Kumar and Toshniwal *Journal of Big Data* (2015) 2:26
DOI 10.1186/s40537-015-0035-y

Journal of Big Data
a SpringerOpen Journal

RESEARCH Open Access

A data mining framework to analyze road accident data

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

*Correspondence:
sachinagnihotri16@gmail.com
[1] Centre for Transportation Systems (CTRANS), Indian Institute of Technology Roorkee, Roorkee 247667, Uttarakhand, India
Full list of author information is available at the end of the article

Abstract

One of the key objectives in accident data analysis to identify the main factors associated with a road and traffic accident. However, heterogeneous nature of road accident data makes the analysis task difficult. Data segmentation has been used widely to overcome this heterogeneity of the accident data. In this paper, we proposed a framework that used K-modes clustering technique as a preliminary task for segmentation of 11,574 road accidents on road network of Dehradun (India) between 2009 and 2014 (both included). Next, association rule mining are used to identify the various circumstances that are associated with the occurrence of an accident for both the entire data set (EDS) and the clusters identified by K-modes clustering algorithm. The findings of cluster based analysis and entire data set analysis are then compared. The results reveal that the combination of k mode clustering and association rule mining is very inspiring as it produces important information that would remain hidden if no segmentation has been performed prior to generate association rules. Further a trend analysis have also been performed for each clusters and EDS accidents which finds different trends in different cluster whereas a positive trend is shown by EDS. Trend analysis also shows that prior segmentation of accident data is very important before analysis.

Keywords: Data mining, Accident analysis, Road accidents, Clustering

Background

Road and traffic accidents are uncertain and unpredictable incidents and their analysis requires the knowledge of the factors affecting them. Road and traffic accidents are defined by a set of variables which are mostly of discrete nature. The major problem in the analysis of accident data is its heterogeneous nature [1]. Thus heterogeneity must be considered during analysis of the data otherwise, some relationship between the data may remain hidden. Although, researchers used segmentation of the data to reduce this heterogeneity using some measures such as expert knowledge, but there is no guarantee that this will lead to an optimal segmentation which consists of homogeneous groups of road accidents [2]. Therefore, cluster analysis can assist the segmentation of road accidents.

Cluster analysis which is an important data mining technique can be used as a preliminary task to achieve various goals. Karlaftis and Tarko [3] used cluster analysis to categorize the accident data into different categories and further analyzed cluster results using Negative Binomial (NB) to identify the impact of driver age on road accidents.

Springer

© 2015 Kumar and Toshniwal. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Ma and Kockelman [4] used clustering as their first step to group the data into different segments and further they used Probit model to identify relationship between different accident characteristics. Poisson models [5, 6] and negative binomial (NB) models [7–9] have been used extensively to identify the relationship between traffic accidents and the causative factors. It has been widely recognized that Poisson models outperform the standard regression models in handling the nonnegative, random and discrete features of crash counts [10, 11].

Regression analysis (such as linear regression models, negative binomial regression models and Poisson regression models) has been the most popular technique in crash analysis because the connection between accidents and factors affecting them can be evidently identified. Using such information, the accident-prone locations can be located by the traffic engineers, and facilities such as illumination and enforcement, can then be effectively applied. However, they have limited capacity to discover new and unanticipated patterns and relationships that are hidden in conventional databases, [12] demonstrates that certain problem may occur while using traditional statistical analysis to analyze datasets with large dimensions such as an exponential increase in the number of parameters with an increase in number of variables and there could be some invalidity of statistical tests as a due to sparse data. Also, Regression models usually have their own model specific assumptions and predefined underlying relationships between dependent and independent variables. Violation of these assumptions may lead the model to provide erroneous results [13]. Hence, we need a different technique that can be used to analyze road accidents properly and can extract better results. Data mining [14] can be described as the set of techniques used for the extraction of implicit, previously unknown and hidden information from the huge amount of data. Data mining is an upcoming area that is being used by the researchers worldwide for the analysis of various types of transportation data. Several data mining techniques such as clustering, classification, association rule mining have been used to analyzed road safety data.

Chang and Chen [13] analyzed national freeway-1 data from Taiwan using CART and negative binomial regression model. Abellan et al. [15] analyzed two lane rural highway data of Granada, Spain using decision rules extracted from decision tree method. Depaire et al. [2] applied latent class clustering on two road user traffic accident data from 1997 to 1999 of Belgium which divides the accident data into seven clusters. Rovsek et al. [16] analyzed crash data from 2005 to 2009 of Slovenia with classification and regression tree (CART) algorithm. Kashani et al. [17] uses CART to analyze crash records obtained from information and technology department of the Iran traffic police from 2006 to 2008.

This paper proposes a framework that is based on the *cluster analysis* using K modes algorithm and *association rule mining* using Apriori algorithm. Using cluster analysis as a preliminary task can group the data into different homogeneous segments. Association rule mining is further applied on these clusters as well as on *entire data set* (EDS) to generate association rules. In the best of our knowledge, it is the first time that both the approaches have been used together for analysis of road accident data. The result of the analysis proves that using cluster analysis as a preliminary task can help in removing heterogeneity to some extent in the road accident data. The paper is organized as follows: In Sect. "Proposed framework", a framework is proposed to analyze the road accident data. Next, a description of the data set used is given. In Sect. "Results and discussion", the

results and findings are elaborated and discussed. Finally, we concluded in Sect. "Conclusion and suggestion".

Proposed framework

To analyze the data, we develop a framework as shown in Fig. 1. The detailed description of the framework is as follows:

Data preprocessing

Data preprocessing [14] is one of the important tasks in data mining. Data preprocessing mainly deals with removing noise, handle missing values, removing irrelevant attributes in order to make the data ready for the analysis. In this step, our aim is to preprocess the accident data in order to make it appropriate for the analysis.

