which results in 87 locations with 7,327 road accidents. These locations are plotted with location id in the increasing number of frequency, as shown in Fig. 1.

Number of clusters to be made is identified using gap statistic as discussed in Sect. 2.3. The values obtained for W_k, log (W_k), E_n^* (log (W_k)) and gap statistic are plotted against the number of clusters in Fig. 2a, b, c respectively. The graphs clearly indicate that there are three clusters for the accident locations. Figure 2a shows a knee curve at cluster 3, Fig. 2b plots the observed and expected values for log(W_k) and Fig. 2c shows that gap value at cluster 3 that maximizes the Gap$_n(k)$.

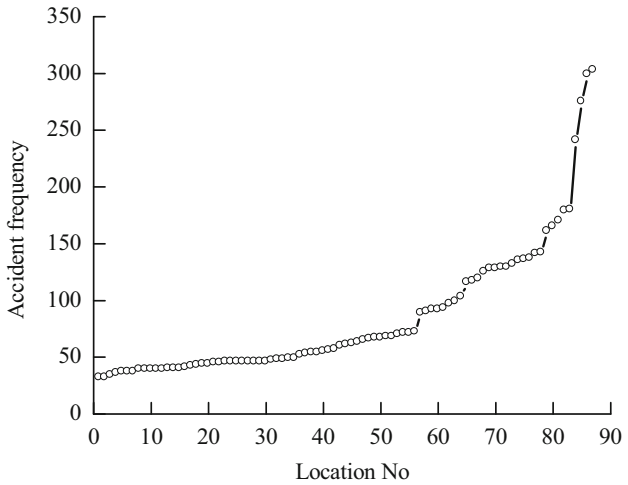

Fig. 1 Accident frequency count of locations in ascending order

Table 2 Road accident attributes used for analysis

Attribute name	Type	Values
Number of victims	Nominal	1, 2, >2
Victim age group	Nominal	Children, Young, Adult, Senior
Victim gender	Binary	Male/Female
Accident category	Nominal	2 wheeler, 3 wheeler, Vehicle_Fall_height, Pedestrian_hit, Multi_Vehicular_Incident (MVI), Vehicle_rollover/skid
Time of accident	Nominal	1,…,12
Day	Nominal	Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday
Month	Nominal	1, 2….,12
Location_No	Ratio	1,……,158
Lighting on road	Binary	Daylight, street light, no light
Roadway feature	Nominal	intersection, curve, slope, other
Accident severity	Nominal	Critical/non-critical
Area around	Nominal	Hospital, colony, market, forest, hills, agriculture land
Road type	Nominal	Highway/non-highway

