**A comparative analysis of heterogeneity in road accident data using data mining techniques**

**Analysis and Predict the Nature of Road Traffic Accident**

**Sachin Kumar1 · Durga Toshniwal2 · Manoranjan Parida3**

Ram Prasanth T\*, Spanglar Diaz V\*, Surendran N\*, Udhayavel V\*, Dr.C.Anand\*\*, Mrs.N.Vasuki\*\*\*

Final year, Dept. of Computer Science and Engineeirng

K.S.R. College of Engineering

\*Final year, Department of Computer Science and Engineering

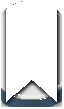
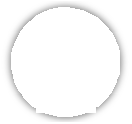
\*\*Associate Professor, Department of Computer Science and Engineering

K.S.R College of Engineering , Tiruchengode.

\*\*\*Assistant Professor, Department of Computer Science and Engineering

IRTT , Erode.

**Abstract** Road accidents are one of the most imperative fac- tors that affect the untimely death among people and economic loss of public and private property. Road safety is a term asso- ciated with the planning and implementing certain strategy to overcome the road and traffic accidents. Road accident data analysis is a very important means to identify various fac- tors associated with road accidents and can help in reducing the accident rate. The heterogeneity of road accident data is a big challenge in road safety analysis. In this study, we are making use of latent class clustering (LCC) and k-modes clus- tering technique on a new road accident data from Haridwar, Uttarakhand, India. The main focus to use both the techniques is to identify which technique performs better. Initially, we applied LCC and k-modes clutering technique on road acci- dent data to form different clusters. Further, Frequent Pattern (FP) growth technique is applied on the clusters formed and entire data set (EDS). The rules generated for each clusters do not prove any cluster analysis technique superior over other. However, it is certain that both techniques are well suited to



1 3

Received: 17 January 2016 / Accepted: 19 August 2016 / Published online: 25 August 2016

Evolving Systems (2017) 8:147–155

DOI 10.1007/s12530-016-9165-5

ORIGINAL PAPER

\* Sachin Kumar [sachinagnihotri16@gmail.com](mailto:sachinagnihotri16@gmail.com)

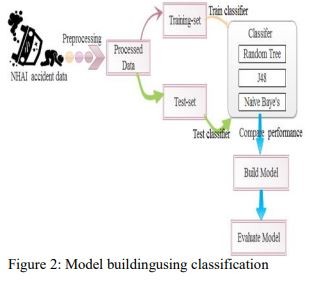
Durga Toshniwal [durgatoshniwal@gmail.com](mailto:al@gmail.com)

Manoranjan Parida [mparida@gmail.com](mailto:mparida@gmail.com)

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

3 Civil Engineering Department (Transportation Group), Indian Institute of Technology Roorkee, Roorkee 247667, Uttarakhand, India



remove heterogeneity of road accident data. The rules gener- ated for each cluster and EDS proves that heterogeneity exists in the entire data set and clustering prior to analysis certainly reduces heterogeneity from the data set and provides better solutions. The rules for Haridwar district reveals some impor- tant information which can used to develop policies to prevent and overcome the accident rate.

**Keywords** Road accident analysis · Heterogeneity · Data mining · FP growth

**1 Introduction**

The increasing number of road and traffic accidents is a challenging issue to the transportation systems. It not only concern with health issues but also associated with eco- nomic burden on the society. Therefore, it is an important task for the safety analysts to carry out a comprehensive study of road accidents to identify the factors that causes an accident to happen, so that preventive actions can be taken to overcome the accident rate and severity of acci- dents consequences. Therefore, a comprehensive study of road accident data is required to identify the several factors associated with road accidents. The main concern with road accident data analysis is to identify the most influential fac- tors affecting road accident frequency and accident sever- ity. The major problem with road accident data analysis is its heterogeneous nature.

Heterogeneity in road accident data is highly undesir- able and unavoidable (Karlaftis and Tarko 1998). The major disadvantage of heterogeneity of road accident data is that certain relationships may remain hidden such as cer- tain accident factors associated with particular vehicle type may not be significant in entire data set; the enormity of

the effect of certain accident related factors may be differ- ent for various conditions; severity levels for an accident contributing factors may be different for different accident types. This heterogeneous nature of road accident data may lead to less accurate results (Kumar and Toshniwal 2015). In order to get more accurate results this heterogeneity of road accident data must be removed. In order to deal with this heterogeneous nature of road accident data, some stud- ies (Ulfarsson and Mannering 2004; Savolainen and Man- nering 2007), divide the data into groups based on some exogenous attributes e.g. accident location, road condition, cause of accident and etc. and analyzed every group sepa- rately to identify several influential factors associated with road accidents in each group. However, this choice is unre- alistic as grouping the data based on certain attributes may results in less important groups (Depaire et al. 2008). Some subgroups can have large number of samples and some can have very low number of samples and thus restricting the application of accident severity models. Other choice is to use some approach such as data mining in order to remove the heterogeneity of the road accident data.