Clustering algorithm

There are several clustering algorithms [14, 18] exist in the literature. The objective of clustering algorithm is to divide the data into different clusters or groups such that the objects within a group are similar to each other whereas objects in other clusters are different from each other [19]. Hierarchical clustering technique (e.g. Ward method, single linkage, complete linkage, etc.), K means and latent class clustering (LCC) have been used in road accident analysis [2, 3, 20–22]. Another clustering technique is K-modes clustering which is an enhanced version of K means algorithm. LCC [23] is widely used clustering technique which provides several cluster selection criteria [24] to determine the number of clusters. Although, LCC has been widely used for analysis of road accidents to identify clusters in accident data [2, 24, 25] but [26] mentioned that if the data contains large number of categorical attributes, LCC can be computationally infeasible

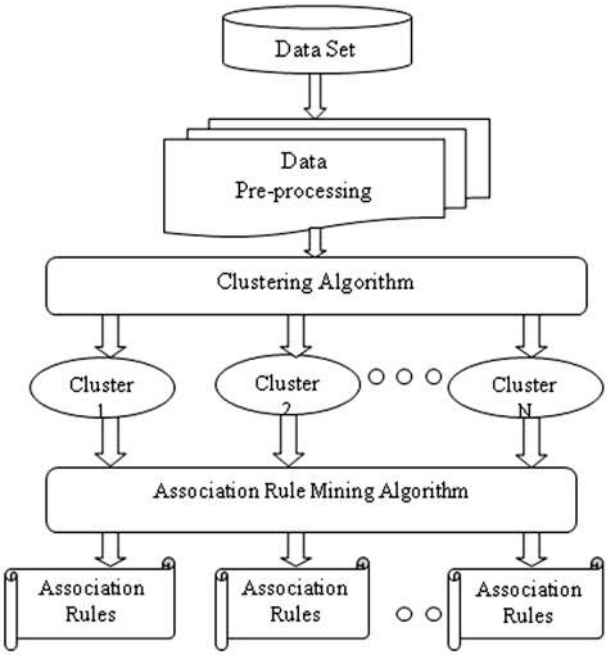

Fig. 1 Proposed framework for analysis

and suggests that K modes algorithm can be a better option and the problem of selecting K in K-modes algorithm can be overcome by using cluster selection criteria of LCC analysis.

In a previous Monte Carlo simulations, it was found that both K modes and LCC have equal efficiency in recovering a known cluster structure [27]. K modes are faster and efficient than LCC in producing locally minimal clustering results. In this paper, we are making use of both K modes clustering and cluster selection criteria of LCC cluster analysis with the following reasons:

a) K modes are a better option for data with large number of categorical attributes.
b) The problem of identifying number of K can be solved by cluster selection criteria used by LCC.
c) K modes can handle large number of data with good efficiency.

Here, we are providing a brief description of the K modes clustering algorithm.

The K-modes clustering technique is an enhanced version of traditional k means algorithm. The major extensions to the k means algorithm to k modes algorithm is the distance measure and the clustering process which are explained below:

Distance measure

Given a data set D, the distance between two objects X and Y, where X and Y are described by N categorical variables, can be computed as follows:

$$\boldsymbol{d}(\boldsymbol{X},\boldsymbol{Y}) = \sum_{\boldsymbol{i}=1}^{\boldsymbol{N}} \boldsymbol{\delta}(\boldsymbol{Xi},\boldsymbol{Yi}) \tag{1}$$

where,

$$\delta(\boldsymbol{X}i,\ \boldsymbol{Y}i) = \begin{cases} 0, & \mathbf{Xi} = \mathbf{Yi} \\ 1, & \mathbf{Xi} \neq \mathbf{Yi} \end{cases} \tag{2}$$

In the above equations, Xi and Yi are the attribute i values in object X and Y. This distance measure is often referred as simple matching dissimilarity measure. The more the number of differences in categorical values of X and Y, more the different two objects are.

K-mode clustering procedure:

In order to cluster the data set D into k cluster, K-modes clustering algorithm perform the following steps:

1. Initially select k random objects as cluster centers or modes.
2. Find the distance between every object and the cluster centre using distance measure defined in Eq. 1.
3. Assign each object to the cluster whose distance with the object is minimum.
4. Select a new center or mode for every cluster and compare it with the previous value of centre or mode; if the values are different, continue with step 2.

Association rules

Association rule mining [28] is a very popular data mining technique that extracts interesting and hidden relations between various attributes in a large data set. Association rule mining produces a set of rules that define the underlying patterns in the data set. The associativity of two characteristics of accident is determined by the frequency of their occurrence together in the data set. A rule $A \rightarrow B$ indicates that if A occurs then B will also occur.

Given a data set D of n transactions where each transaction $T \in D$. Let $I = \{I_1, I_2, \ldots I_n\}$ is a set of items. An item set A will occur in T if and only if $A \subseteq T$. $A \rightarrow B$ is and association rule, provided that $A \subset I$, $B \subset I$ and $A \cap B = \emptyset$.

Agrawal and Srikant [29] proposed an algorithm known as Apriori algorithm to find the association rules from large datasets. The pseudo-code for traditional association rule mining algorithm for frequent itemset generation is as follows:

L_k = {Frequent itemset of size k}
C_k = {Candidate itemset of size k}
L_1 = {frequent 1 itemsets};
k=1;
while ($L_k - 1 \neq \Phi$) then
    C_{k+1} = candidates generated from L_k
    For each transaction t ∈ D do
        Increment the count of candidates in C_k+1 that also contained in t
    L_k+1 = candidates in Ck+1 with minimum support
    k=k+1;
Return $U_k L_k$

Further association rules are generated from the frequent itemsets and strong rules based on interestingness measures are taken for the analysis.

Interestingness measures

An association rule is considered as a strong rule if is satisfies the minimum threshold criteria, i.e., confidence and support. A minimum support S of a rule $A \rightarrow B$ indicates that in x % of all transactions A and B together occurs and it can be calculated using Eq. (3); whereas a confidence C of a rule indicates that in C % of all transaction when A occur then B also occurs and it can be calculated using Eq. (4). Lift is another interestingness measure of a rule, which can be calculated using Eq. (5). A value greater than 1 for the lift measures indicates that the appearance of A and B together is more than expected whereas a value lower than 1 indicates reverse of the concept. So a rule is considered as strong if it has a value greater than 1 for the lift parameter.

$$\textit{Support} = P(A \cap B) \tag{3}$$

$$\textit{Confidence} = P(A|B) = \frac{P(A \cap B)}{P(A)} \tag{4}$$

$$Lift = \frac{P(A \cap B)}{(P(A)P(B))} \tag{5}$$

Data set description

Accident data for this research were obtained from GVK-Emergency Management Research Institute, Dehradun. The data set consists of 11,574 road accidents for 6 years period from 2009 to 2014, in Dehradun District of Uttarakhand State. After preprocessing of the data, 11 variables were identified satisfactory for the research. The data set comprised of accident characteristics (time, day, month, type of accident, number of injured victims), victims age and gender, road type, road feature and area around accident site. The brief information about this data is given in Table 1.