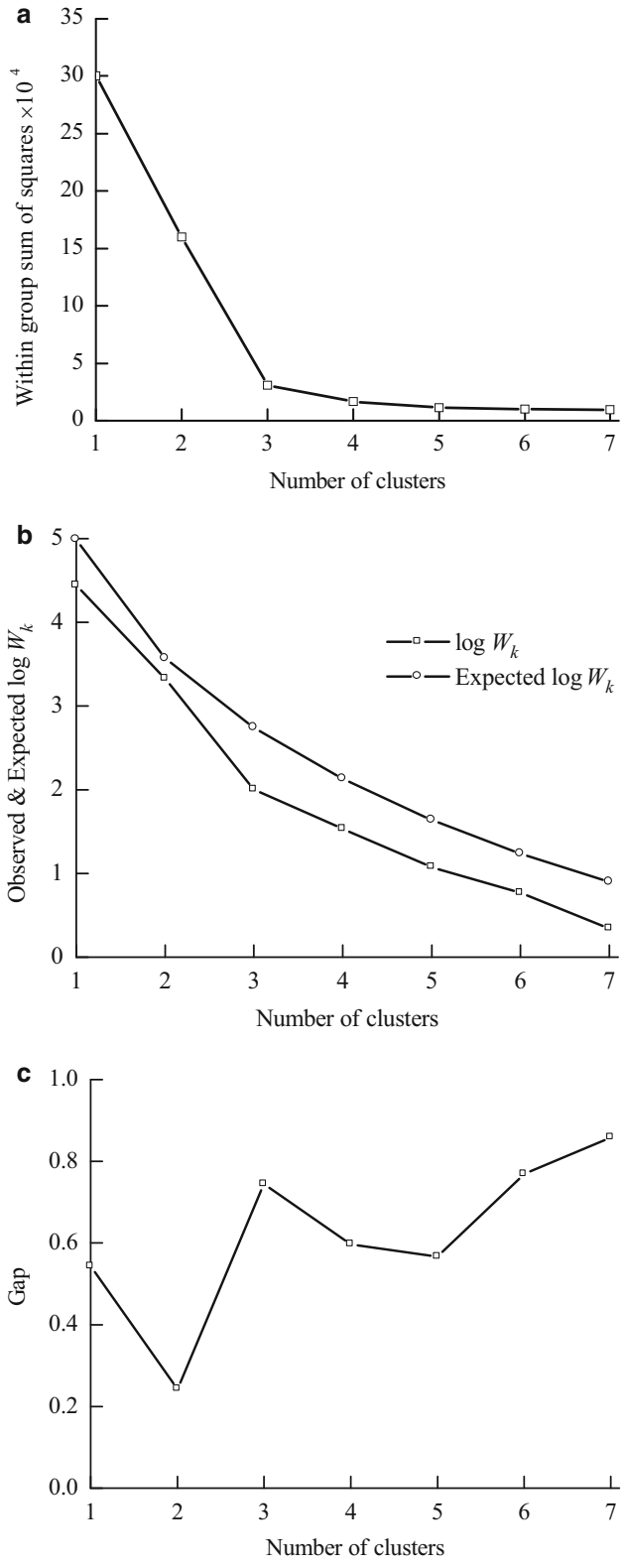

Fig. 2 **a** Within group sum of squares function W_k against the number of clusters. **b** Observed log (W_k) and Expected $E_n^*\{\log(W_k)\}$. **c** Gap statistic curve

Further, *k*-means clustering technique was applied on the accident frequency counts of accident locations to get three clusters of 87 locations as given in Table 3. We renamed three clusters as high-frequency accident location (HFAL), moderate-frequency accident locations (MFAL) and low-frequency accident locations (LFAL) based on their frequency count.

In LFAL cluster, there are 56 locations with a frequency count of less than 75, where a total of 2,646 (36.11 %) accidents occurred. Similarly, 22 locations with accident count in the range of 75–150 are in MFAL cluster which consists of 35.06 % of accidents. There are eight locations in HFAL cluster where around 28.82 % of the total accidents have occurred.

3.2 Mining association rules

Initially, we divided the locations into three categories namely HFAL, MFAL and LFAL. We used Apriori [22] algorithm in Weka 3.6.12 in order to generate rules. In order to identify strong rules, a minimum support of 5 % has been taken. The following tables show the association rules generated for each category.

Tables 4, 5 and 6 show the association rule generated for HFAL, MFAL and LFAL, respectively. Association rules provide an understanding of correlation between different attribute values that occur together when an accident occurs. Several rules are generated for each category but only some interesting rules (based on lift value) have been chosen to show in this paper. The rules for HFAL, MFAL and LFAL are discussed as follows:

3.3 Association rules for HFAL

The association rules reveal that HFAL locations are primarily hill roads towards hill stations with continuous traffic flow. Most of the accidents at these locations are severe accidents. These locations are highly sensitive for two-wheeler accidents. Other type of accidents such as pedestrian hit, multi-vehicular accidents are very less as compared to two-wheeler accidents. The major accidents have happened at curve and slope on hilly roads. The other locations with high-frequency accidents are markets with high traffic volume. Strong rules with high lift value disclose that intersections at market on highway roads are the main locations where accidents have occurred. The rules suggest that highways are the high-frequency accident locations, and intersections on highways which come across market locations are more prone to severe accidents.

Table 3 Accident frequency distribution and categorization

Cluster id	Number of locations	Accident frequency	Total accidents	% Accidents	Category name
1	56	30–75	2,646	36.11	LFAL
2	22	75–150	2,569	35.06	MFAL
3	8	150–303	2,112	28.82	HFAL