Data mining (Han and Kamber 2001; Kumar and Tosh- niwal 2016), is an evolutionary technique which has been used in the field of transportation systems. Barai (2003) mentioned that data mining has a variety of applications in transportation engineering. Transportation safety is one of the important areas of transportation systems which are actively involved in safety from road and traffic accidents on road. Various research studies on this issue have been done previously using traditional statistical techniques (Joshua and Garber 1990; Chen and Jovanis 2002; Abdel- aty and Radwan 2000). However, traditional statistical tech- niques have their own assumptions regarding dependent and independent data attributes. Any wrong selection of these attributes can lead to erroneous outcomes. Data mining techniques include clustering, classification, association rule mining and anomaly detection. Clustering techniques such as k-means clustering, k-modes clustering and hierarchical clustering are very popular algorithms in several domains.

In road accident data analysis, it is suggested that prior segmentation is very much useful in producing good results. Previously, Ulfarsson and Mannering (2004) and Islam and Mannering (2006), tried to group the data into homogenous subgroups based on some expert knowledge, methodologies. However, Ona et al. (2013), suggests that these factors can segment the data into workable groups but this cannot be guaranteed that the subgroups will comprise of homogeneous group of accidents. Therefore, data min- ing techniques such as cluster analysis have been used to remove the heterogeneity of road accident data. Some stud- ies (Depaire et al. 2008; Ona et al. 2013; Sasidharan et al.

2015), used latent class clustering to reduce the heteroge-

used k-modes clustering to reduce the heterogeneity of the data. The authors claimed that these clustering algorithms are very much useful to reduce the heterogeneity in road accident data and providing good results that can be uti- lized for accident preventive efforts. The next task after prior segmentation of road accident data is to select a rep- resentative variable or target variable in order to perform classification of the data. Most of the road accident data analysis work selects the severity or criticality of the road accident as a target variable for classification.

In this paper, we are presenting a comparative study of the latent class clustering and k-mode clustering technique on a newly available road accident data of Haridwar district of Uttarakhand, India. As both authors claim that their clus- tering technique is highly efficient in dividing the accident data into homogeneous subgroups; this paper attempts to identify which of the above two techniques perform better on our new road accident data set. Further, the association rule mining using Frequent Pattern (FP) growth technique is applied on the subgroups identified by k-modes and latent class clustering techniques to identify several factors associated with road accidents in Haridwar district.

**2 Methodology**

In this section, the paper explains the k-modes and latent class clustering technique for cluster analysis. Further, various cluster selection criteria are discussed followed by association rule mining technique using FP growth algorithm.

**2.1 k‑Modes clustering**

k-Modes clustering (Chaturvedi et al. 2001), approach is an enhanced version of traditional k-means algorithm with an amendment of distance measure, iteration process and cluster center representation. k-Mode clustering is mainly proposed to analyze categorical dataset. The k-modes algo- rithm used a simple matching similarity measure criterion for clustering of categorical data. Let A and B be two quali- tative data objects categorized by x categorical attributes. The simple similarity matching criterion between A and B is the number of matching attribute values of the two data objects. The greater the number of matches is, more the similarity of two objects. Unlike k-means algorithm, k-modes algorithm uses mode instead of means for clus- tering purpose. The k-mode algorithm is quite efficient in handling large categorical data. The distance function of k-modes algorithm can be defined as,

*x*

*d*(*A*, *B*) = δ(*Ai*, *Bi*)

neity of road accident data. (Kumar and Toshniwal 2015),

1 3

148 Evolving Systems (2017) 8:147–155

*i*=1

(1)

where,

1, *If* (*Ai* = *Bi*)

δ(*Ai*, *Bi*) = 0, *If* (*Ai* �= *Bi*)

(2)

used both gap statistics and different information criterion to identify the number of clusters in our data set and also to verify and validate the results off all approaches.