Results and discussion

Cluster analysis

The basic requirement for cluster analysis is to determine the number of clusters to be formed by clustering algorithm. To achieve the solution for this, we used several information criteria such as Akaike Information Criteria (AIC) [30] Bayesian Information Criterion (BIC) [31] and Consistent AIC (CAIC) [32]. We generated 15 models for 1 cluster to 15 clusters. Figure 2 illustrates the evolution of BIC, AIC and CAIC for the 15 models generated. It shows that there is a reduction in the values of AIC, BIC and CAIC with an increase in the number of clusters. Based on the Fig. 2 (a low score is considered as good), we select the model with 6 clusters as there is no improvement after this. Our selection also follows the approach used by previous studies [2, 24].

After getting number of clusters to be made, we used K-modes algorithm using R statistical software to segment the accident data set. After getting appropriate segmentation of the data set, our next task is the characterization of each cluster. A thorough analysis of each cluster reveals that accident variables that categorized the clusters were TOA, RTY, ROF and ARA. The brief description of cluster is given below:

Cluster 1 (C1)

It consists of 69 % of two wheeler accidents which are distributed on intersections near markets, hospitals, local colonies across highways and non-highway roads. Those accidents which occurred on intersections and curves on highways involved one injury only. Two wheeler accidents at non-highway locations are mostly involved two injuries.

Cluster 2 (C2)

It consists of two wheeler accidents that occurred on highway that goes through a hill area, forest area or agriculture land area. In this cluster 64 % of accidents involved more than two injuries and 26 % accidents involved 1 injury and rest involved more than 2 injuries.

Cluster 3 (C3)

It consists of all accidents which were due to vehicle falling down from height. Around 80 % of these cases are critical where ARA was hill. Rest of the accidents of this category

Table 1 Road accident attributes

S. no.	Attribute	Code	Value	Total	Criticality	
					Critical	Non-critical
1	NOV: number of injury	1	1 injury	5932	689	5243
		2	2 injuries	2598	451	2147
		+2	>2 injuries	3044	114	2930
2	AGE: age	CHL	< 18 years	988	268	720
		YNG	18–30 years	5954	654	5300
		ADL	30–60 years	3045	165	2880
		SNR	>60 years	1587	167	1420
3	GND: gender	M	Male	8625	952	7673
		F	Female	2949	302	2647
4	TOD: time of day	T1	[0–4]	678	45	633
		T2	[4–8]	1032	164	868
		T3	[8–12]	1358	258	1100
		T4	[12–16]	1972	126	1846
		T5	[16–20]	3768	245	3523
		T6	[20–24]	2766	416	2350
5	MON: month	WNT	Winter	2822	325	2497
		SPR	Spring	2787	312	2475
		SMR	Summer	3144	368	2776
		ATM	Autumn	2821	249	2572
6	LOR: lighting on road	DLT	Day light	3850	268	3582
		DUS	Dusk	3203	429	2774
		RLT	Road light	1665	126	1539
		NLT	No light	2856	431	2425
7	ROF: roadway feature	INT	Intersection	3526	374	3152
		SLP	Slope	1157	212	945
		CUR	Curve	2827	266	2561
		UNK	Unknown	4064	402	3662
8	RTY: road type	HIW	Highway	6032	785	5247
		NHW	Non-highway	5542	469	5073
9	ASV: accident severity	CR	Critical	1254	1254	0
		NC	Non-critical	10320	0	10320
10	ARA: area around	AGL	Agriculture land	1984	289	1695
		MAR	Market	2069	145	1924
		COL	Colony	3250	119	3131
		FOR	Forest	1165	267	898
		HIL	Hill area	2354	345	2009
		HOS	Hospital	752	89	663
11	TOA: type of accident	TWH	Two wheeler	3688	194	3494
		THW	Three wheeler	255	55	200
		MVH	Multi-vehicular	855	64	791
		VFH	Vehicle fall height	2132	398	1734
		VRO	Vehicle roll over	1356	129	1227
		PH	Pedestrian hit	1580	265	1315
		NM	Non-motorized	254	12	242
		MC	Multi-casualty	364	16	348
		FO	Fixed object/divider hit	987	121	866
		OT	Others	103	0	103

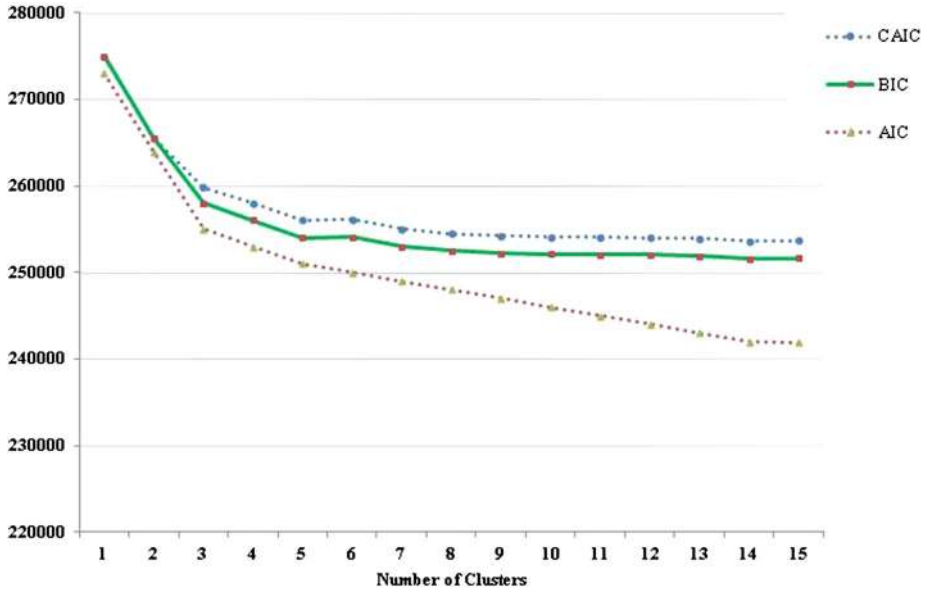

Fig. 2 AIC, BIC and CAIC representation for different set of clusters

belongs to non-critical injury. About 68 % of these accidents involved more than two injuries and rest accidents were two injuries involved.

Cluster 4 (C4)

It consists of accidents involving multiple vehicle accidents and divider hit/fixed object hit cases. The accidents that are mostly happened in night time on highways are critical accidents whereas accidents at other locations such as market, colonies at night time are non-critical in this cluster.

Cluster 5 (C5)

It consists of accidents involving pedestrian hit cases. Most of the pedestrian hit cases have happened in market, near hospitals, and other populated areas. Pedestrian hit accidents at night time were critical whereas at day time these accidents have minor injuries. Pedestrian hit cases are distributed among all areas.

Cluster 6 (C6)

It consists of the accidents involving vehicle roll-over cases. Vehicle roll-over cases were found at curves and slopes on highways. It has been observed that 40 % of these accidents have happened on the forest and agriculture land areas. About 55 % of vehicle roll over cases are found at unknown road features.