Table 4 Association rules for HFAL

Rule no.	Rule body	Conf.	Lift	Lev.	Conv.
1.	{Hills} → {Curve, Highway}	0.56	5.33	0.05	2.03
2.	{Curve, Highway} → {Hills}	0.6	5.33	0.05	2.03
3.	{Time = 12, Agriculture Land} → {Others, No light}	0.9	4.27	0.05	7.34
4.	{Intersection, Highway} → {Daylight, Market}	0.48	3.07	0.05	1.6
5.	{Market, 2 wheeler} → {Intersection, Highway}	0.42	2.71	0.04	1.45
6.	{Intersection, Highway} → {Market}	0.72	2.61	0.07	2.53
7.	{Others, Agriculture Land} → {No light, 2 wheeler}	0.39	2.07	0.06	1.32
8.	{No light, Non-highway} → {Agriculture Land}	0.97	1.98	0.06	13.35
9.	{Pedestrian, Market} → {Intersection}	0.56	1.81	0.02	1.54
10.	{Intersection} → {Non-Highway, 2 wheeler}	0.39	1.46	0.04	1.2
11.	{Others, Vechicle_Rollover} → {Agriculture Land}	0.71	1.45	0.02	1.7
12.	{Intersection, Market, 2 wheeler} → {Highway}	0.81	1.44	0.02	2.19
13.	{Intersection, Market} → {Highway}	0.79	1.42	0.03	2.1
14.	{Daylight, Highway} → {Multi_Vehicular_Incident}	0.56	1.18	0.01	1.03
15.	{Highway, Agriculture Land → 2 wheeler}	0.45	1.07	0.01	1.02

Table 5 Association rules for MFAL

Rule no.	Rule body	Conf.	Lift	Lev.	Conv.
16.	{Colony} → {Non-highway, Pedestrian}	0.48	3.22	0.01	1.14
17.	{Intersection, Market} → {Highway, Pedestrian}	0.44	3.11	0.01	1.11
18.	{Curve, Agriculture Land} → {Multi_Vehicular_Incident}	0.36	2.31	0.01	1.29
19.	{Nolight, Non-highway → {Agriculture Land, Multi_Vehicular_Incident}	0.71	2.11	0.01	1.1
20.	{Role_over} → {Other, Agriculture Land}	0.64	2.08	0.03	1.89
21.	{Daylight, Highway} → {Curve, Fall_height}	0.47	1.93	0.01	1.03
22.	{Daylight, 3 wheeler} → {Other}	0.97	1.87	0.01	9.09
23.	{Street Light, 2 wheeler} → {Intersection}	0.55	1.8	0.02	1.52
24.	{Agriculture Land, 2 wheeler} → {Other, No light, Non-highway}	0.68	1.78	0.02	1.52
25.	{Intersection, Non-highway} → {Daylight, Agriculture Land}	0.42	1.77	0.03	1.3
26.	{Intersection, Daylight, 2 wheeler} → {Agriculture Land, Non-highway}	0.46	1.7	0.02	1.34
27.	{Daylight, 2 wheeler} → {Time = 9}	0.41	1.6	0.02	1.1
28.	{2 wheeler} → {Intersection, Agriculture Land}	0.49	1.55	0.04	0.49
29.	{Market} → {Daylight, Highway}	0.45	1.46	0.04	1.25
30.	{Pedestrian, Colony} → {Intersection}	0.58	1.43	0.02	1.54

3.4 Association rules for MFAL

In MFAL locations, colonies on non-highway roads and intersections in market on highway roads are dangerous for pedestrians. Curve or bends on roads surrounded with agriculture land on non-highway roads are dangerous for multi-vehicle accidents. Roads that pass through agriculture land are also found dangerous to vehicle role over

Table 6 Association rules for LFAL

Rule no.	Rule body	Conf.	Lift	Lev.	Conv.
31.	{Intersection, Market} → {Street Light}	0.8	5.53	0.03	4.06
32.	{Curve, Highway} → {Hills, 2 wheeler}	0.26	4.47	0.02	1.26
33.	{Month = 10, Highway} → {Market}	0.57	2.06	0.02	1.63
34.	{Other, Daylight} → {Colony}	0.43	2.04	0.03	1.15
35.	{Agriculture Land} → {Time = 10,Other, Non-highway, 2 wheeler}	0.34	1.85	0.01	1.02
36.	{Intersection} → {Pedestrian}	0.38	1.81	0.02	1.1
37.	{Pedestrian_hit} → {Hospital, Highway}	0.37	1.77	0.02	1.24
38.	{Curve,Fall_height} → {Highway}	0.97	1.73	0.01	8.96
39.	{Day = Monday, Non-highway, 2 wheeler} → {Agriculture Land}	0.67	1.38	0.01	1.49
40.	{Daylight} → {Curve, Hills, Highway, Fall_height}	0.54	1.12	0.2	1
41.	{Agriculture Land, Non-highway} → {Day = Saturday, 2 wheeler}	0.59	1.02	0	1
42.	{Time = 11, Other, 2 wheeler} → {Highway}	0.56	1.01	0	0.99
43.	{Time = 9, Daylight, Agriculture Land} → {Non-highway}	0.44	1.01	0	0.99
44.	{Day = Wednesday} → {Daylight, Agriculture Land, 2 wheeler}	0.45	1.01	0	1
45.	{Month = 11, 2 wheeler} → {Non-highway}	0.44	1.01	0	0.98