The AIC, BIC and CAIC criterion can be denoted as

Given a set of categorical data objects D defined by n attributes A1, A2, …, An. A mode of D = {D1, D2, …, Dn} is a vector V = {v1, v2, …, v3} that minimize

*n*

AIC = − 2logL + 2p,

BIC = −2logL + p log (n)

CAIC = −2logL + p(log(n) + 1)

(5) (6)

(7)

*d*(*D*, *V*) = *d*(*Di*, *V*)

*i*=1

(3)

where, p is the number of model parameters (Akaike 1987), n is the sample size. The values of AIC, BIC and CAIC are

The iterative process of k-modes algorithm is similar to k-means algorithm.

**2.2 Latent class clustering**

compared across several possible cluster values.

The gap statistic (Tibshirani et al. 2001), is a cluster identification method that can be used with any clustering technique. It can be defined as

LCC technique (Vermunt and Magidson 2002), is a clus- ter analysis techniques widely used technique for the segmentation of road accident data.LCC is a probability

GapnK = E ∗ n{log(Wk)}− log(Wk)

where, Wk is defined by,

(8)

based cluster analysis technique. The objects in each clus- ter formed by LCC are assigned to that cluster based on probability measured wi maximum likelihood technique.

*k*

*Wk* =

*i*=1

1

*Di*

2*ni*

(9)

LCC is different from other clustering techniques as it is available to be used with any type of data variables such as qualitative, quantitative or mixtures of both. LCC does not require any prior standardization that affects the results. Several statistical criteria such as Akaike Information Cri- teria (AIC), Bayesian Information Criteria (BIC) and Con- sistent AIC (CAIC) are available to be used with LCC. The basic form of LC cluster model is given below,

*N*

Di is the sum of pair wise distances for all points in clus- ter i and Wk is the combined sum of squares around the cluster means for each cluster. Gapn K is the difference of expected and observed value of log(Wk). The size of the cluster can be selceted at the value maximizing gap.

**2.4 FP growth technique**

Association rule mining is a popular data mining tech-

*F*(*Yi*|θ ) = π*jFj* *Yi*|θ*j*

*j*=1

(4)

nique that is based on market basket data analysis. Previous

studies (Kumar and Toshniwal 2016; Geurts et al. 2003), used association rule mining using Apriori algorithm for

where, Yi indicates the score of an object on a set of vari- ables under observation. N is the number of clusters to be formed and πj is the prior probability of an object’s mem-

1 3

Evolving Systems (2017) 8:147–155 149

bership to a cluster j. θj is the model parameters of cluster j

and F(YiΙθ) is the mixture probability density.

**2.3 Number of cluster selection**

Cluster analysis is a process of segmenting the data set into homogeneous groups of clusters. The primary requirements for any cluster analysis task to find the number of clusters to form. Various approaches are exists in literature to iden- tify the number of clusters e.g. gap statistic and different information criteria such as AIC, BIC and CAIC (Akaike

1987; Raftery 1986; Fraley and Rafetery 1998). We have

road accident data analysis. The major problem with Apri- ori algorithm is that it uses candidate item set generation and then tests whether these item sets are frequent or not. Hence, Apriori algorithm is computationally expensive as it reuires multiple database scans in order to generate can- didate sets. The another association rule mining technique is FP growth algorithm which is proposed in 2000 (Han et al. 2000). The difference between FP growth and Apriori is that it is computationally faster than Apriori as it does not require candidate generation. FP growth algorithm uses a special data structure known as FP tree, which preserve the itemset association information. We used support, con- fidence and lift interesting measures to extract strong rules from data set. Once FP tree construction is done, the work- ing of FP growth algorithm is given as follows:

**Algorithm: FP-Growth**

*Input*: A data set D, denoted by FP-tree constructed and threshold value for support

*Output*: The set of frequent patterns

*Method*: call FP-Growth(FP-tree, [ ] ) // [ ] represents NULL Procedure FP-Growth(Tree, a)

{

If Tree contains a single prefix path then // Mining single prefix-path FP-tree

{

} Else

Let P be the single prefix-path part of Tree;

Let Q be the multipath part with the top branching node replaced by a null root; For each combination (denoted as A) of the nodes in the path P do

Generate pattern A∪ a with support = minimum support of nodes in A;

Let frequent-pattern-set(P) be the set of patterns so generated;

Let Q be Tree;