The size and description of each cluster is tabularized in Table 2.

All these clusters are further analyzed using association rule mining to find the correlation among different attributes in the data.

Table 2 Size and description of clusters

Cluster 1	Cluster description	Count	Size (%)
1	Two wheeler accidents on road intersections and curves near colonies and markets	3181	27.48
2	Two wheeler accident occurred on highways near hill, forest and agriculture land area	1772	15.31
3	All fall height accidents with two or more injuries	1928	16.66
4	Multiple vehicle accidents and fixed object hit accidents in no light condition	1394	12.04
5	Pedestrian hit cases	1746	15.08
6	Vehicle roll-over accidents	1553	13.42

Association rule mining

Apriori algorithm [28] has been applied on every cluster to generate association rules. In order to generate association rules with minimum 30 % support values are generated for each cluster and EDS. These rules are also evaluated on the basis of confidence and lift measures. The strong rules with high lift value are considered for analysis. The strong 10 rules for each cluster and EDS have been shown in this paper. Table 3 shows the association rules generated in descending order of the lift value.

Association rule for cluster 1 shows that two wheeler accidents are mainly occurs on specific road segments such as intersections at community areas, i.e., colony, markets and hospitals. Intersections in colonies near highways are more prone to two wheeler accidents than colonies on non-highways. Also market areas are more likely to have two wheeler accidents with two or more injuries at evening around 4:00 p.m. to 8:00 p.m. Rules revealed that hospitals area are also associated with two wheeler accidents but most of the accidents at this place have happened at night time after 8:00 p.m.

Association rules for cluster 2 indicates that forest area and agriculture land area that are aside of certain highways are dangerous for two wheeler accidents as sudden bend, slope at night time can cause imbalance of driver and may cause accidents. Rules show that curves on hilly highways involves two injuries and mostly young people are involved in such accidents. Also, no light areas such as forest are also prone to accidents in night time. Highways with agriculture land area aside are found to be accident prone areas.

Association rule for cluster 3 shows that most of the vehicle-fall from height accidents involved more than 2 injured. It is found that vehicles falling from height on hilly highways are severe accident where more than two injured persons are there. The reason might be the vehicle type is four-wheeler or similar category which transports more than 2 persons at a time. Also, it shows that mostly vehicles fall from height from hill location are due to a curve on road that is the main characteristics of the hills.

Association rules for cluster 4 indicate that multi-vehicular and fixed object/divider hit accidents are mostly occurred at night time on highway roads. Intersections on highways are another road feature for such type of accidents. Mostly the areas with no light condition are more accident prone in night time and results in critical accidents. Rules show that curve at agriculture land and forest area and intersection at highways are more dangerous at night time as it is difficult for a speedy vehicle to judge the vehicles from opposite side and fixed object to avoid collision. Some rules revealed that these accidents also

Table 3 Cluster-wise association rules

Rule no.	Rule body	Support	Confidence	Lift
Cluster 1				
1	{HIW, INT, COL} →{1}	0.54	0.75	5.24
2	{HIW, CUR} → {1, DUS}	0.45	0.86	4.47
3	{NHW, INT, MAR} → {>2, DLT}	0.65	0.61	2.31
4	{NHW, INT} → {COL}	0.38	0.52	2.63
5	{HIW, MAR} → {+2}	0.35	0.55	1.54
6	{NHW, COL} → {+2}	0.58	0.65	1.26
7	{INT, COL} → {NLT}	0.62	0.66	1.23
8	{MAR, DUS} → {INT, T5}	0.36	0.54	1.20
9	{HIW, HOS} → {T6}	0.54	0.84	1.14
10	{MAR, T6} → {HIW}	0.47	0.69	1.11
Cluster 2				
11	{HIW, SLP} → {HIL}	0.63	0.8	3.16
12	{HIW, NLT} → {FOR}	0.56	0.74	3.14
13	{HIW, AGL} → {+2}	0.40	0.68	2.75
14	{HIW}→ {AGL}	0.54	0.75	2.71
15	{FOR, T6] → {CUR, NLT}	0.56	0.74	1.98
16	{CUR} → {HIL, T2]	0.69	0.71	1.95
17	{HIL, CUR} → {+2}	0.36	0.65	1.65
18	{AGL, CUR} → {T5}	0.42	0.58	1.35
19	{YNG, HIL} → {T3, CUR}	0.45	0.64	1.23
20	{FOR, UNK} → {NLT}	0.39	0.46	1.15
Cluster 3				
21	{HIW, HIL} → {+2, CR}	0.78	0.90	3.18
22	{CUR, HIL, ADL} → {HIW}	0.64	0.85	2.93
23	{HIW, HIL, +2} → {CR}	0.85	0.95	2.87
24	{HIW, CUR} → {HIL}	0.82	0.88	2.58
25	{FOR} → {NC}	0.78	0.65	1.78
26	{NHW} → {FOR}	0.45	0.50	1.76
27	{HIL, ADL} → {CR}	0.42	0.65	1.64
28	{T2, HIL} → {NLT}	0.39	0.46	1.30
29	{CUR} → {HIL, T5}	0.46	0.80	1.25
30	{HIL, SLOPE} → {HIW}	0.35	0.74	1.22
Cluster 4				
31	{HIW, NLT} → {INT, T6}	0.65	0.78	4.85
32	{HIW, CUR} → {NLT, T1}	0.78	0.85	3.81
33	{NLT} → {INT}	0.74	0.7	3.77
34	{HIW, DAY} → {INT, NC}	0.70	0.84	2.73
35	{NHW, NLT} → {SLP}	0.40	0.65	3.38
36	{T6, AGL} → {NLT, CR}	0.55	0.65	2.56
37	{CUR, T1, HIW} → {FOR}	0.36	0.46	2.16
38	{FOR} → {NLT, HIW}	0.54	0.74	1.98
39	{RLT, MAR} → {INT, NC}	0.57	0.66	1.80
40	{HIW} → {NLT, AGL}	0.56	0.84	1.65
Cluster 5				
41	{NHW, MAR} → {YNG}	0.48	0.87	3.22
42	{COL, INT} → {NHW, DUS}	0.56	0.92	3.11
43	{INT, NLT} → {AGL, T2}	0.58	0.74	2.31
44	{HIW, INT} → {MAR}	0.38	0.46	2.11