accidents but road feature are unknown. Fall height accidents are associated with curve on roads. Three-wheeler accidents have happened at day time but the road features are unknown. Two-wheeler accidents are scattered on non-highway roads where area around is mostly agriculture land and road feature is intersection.

3.5 Association rule for LFAL

In LFAL locations, rule shows that pedestrians are more prone to accidents at intersections on road. In these locations, most of the accidents have occurred after sunset and before sunrise. Two-wheeler accidents mostly occur in night time. In LFAL locations, October and November month have more accidents than any other month. In India, these months have more national festivals such as Vijaydashmi and Diwali, and these festival celebrations cause congested traffic and rush on roads. This congestion and rush could be a cause of more accident rate in these months. Monday and Saturday that are first day of working week and last day of working week have more two-wheeler accidents in comparison to other week days.

The association rules for HFAL, MFAL and LFAL are slightly similar but have different interesting measures. Hence we tried to identify the reasons why the accidents mostly occurred at HFAL locations and identified that most of the HFAL locations are on the highway towards the tourist hill stations where traffic movement are high every day of the week. Another HFAL location was found to be non-highway roads connecting two or more cities which contain all type of transport activities like highways and also non-highways features such as local colonies, markets. Hence these locations contain highway and non-highway traffic on these roads making them vulnerable to road accidents. MFAL locations are mainly highways that go through agriculture land, forest areas and non-tourist hill areas. Usually, these roads are free from side road traffic such as pedestrians on the road. So we can expect vehicles with high speed at these locations. LFAL locations consist of both highway and non-highway roads in the city premises and contain colonies, congested market, hospitals etc. As these locations are scattered throughout the city area and similarly city traffic, pedestrians are also distributed as per their localities in the city, there the accident counts are relatively low as compared to HFAL and MFAL. However, the association rule concludes the following information related to the circumstances of the accident occurrence:

The association rule reveals that curvy roads in hilly regions are more prone to accidents for every type of vehicles. So, one should carefully drive on these locations in order to prevent the chance of accidents. Government officials should also do some prevention work such as putting hoardings about sudden bend and slopes on road. The roads that pass through the agricultural land are risky at night as there are no road side lights. Various intersections are not visible on these roads at night time, which create the chances for an accident as intersection is a place where several vehicles can encounter with each other at the same time. Road lightning in these locations can help in overcoming the accident rates at night time.

Pedestrian hit cases are mostly found near intersections in the market and colonies. Market and colonies are highly congested areas where mostly pedestrians are found. The reason of pedestrian hit cases is attributed to the parked vehicles in the market roads, which makes the pedestrian walk difficult. Hence, there must be different vehicle

parking zones so that vehicles cannot enter the congested area and thereby pedestrian hit accidents can be prevented in these areas.

Vehicle rollover accidents are mainly found in roads through agriculture land. Although the road type feature is unknown for this type of accidents but based on authors personal knowledge, potholes and bad road conditions can be the reason for these accidents. Potholes are usually rare on highways but it can cause severe accidents when a high speed vehicle comes into its contact. So, government should periodically monitor the conditions on road and provide proper maintenance to overcome these accidents. Several other rules indicate that in the night time highways become more sensitive to road accidents for all type of accidents. Other road accident features such as the speed of the vehicle at the time of road accident, condition of the road, driver characteristics and weather information are very much helpful to extract other useful information which can better explain the circumstances of accident occurrence.