For each item Xi in Q do // Mining multipath FP-tree

{

Generate pattern A = Xi∪X with support = Xi .support;

Construct A’s conditional pattern-base and then A’s conditional FP-tree Tree A;

If Tree A ≠ Ø then

Call FP-Growth(Tree A,A);

Let frequent-pattern-set(Q) be the set of patterns so generated;

}

Return (frequent-pattern-set(P) ∪ frequent-pattern-set(Q) ∪ (frequent-pattern-set(P) ×

frequent-pattern-set (Q)))

}

**2.5 Data set**

150 Evolving Systems (2017) 8:147–155

The data set used for this study is provided by GVK-EMRI, Dehradun (EMRI 2015). This data set covers data from entire state of Uttarakhand from January 2009 to December

2014. In this paper, we have selected 4570 road accident records with 11 accident attributes of Haridwar district for analysis. The brief description of road accidents are given in Table 1:

**3 Results and discussion**

**3.1 Cluster analysis**

The primary task of cluster analysis is to determine the number of cluster that can be formed in the data set. We generated ten different models for cluster 1 to cluster 10 using both the gap statistic and AIC, BIC and CAIC crite- ria. The Figs. 1 and 2 illustrates the results of Gap statistic and AIC, BIC and CAIC criteria.

Figure 1 illustrates the representation of AIC, BIC

and CAIC criteria for ten cluster models. A drop in the

1 3

values of AIC, BIC and CAIC are found with an increase in the number of cluster models. Based on some previ- ous studies (Depaire et al. 2008; Ona et al. 2013), we have confirmed the model with four clusters. In order to validate this, another cluster selection criteria i.e. gap statistic has been used. Figure 2a shows elbow plot and Fig. 2b and c illustrates the cluster selection results using Gap statistic. The Fig. 2a illustrates that there is a knee curve at model with cluster 4. Figure 2b and c indicates that gap value at cluster model 4 maximizes the value of Gapn K. Based on these results, we have selected the model with four numbers of clusters as all cluster selection criteria are agreed with one model with four clusters.

Both k-modes and LCC has been used to form homoge- neous segments from the road accident data.

The clusters identified and their size is given in Table 2. We formed all four clusters using both LCC and

k-modes clustering algorithm. The number of data instances assigned to their respective clusters was differ- ent for LCC and k-modes which has shown in Table 2. The cluster 1 consists of two wheeler accidents that were hap- pened at intersection on both highway and local roads. The

**Table 1** Road accident data attributes

S. no. Attribute Attribute values Code Critical Non-critical Total

1 Number of injury: NOI 1 injury 1 258 1442 1700

2 injuries 2 348 1247 1595

>2 injuries +2 520 755 1275

2 Age of victim: AOV <18 years CHD 230 450 680

18–30 years YNG 324 1318 1642

30–50 years ADU 297 1003 1300

>50 years SNR 275 673 948

3 Gender: GEN Male M 758 2698 3456

Female F 368 746 1114

4 Time of day: TOD 0–3 T1 59 39 98

3–6 T2 102 163 265

6–9 T3 195 463 658

9–12 T4 68 653 721

12–15 T5 82 616 698

15–18 T6 97 544 641

18–21 T7 254 733 987

21–24 T8 269 233 502

5 Month: MON Winter WNT 290 835 1125

Spring SPR 284 701 985

Summer SUM 297 1303 1600

Autumn ATM 255 605 860

6 Lighting condition: LIG Day light DLT 198 1621 1819

Dusk DUS 289 708 997

Road light RLT 246 843 1089

No light NLT 393 272 665

7 Roadway feature: ROF Intersection INT 469 1450 1919

Slope SLP 65 133 198

Curve CUR 374 884 1258

Unknown UNK 218 977 1195

8 Road type: ROT Highway HIW 871 2644 3515

Local LOC 255 800 1055

9 Accident severity: ASV Critical CR 1126 0 1126

Non-critical NC 0 3444 3444

10 Surrounding area: SUA Agriculture land AGL 298 1028 1326

Market MAR 215 580 795

Colony COL 196 449 645

Forest FOR 252 1028 1280

Hill area HIL 155 213 368

Hospital HOS 10 146 156

11 Type of accident: TOA Two wheeler TWH 401 1914 2315

Three wheeler THW 29 96 125

Vehicle fall height VFH 198 187 385

Vehicle rollover VRO 139 251 390

Pedestrian hit PDH 152 698 850

Fixed object/divider hit FOH 207 298 505

second cluster 2 contains the two wheeler accidents which were happened at different road features other than inter- section on local and highway roads. Cluster 3 represents

1 3

Evolving Systems (2017) 8:147–155 151

the road accidents that occurred in forest and hilly areas. Cluster 4 represents road accidents at other surrounding areas.