Table 3 continued

Rule no.	Rule body	Support	Confidence	Lift
45	{T3, DAY} → {NC, INT}	0.65	0.69	2.05
46	{NLT, INT} →{CR, T6}	0.39	0.82	1.95
47	{HOS, NLT} → {T6}	0.54	0.81	1.80
48	{MAR} → {NC, DAY}	0.46	0.64	1.78
49	{RLT, INT} → {MAR}	0.51	0.79	1.50
50	{AGL, ADL} → {NLT}	0.36	0.78	1.45
Cluster 6				
51	{FOR, SLP} → {NLT}	0.45	0.65	3.62
52	{AGL, DUS} → {CUR}	0.35	0.54	3.58
53	{FOR} → {NLT, UNK}	0.65	0.78	2.46
54	{HIW, T2} → {AGL}	0.58	0.81	2.42
55	{AGL, UNK} → {NHW}	0.64	0.85	2.12
56	{UNK} →{DAY, COL}	0.58	0.65	1.94
57	{RLT, UNK} → {NC, MAR}	0.45	0.75	1.68
58	{ADL, AGL} → {UNK}	0.64	0.68	1.42
59	{AGL, INT} → {NC}	0.39	0.75	1.35
60	{FOR} → {CUR}	0.40	0.68	1.34
Entire data set (EDS)				
61	{HIW, INT} → {NC, MAR}	0.40	0.75	5.45
62	{FOR, NLT, M} → {HIW, TWH}	0.52	0.65	4.35
63	{HIW, NC} → {AGL, DAY}	0.64	0.79	4.25
64	{NHW, T5} → {NC, COL}	0.38	0.65	4.36
65	{HIW, HIL} → {NLT, NC}	0.42	0.74	3.89
66	{HIW, YNG} → {NC, TWH}	0.46	0.65	3.48
67	{NHW, MAR} → {M, NC}	0.47	0.65	3.26
68	{HIW, UNK} → {NC, TWH}	0.56	0.69	2.98
69	{NHW, T6, UNK} → {NC}	0.65	0.62	2.46
70	{NHW, COL} → {NC, M}	0.39	0.74	2.15

occurred at intersections in market area with road light condition but these are non-critical accidents.

Association rule for cluster 5 shows that local colonies on non-highways locations are the major places of pedestrian hit cases. Other places where pedestrian hit cases are found are the market locations on non-highway roads. The reasons may be that most of the pedestrians are found at these places. Pedestrian hit accidents at night time are found as critical. Rules show that hospital areas with no light conditions after evening become more prone to pedestrian hit accidents. Intersections at market area are also found dangerous for pedestrians. Pedestrian hit accidents that have occurred at agriculture land area involved mostly adults where accidents at market area involved mostly young people.

Association rules for cluster 6 indicate that vehicle roll-over accidents are occurred at night in forest area and roads near agriculture land areas. A slope in forest road and a curve on road is the reasons involved in these accidents. A forest road is more prone to vehicle roll-over accidents in night time. Although rules revealed that vehicle roll-over accidents are scattered at every road condition and road type, but most of these

accidents have happened on forest areas and agriculture land areas on both highway and non-highways. Rules shows that UNK road feature are highly involved in this cluster. Our survey reveals that the locations where these accidents occurred were having different size of potholes and bad road surface that probably causes these accidents.

Association rule for EDS are also generated to distinguish between findings using clustering and without clustering. An association rule for EDS does not reveal enough information that can be important to identify factors affecting road accidents. The rules only show that accidents are scattered at every type of road conditions and does not identify any critical accidents. Few rules are there which focused on two wheeler accidents as the number of two wheeler accidents is comparatively high.

Hence, association rules generated for every cluster identifies the different accident prone circumstances for every cluster. Our results show that performing cluster analysis as a preliminary task can identify more important findings which can remain hidden if only entire data set is analyzed. In general, the major differences identified between the clusters and EDS are given as follows:

- Only two wheeler accidents are identified in EDS that satisfies minimum support of 30 %, other accident type remain hidden.
- Rules for EDS do not reveal the obvious impact of road features on accidents such as it only shows that intersections are accident prone for every accident type, but rules for clusters shows that its probability of being accident prone varies for different clusters.
- Forming cluster before rule generation gives various rules that are mainly associated with that cluster, but rules for EDS only shows a common association for each accident type which is not interesting.
- A majority of unknown road feature is there in EDS rules but after cluster analysis it seems that its impact is associated with few clusters.

Trend analysis

Monthly analysis

For every cluster and EDS, we performed a trend analysis on monthly road accident counts for each cluster. Figures 3a–d and 4a, b illustrates the month wise trend for cluster 1 to cluster 6, respectively. Figure 4c illustrates the trend for EDS. The trend for EDS shows a positive trend which when compared to different cluster's trend is found different. Cluster 1 and cluster 5 have strong positive trend. Cluster 2 and cluster 3 has slight positive trend. Cluster 4 has a negative trend and cluster 6 has approximately straight positive trend. All these trends are different from EDS trend. Hence results of month wise trend analysis also indicate that clustering of data prior to analysis can reveal important information which can be hidden if only EDS is analyzed.

Hourly analysis

Next to monthly analysis, we also performed hourly trend analysis of road accidents for all clusters and EDS. The hourly analysis of clusters and EDS are shown in Figs. 5 and 6. Figures 5a–d and 6a, b shows hourly analysis of cluster 1 to cluster 6, respectively, and Fig. 6c shows analysis for EDS. Figures 5 and 6 illustrates that C1, C2 and C5 shows a