Figure 3 illustrates the time-wise distribution of road accidents in HFAL, MFAL and LFAL locations. It indicates that accident pattern in these locations are almost similar except that for MFAL locations there is a downfall at 14:00–16:00 h while an there is an increase in the LFAL and MFAL locations. Figure 4 illustrates the month-wise distribution of road accidents in HFAL, MFAL and LFAL

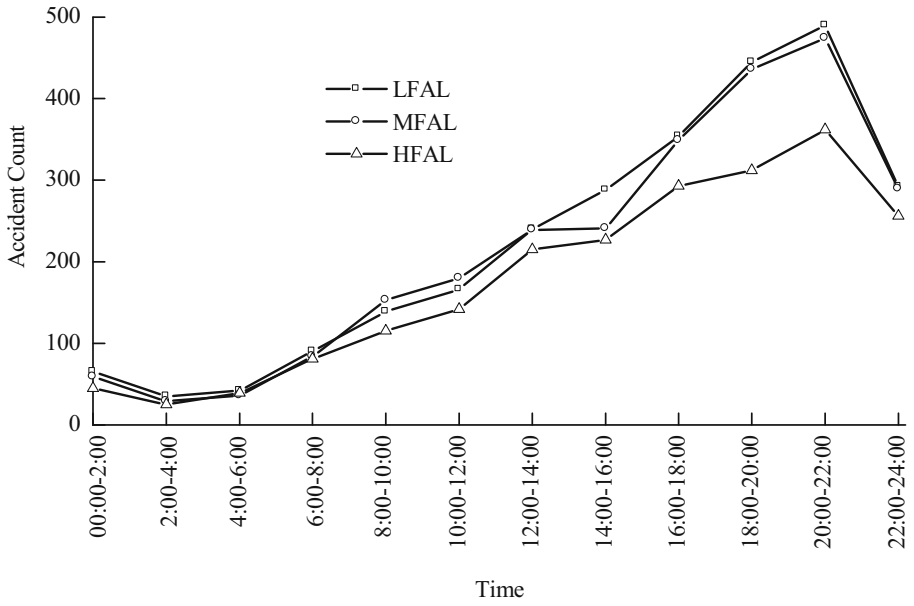

Fig. 3 Time-wise distribution of accidents

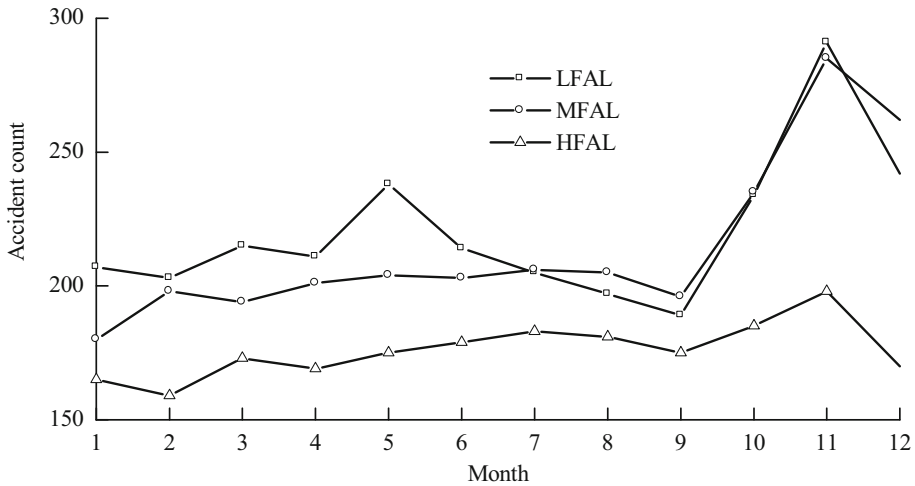

Fig. 4 Month-wise distribution of accidents

locations. Month-wise illustration of accidents shows that there is a different pattern of road accidents. As mentioned that HFAL locations contain tourist hill stations, Fig. 4 illustrates that there is an increase in accident rate from April to July. The reason can be most of the people visit tourist places in summer duration. LFAL locations on September show the minimum peak of accident count. As most of the rainfall occurs in September month, the people are happier to stay at home and hence very few accidents are reported at that duration.

Though authors try to extract the best characteristics of accident occurrence in the available data through data mining techniques, other accident features such as speed of the vehicle at the time of accident, weather information, traffic volume information can surely strengthen our results, which are not available with the data.

4 Conclusion

Data mining has been proven as a reliable technique in analyzing road accident data. Various authors used data mining techniques to analyze road accident data of different countries. Various data mining techniques such as clustering, classification and association rule mining are widely used in the literature to identify reasons that affect the severity of road accidents. It is the first time that *k*-means algorithm is used to identify high- and low-frequency accident locations based on accident count as it provides some technical measures to divide the accident locations based on threshold values. Association rule mining is a very popular technique that can be used to identify the relationship among different sets of attributes that frequently occur together when an accident takes place. In our study, we applied association rule mining algorithm on different groups of accident locations. The rules generated for every group exposed the various factors associated with road accidents in these locations. Although certain rules are quite similar in each group, they have different interest scores for each cluster. The road accident dataset and its analysis using *k*-means clustering and association rule mining algorithm illustrate that this approach can be reused on other accident data with more attributes to identify various other factors associated with road accidents. Although this data mining approach is quite sufficient to uncover reasonable information from the selected data set, the results remain at very general level as source data does not contain other accident related information such as the speed of vehicle at the time of accident, weather information, road surface condition. The data with more number of attributes can reveal more information using our approach.