**140000**

**135000**

**130000**

**125000**

**120000**

**115000**

AIC BIC CAIC

than Apriori algorithm. A support of 20 % has been chosen to generate strong rules.

The rules generated of each cluster and EDS using FP growth algorithm were analyzed and it was found that rules for clusters obtained from LCC and k-modes was though similar with different interest values. Some relevant rules for EDS and C1–C4 are given in Table 3.

**110000**

1 3

152 Evolving Systems (2017) 8:147–155

**0 2 4 6 8 10 12**

**Number of cluster models**

**Fig. 1** Representation of AIC, BIC and CAIC for different cluster models

**a**

Within group sum of squares Wk×105

**3.5**

**3**

**2.5**

**2**

**1.5**

**1**

**0.5**

**0**

**0 2 4 6 8 10 12**

Number of cluster models

**b**

**Observed and Expected log(Wk)**

log(Wk) Elog(Wk)

**6**

**5**

**4**

**3**

**2**

**1**

**0**

**0 2 4 6 8 10 12**

Number of cluster models

**c**

**0.9**

**0.8**

**0.7**

**0.6**

**0.5**

**Gap**

**0.4**

**0.3**

**0.2**

**0.1**

**0**

**0 2 4 6 8 10 12**

**Number of cluster models**

**Fig. 2 a** Representation of Wk for different cluster models (elbow plot), **b** observed and expected log(Wk) against different number of clusters, **c** gap statistic curve

**3.2 Rule mining**

Association rule mining is a popular technique that is used to identify the correlation between values of different attrib- utes for a data set. FP growth algorithm is an association rule mining technique which is computationally efficient

As mentioned in Table 3 that C1 consists of two wheeler road accidents that mainly occurred at intersections on road network of Haridwar district. The rules generated for these clusters exposed different other factors that are associated with road accident in this cluster. Some relevant and strong rules are given in Table 3. A rule in C1 i.e. “If {TOD = T7

AND ROT = HIW} Then {ASV = CR}” simply states that two wheeler road accidents at intersection on high- ways between 6:00 PM and 9:00 PM was critical accidents which involves severe injury or fatal injury. Another rule “If

{SUA = COL AND NOV = +2} Then {AOV = CHD}” tells that a two wheeler accidents that occurred at intersec- tions in colony and which involves more than two victims injured then the age of the victims were between 0 and

14 years.

Another cluster C2 contains two wheeler accidents which happened on different road sections other than intersections on roads either highway or local. In this clus- ter, a rule “If {AOV = YNG AND SUA = HILL AND ROF = CUR} Then {ASV = CR}” states that if a two wheeler accident occurs in at curve on hilly roads and where age of victim was between 18 and 30 years then that accident was critical. Another rule “If {SUA = FOR AND TOD = T8} Then {ASV = CR}” states that a two wheeler accident in night time or between 9:00 PM and 11:00 PM in forest area was critical accidents.

Cluster C3 represents all road accidents occurred at for- est or hill regions except two wheeler accidents. In C3, rule “If {ROT = HIG AND LIG = DUS AND NOV = +2} Then {TOA = VFH}” indicates that road accidents in this clusters that occurred in dusk lightning conditions on highway road with more than two injured victims were mostly vehicle fall from height accidents. Also, “If

{ROT = LOC AND TOD = T9} Then {ROF = UNK AND TOA = VRO}” indicates that in night duration between 9:00 PM and 12:00 PM, if an accidents occurred on local roads and if road feature is unknown then the acci- dents would be of vehicle roll over type. Other rule in C3 shown in Table 3 indicates that vehicle fall height acci- dents and fixed object hit accidents that occurred during

9:00 PM–3:00 AM were critical accidents with severe or fatal injuries.

Cluster C4 represents all other road accidents except two wheeler accidents which have occurred in different regions such as roads near markets, colonies, hospitals, agriculture land and etc. but not in hilly or forest areas.