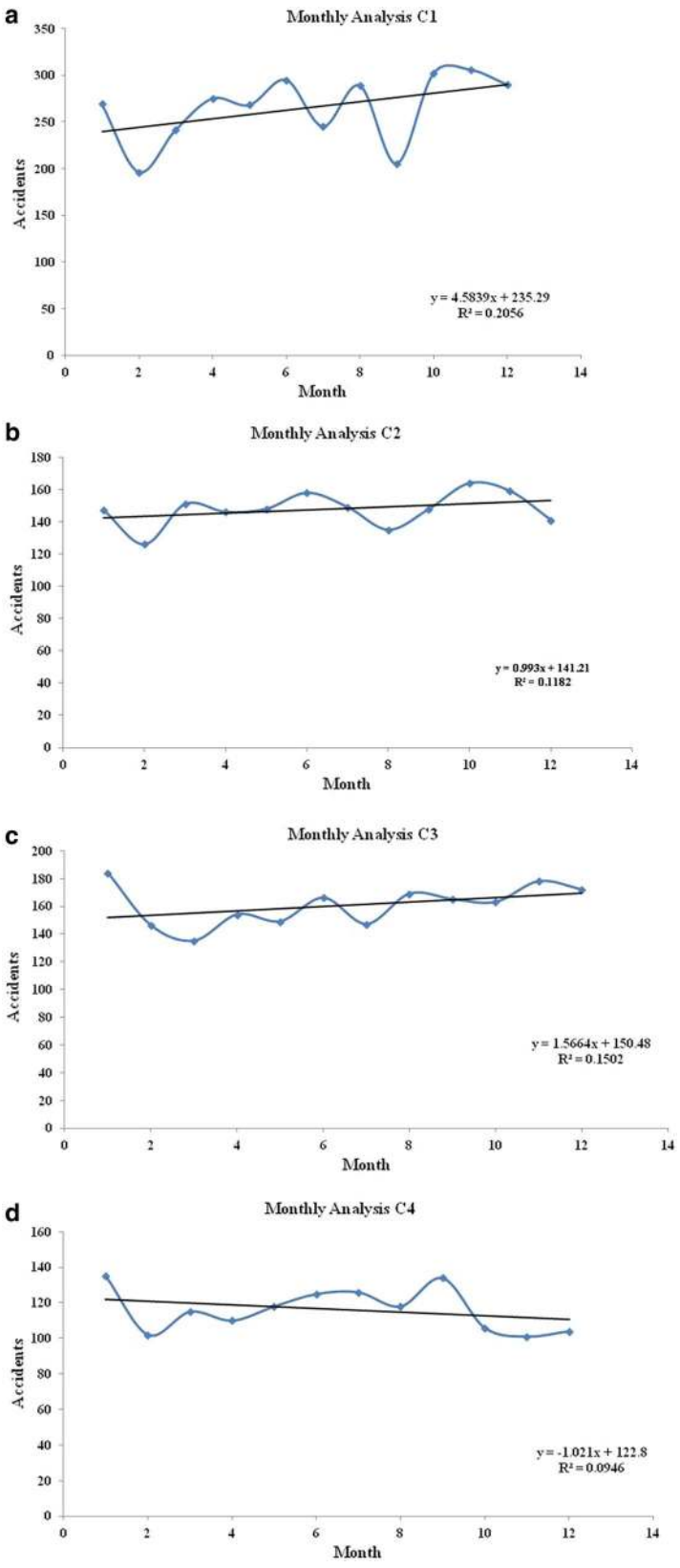

Fig. 3 **a** Month wise trend analysis of cluster 1. **b** Month wise trend analysis of cluster 2. **c** Month wise trend analysis of cluster 3. **d** Month wise trend analysis of cluster 4

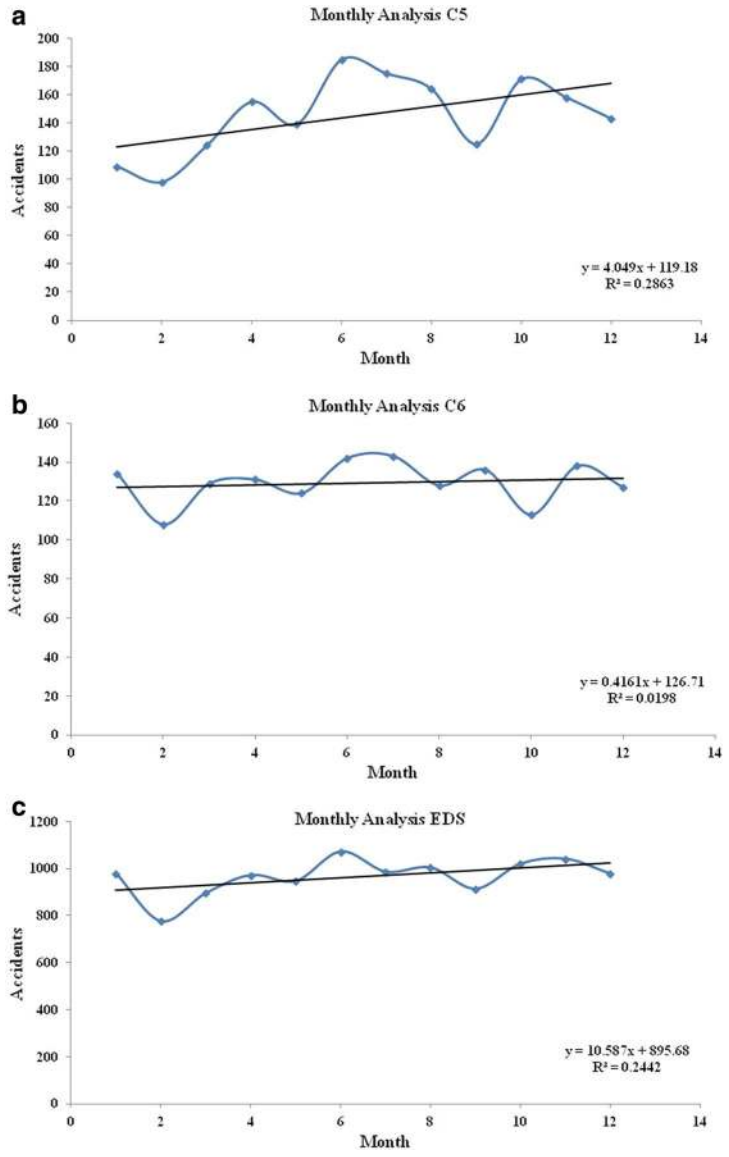

Fig. 4 **a** Month wise trend analysis of cluster 5. **b** Month wise trend analysis of cluster 6. **c** Month wise trend analysis of EDS

trend that are rather similar to EDS whereas C3 and C6 has although positive but slightly different trend than EDS. The hourly trend for C4 is different from every other cluster and EDS as it shows a negative trend. We could see that C4 also has a negative trend for monthly analysis. Hence, our results show that using cluster analysis as a preliminary task for accident data analysis can surely results in unknown findings which are very difficult if only whole data set is analyzed. Also, use of cluster analysis as an initial task for any accident data analysis removes the heterogeneity of the data to some extent, which makes further analysis of the data easier. There our findings have the same opinion with past studies [2, 3, 20, 22] that in order to improve the homogeneity in the data, it is advisable to perform clustering on the road accident data set being used for analysis.