Acknowledgments We are thankful to GVK-EMRI Dehradun for providing data for our research.

References

1. MORTH (2014) Road Accidents in India 2013. New Delhi: Ministry of Road Transport and Highways Transport Research Wing, Government of India, August 2014. http://morth.nic.in/showfile.asp?lid=1465. Accessed 20 May 2015
2. Kononov J, Janson BN (2002) Diagnostic methodology for the detection of safety problems at intersections. Transp Res Rec. doi:10.3141/1784-07
3. Lee C, Saccomanno F, Hellinga B (2002) Analysis of crash precursors on instrumented freeways. Transp Res Rec. doi:10.3141/1784-01
4. Chen W, Jovanis P (2000) Method for identifying factors contributing to driver-injury severity in traffic crashes. Transp Res Rec. doi:10.3141/1717-01
5. Barai S (2003) Data mining application in transportation engineering. Transport 18:216–223. doi:10.1080/16483840.2003.10414100
6. Tan PN, Steinbach M, Kumar V (2006) Introduction to data mining. Pearson Addison-Wesley, Boston
7. Ponnaluri RV (2012) Road traffic crashes and risk groups in India: analysis, interpretations, and prevention strategies. IATSS Res 35:104–110. doi:10.1016/j.iatssr.2011.09.002
8. Kumar CN, Parida M, Jain SS (2013) Poisson family regression techniques for prediction of crash counts using Bayesian inference. Proc Soc Behav Sci 104:982–991. doi:10.1016/j.sbspro.2013.11.193
9. Parida M, Jain SS, Kumar CN (2012) Road traffic crash prediction on national highways. Indian Highw Indian Road Congr 40:93–103
10. Geurts K, Wets G, Brijs T, Vanhoof K (2003) Profiling of high frequency accident locations by use of association rules. Transp Res Rec. doi:10.3141/1840-14
11. Tesema TB, Abraham A, Grosan C (2005) Rule mining and classification of road accidents using adaptive regression trees. Int J Simul 6:80–94
12. Abellan J, Lopez G, Ona J (2013) Analysis of traffic accident severity using decision rules via decision trees. Expert Syst Appl 40:6047–6054. doi:10.1016/j.eswa.2013.05.027
13. Depaire B, Wets G, Vanhoof K (2008) Traffic accident segmentation by means of latent class clustering. Accid Anal Prev 40:1257–1266. doi:10.1016/j.aap.2008.01.007
14. Kashani T, Mohaymany AS, Rajbari A (2011) A data mining approach to identify key factors of traffic injury severity. Promet-Traffic Transp 23:11–17. doi:10.7307/ptt.v23i1.144
15. Kwon OH, Rhee W, Yoon Y (2015) Application of classification algorithms for analysis of road safety risk factor dependencies. Accid Anal Prev 75:1–15. doi:10.1016/j.aap.2014.11.005
16. MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, volume 1:

Statistics, pp 281–297, University of California Press, Berkeley, 1967. http://projecteuclid.org/euclid.bsmsp/1200512992
17. Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a data set via the gap statistic. J R Statist Soc B 63:411–423. doi:10.1111/1467-9868.00293
18. Han J, Kamber M (2001) Data mining: concepts and techniques. Morgan Kaufmann Publishers, Burlington
19. Piatetsky-Shapiro G (1991) Knowledge discovery in databases. In: Piateski G, Frawley W (eds) Discovery, analysis, and presentation of strong rules. AAAI/MIT Press, Menlo Park, pp 229–248
20. Brin S, Motwani R, Ullman JD, Tsur S (1997) Dynamic itemset counting and implication rules for market basket data. In: Proceedings of the 1997 ACM SIGMOD international conference on management of data, Tucson, Arizona, pp 255–264. doi:10.1145/253260.253325
21. http://www.emri.in
22. Agrawal R, Srikant R (1994) Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th international conference on very large data bases, pp. 487–499