**Table 2** Cluster description and number of data instances in each cluster

Cluster ID Cluster description Number of data instances using

k-modes LCC

C1 Two wheeler accidents at intersections on local road and highways 1285 1218

C2 Two wheeler accidents at other road sections on local road and highways 1030 1097

C3 Other road accidents that occurred in forest area and Hill 1024 1135

C4 Other road accidents that occurred in other surrounding area 1231 1120

**Table 3** Association rules relevant to road accidents for each cluster and EDS

Group Rule k-modes cluster LCC cluster

Confidence Lift Confidence Lift

C1 If {TOD = T8 AND ROT = HIW} Then {ASV = CR} 0.80 3.25 0.87 3.85

If {LIG = RLT AND SUA = MAR} Then {ASV = NC} 0.75 4.20 0.84 4.35

If {LIG = DLT AND SUA = COL} Then {ASV = NC} 0.74 3.45 0.75 3.65

If {SUA = COL AND NOV = +2} Then {AOV = CHD} 0.74 2.86 0.76 1.95

C2 If {AOV = YNG AND SUA = HILL AND ROF = CUR} Then {ASV = CR} 0.89 3.15 0.82 3.65

If {SUA = FOR AND TOD = T8} Then {ASV = CR} 0.79 3.25 0.81 3.48

If {ROT = LOC AND SUA = AGL AND GEN = F} Then {ASV = NC} 0.74 4.25 0.68 4.36

If {ROT = LOC AND ROF = UNK AND SUA = AGL} Then {ASV = CR} 0.68 2.24 0.78 2.56

C3 If {ROT = HIG AND LIG = DUS AND NOV = +2} Then {TOA = VFH} 0.74 4.15 0.84 3.90

If {ROT = LOC AND TOD = T9} Then {ROF = UNK AND TOA = VRO} 0.73 3.16 0.75 3.25

If {ROT = HIG AND TOD = T1 AND TOA = FOH} Then {ASV = CR} 0.71 3.87 0.78 3.95

If {TOA = VFH AND TOD = (T1 OR T8)} Then {ASV = CR} 0.65 2.36 0.63 2.44

C4 If {ROF = INT AND SUA = MAR} Then {TOA = PDH} 0.77 3.68 0.78 3.78

If {SUA = MAR AND NOV = +2} Then {TOA = THW} 0.66 3.74 0.62 3.45

If{TOA = PDH AND TOD = T9} Then {ASV = CR} 0.65 2.89 0.60 3.65

If {ROT = HIG AND TOA = VRO AND NOV = +2} Then {ASV = CR} 0.49 2.90 0.55 2.82

Confidence Lift

EDS If {NOV = 1 AND ROF = INT} Then {TOA = TWH} 0.85 4.56

If {SUA = AGL AND LIG = DLT} Then {TOA = TWH AND ASV = NC} 0.65 4.34

If {ROF = INT AND SUA = MAR} Then {TOA = TWH} 0.65 4.13

If {ROT = HIG AND SUA = FOR AND LIG = DLT} Then {ASV = CR} 0.62 3.58

If {TOA = TWH AND GEN = M} Then {AGE = YNG} 0.61 4.26

If {SUA = MAR AND ROT = HIG} Then {TOA = PDH} 0.44 1.12

If {ROT = HIW AND ROF = INT} Then {ASV = NC} 0.74 3.45

In this cluster, rule “If {ROF = INT AND SUA = MAR} Then {TOA = PDH}” indicates that intersections in mar- ket areas are more prone to road accident for pedestrians. Another rule “If {SUA = MAR AND NOV = +2} Then

1 3

Evolving Systems (2017) 8:147–155 153

{TOA = THW}” indicates that if an accident occurred in market and more than two victims are injured then the accident type will be three wheeler accidents in most cases. Rule “If{TOA = PDH AND TOD = T9} Then

{ASV = CR}” indicates that pedestrian hit accidents dur- ing 9:00 PM–12:00 AM was mostly critical accidents with

severe or fatal injury. Another rule “If {ROT = HIG AND TOA = VRO AND NOV = +2} Then {ASV = CR}” sim- ply indicates that vehicle rollover accidents on highways with more than two injuries were mostly critical.

The major differences found in rules within clusters and

EDS are discussed below:

• In EDS, the majority of rules represent the road acci- dents with two wheeler accidents. The reason can be the high number of two wheeler road accidents in com-

parison to other type of accidents. The other type of accident that is covered by association rules in EDS is pedestrian hit accidents. Other type of road accidents are not exposed by the association rules as they do not satisfy the minimum threshold level. But after prior seg- mentation of the data, all types of accidents are covered by association rules in C1–C4.