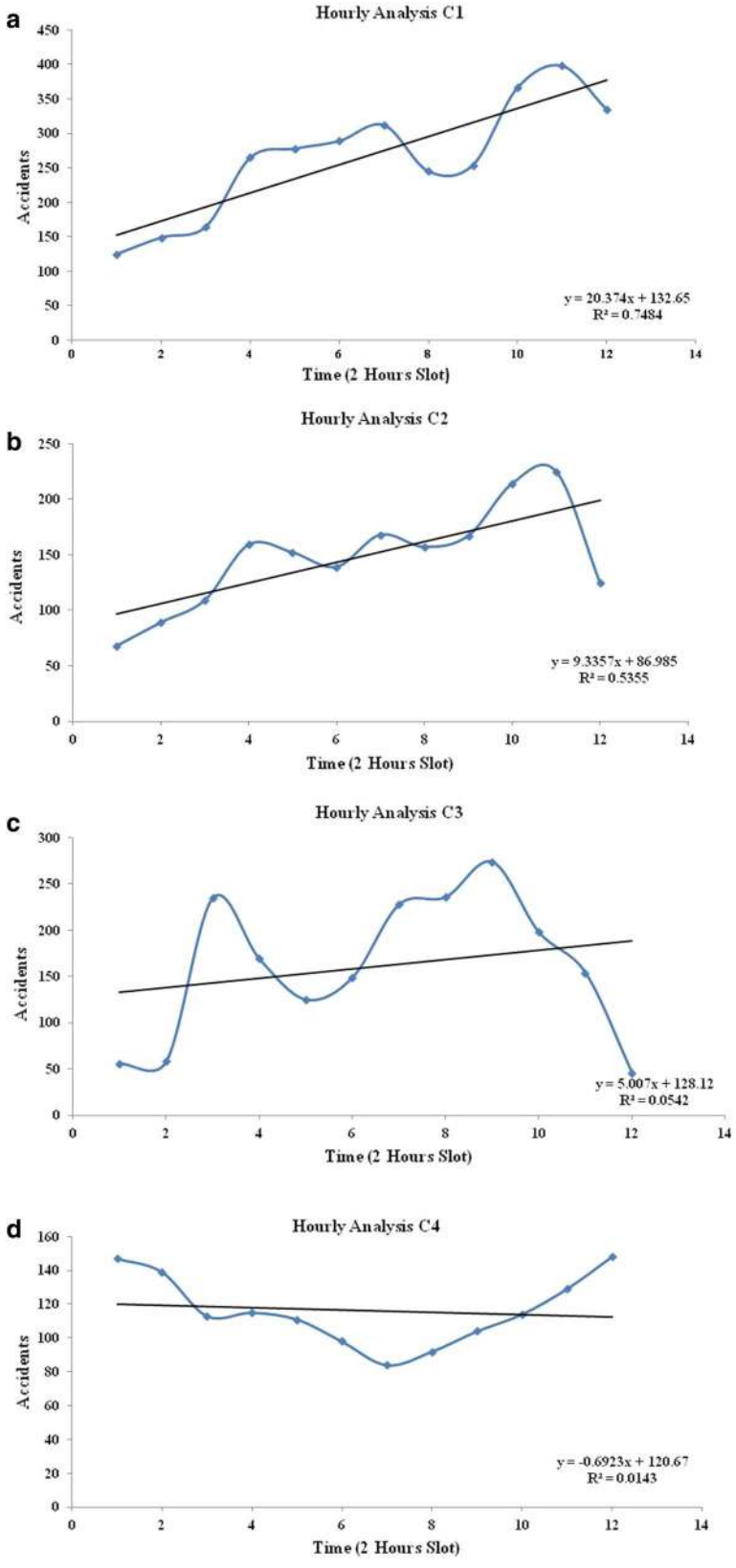

Fig. 5 **a** Time wise trend analysis of cluster 1. **b** Time wise trend analysis of cluster 2. **c** Time wise trend analysis of cluster 3. **d** Time wise trend analysis of cluster 4

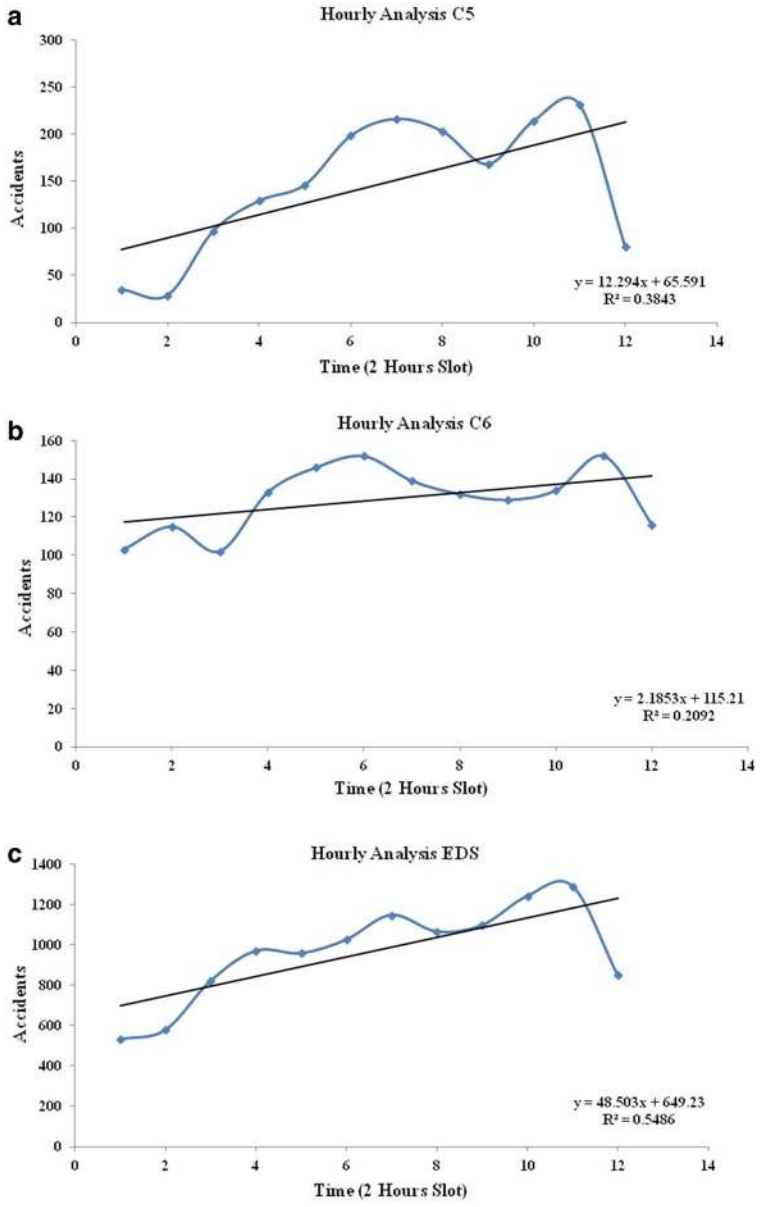

Fig. 6 **a** Time wise trend analysis of cluster 5. **b** Time wise trend analysis of cluster 6. **c** Time wise trend analysis of EDS

Conclusion and suggestions

In this paper, we proposed a framework for analyzing accident patterns for different types of accidents on the road which makes use of K modes clustering and association rule mining algorithm. The study uses 11,574 accidents that have occurred on Dehradun district road network during 2009 to 2014. K modes clustering find six cluster (C1–C6) based on attributes accident type, road type, lightning on road and road feature. Association rule mining have been applied on each cluster as well as on EDS to generate rules. Strong rules with high lift values are taken for the analysis. Rules for every cluster reveal the circumstances associated with the accidents within that cluster. These rules are

compared with the rules generated for the EDS and comparison shows that association rules for EDS does not reveal appropriate information that can be associated with an accident. More information can be identified if more feature are available that is associated with an accident. To strengthen our methodology, we also performed trend analysis of all clusters and EDS on monthly and hourly basis. The results of trend analysis also supports our methodology that performing clustering prior to analysis helps in identify better and useful results that we cannot obtained without using cluster analysis.

Authors' contributions

DT contributed for the underlying idea, helped drafting the manuscript and played a pivotal role guiding and supervising throughout, from initial conception to the final submission of this manuscript. SK developed and implemented the idea, designed the experiments, analyzed the results and wrote the manuscript. Both authors read and approved the final manuscript.

Author details

[1] Centre for Transportation Systems (CTRANS), Indian Institute of Technology Roorkee, Roorkee 247667, Uttarakhand, India. [2] Computer Science and Engineering Department, Indian Institute of Technology Roorkee, Roorkee 247667, Uttarakhand, India.

Acknowledgements

We are thankful to GVK-Emergency Management Research Institute Dehradun, Uttarakhand to provide data for our research. We are also thankful to MHRD to provide scholarship to do research for the Ph.D. program.

Competing interests

The authors declare that they have no competing interests.

Received: 12 October 2015 Accepted: 9 November 2015
Published online: 21 November 2015

References

1. Savolainen P, Mannering F, Lord D, Quddus M. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accid Anal Prev. 2011;43:1666–76.
2. Depaire B, Wets G and Vanhoof K. Traffic accident segmentation by means of latent class clustering, accident analysis and prevention, vol. 40. Elsevier; 2008.
3. Karlaftis M, Tarko A. Heterogeneity considerations in accident modeling. Accid Anal Prev. 1998;30(4):425–33.
4. Ma J, Kockelman K. Crash frequency and severity modeling using clustered data from Washington state. In: IEEE Intelligent Transportation Systems Conference. Toronto Canadá; 2006.
5. Jones B, Janssen L, Mannering F. Analysis of the frequency and duration of freeway accidents in Seattle, accident analysis and prevention, vol. 23. Elsevier; 1991.
6. Miaou SP, Lum H. Modeling vehicle accidents and highway geometric design relationships, accident analysis and prevention, vol. 25. Elsevier; 1993.
7. Miaou SP. The relationship between truck accidents and geometric design of road sections–poisson versus negative binomial regressions, accident analysis and prevention, vol. 26. Elsevier; 1994.
8. Poch M, Mannering F. Negative binomial analysis of intersection-accident frequencies. J Transp Eng. 1996;122.
9. Abdel-Aty MA, Radwan AE. Modeling traffic accident occurrence and involvement. Accid Anal Prev Elsevier. 2000;32.
10. Joshua SC, Garber NJ. Estimating truck accident rate and involvements using linear and poisson regression models. Transp Plan Technol. 1990;15.
11. Maher MJ, Summersgill I. A comprehensive methodology for the fitting of predictive accident models. Accid Anal Prev Elsevier. 1996;28.
12. Chen W, Jovanis P. Method of identifying factors contributing to driver-injury severity in traffic crashes. Transp Res Rec. 2002:1717.
13. Chang LY, Chen WC. Data mining of tree based models to analyze freeway accident frequency. J Saf Res Elsevier. 2005;36.
14. Tan PN, Steinbach M, Kumar V. Introduction to data mining. Pearson Addison-Wesley; 2006.
15. Abellan J, Lopez G, Ona J. Analyis of traffic accident severity using decision rules via decision trees, vol. vol. 40. Expert System with Applications: Elsevier; 2013.
16. Rovsek V, Batista M, Bogunovic B. Identifying the key risk factors of traffic accident injury severity on Slovenian roads using a non-parametric classification tree, transport. UK: Taylor and Francis; 2014.
17. Kashani T, Mohaymany AS, Rajbari A. A data mining approach to identify key factors of traffic injury severity, promet-traffic & transportation, vol. 23; 2011.
18. Han J, Kamber M. Data Mining: Concepts and Techniques. USA: Morgan Kaufmann Publishers; 2001.
19. Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc. 2002;97(458):611–31.
20. Sohn SY. Quality function deployment applied to localtraffic accident reduction. Accid Anal Prev. 1999;31:751–61.
21. Ng KS, Hung WT, Wong WG. An algorithm for assessing the risk of traffic accident. J Saf Res. 2002;33:387–410.

22. Pardillo-Mayora JM, Domínguez-Lira CA, Jurado-Pina R. Empirical calibration of a roadside hazardousness index for Spanish two-lane rural roads. Accid Anal Prev. 2010;42:2018–23.
23. Vermunt JK, Magidson J. Latent class cluster analysis. In: Hagenaars JA, McCutcheon AL, editors. Advances in latent class analysis. Cambridge: Cambridge University Press; 2002.
24. Oña JD, López G, Mujalli R, Calvo FJ. Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks, accident analysis and prevention, vol. 51; 2013.
25. Kaplan S, Prato CG. Cyclist-motorist crash patterns in denmark: a latent class clustering approach. Traffic Inj Prev. 2013;14(7):725–33.
26. Chaturvedi A, Green P, Carroll J. K-modes clustering. J Classif. 2001;18:35–55.
27. Goodman LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrica. 1974;62.
28. Agrawal R, Srikant R. Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on very large data bases; 1994. pp. 487–99.
29. Akaike H. Factor analysis and AIC. Psychome. 1987;52:317–32.
30. Raftery AE. A note on Bayes factors for log-linear contingency table models with vague prior information. J Roy Stat Soc B. 1986;48:249–50.
31. Fraley C, Raftery AE. How many clusters? Which clustering method? Answers via model-based cluster analysis. Comput J. 1998;41:578–88.
32. Wong SC, Leung BSY, Loo BPY, Hung WT, Lo HK. A qualitative assessment methodology for road safety policy strategies. Accid Anal Prev. 2004;36:281–93.

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Immediate publication on acceptance
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com