154 Evolving Systems (2017) 8:147–155

• Road accidents at hilly regions are not covered in EDS as the number of accidents at hilly regions is compara- tively low, but some rules are generated covering road accidents at hilly regions in some clusters.

• Various rules in EDS shows a general nature, for example a rule “If {ROT = HIW AND ROF = INT} Then {ASV = NC}” in EDS shows that road accident occurred at intersection on highway were non-critical. This information is fair but incomplete. Although this rule has high confidence and lift value that makes this rule very strong, we cannot lead to a conclusion that all type of accidents at intersection on highways were not critical. Instead, association rules for every cluster are a representative to its cluster. So one can be assured that a rule in a cluster makes sense to its cluster. For exam- ple, the same rule if it was in C1, we can conclude that two wheeler accidents at intersection on highways were not critical. This statement makes some sense. Hence, segmenting the data set using some cluster analysis technique makes the analysis more precise and easy to understand.

These results are quite similar to the results in a previ- ous study of Dehradun district which is adjacent to district Haridwar. This strengthens the concept that road accident data of any region can certainly help in understanding the factors associated with road accidents in nearby regions also. Hence, we have reached to a conclusion that the pre- vious approaches (Depaire et al.2008; Ona et al. 2013; Sasidharan et al. 2015), that were suggesting about prior segmentation of accident data before its analysis is very important and it can help in extracting important informa- tion associated with road accidents which can remain hid- den otherwise.

**3.3 Comparison of k‑modes and LCC over cluster analysis**

In this study, authors attempted to use both LCC and k-modes clustering technique to segment the real world road accident data. The clusters formed by both the tech- nique have different number of data instances as given in Table 2. LCC uses maximum likelihood technique to meas- ure the probability of data objects before assigning them to any subgroup whereas k-modes technique uses distance

1 3

metric defined by Eqs. (1) and (2). There are various asso- ciation rules generated using FP growth algorithm for all clusters formed by LCC and k-modes clustering technique which are although similar but they have different interest- ing measures. For example, a rule “If {SUA = COL AND NOV = +2} Then {AOV = CHD}” in cluster C1 has higher confidence (0.76) but lower lift value (1.95) for clus- ter formed by LCC.

So for a real world accident data set of nominal attrib- utes, one cannot conclude which technique is superior over other. As both techniques are previously used and provided good solution to overcome the heterogeneity in road acci- dent data. We can conclude that LCC and k-modes cluster- ing technique are both available to deal with road accident data with categorical attributes and these techniques pro- vide good solutions in finding associated factors that are associated with road accidents.

**4 Conclusion and suggestion**

This paper presents a comparative study of k-modes clus- tering and LCC on a new road accident data set from Haridwar district of Uttarakhand, India. This study uses

4570 road accident records of Haridwar district obtained from emergency transport service controlled by GVK- EMRI, Dehradun. The number of attributes that has been used in the analysis was 11 which were associated with road accidents. The information criteria (AIC, BIC and CAIC) and gap statistic are used to identify the number of clusters to be made. Based on the results obtained from cluster selection criteria four clusters C1–C4 were identi- fied by k-modes and LCC. The clusters identified by both the techniques have different number of road accidents in each cluster. Further, FP growth technique is applied to each cluster and EDS to generate association rules which can define the correlation between the values of different attributes in the data. There is no major difference found in the association rules generated by FP growth algo- rithm except that the rules have different confidence and lift value for the clusters formed by k-modes and LCC. Although Chaturvedi et al. (2001), provided that k-modes are better than LCC on categorical data, we do not found any differences that shows that k-modes are better than LCC especially in road accident data except computa- tional speed. There is no doubt that both the cluster analy- sis technique performs well in reducing the heterogeneity of road accident data. Also the association rules gener- ated is providing information about various types of road accidents and their associated factors. Also, these results are quite similar to Dehradun district which is adjacent to Haridwar district. It reveals that certain factors that affects

Evolving Systems (2017) 8:147–155 155

the road accident in one district also affects the road acci- dents in adjacent districts. This can be very useful to over- come the accident rates in Haridwar and surrounding dis- trict of Uttarakhand.

**Acknowledgments** We are thankful to GVK-Emergency Manage- ment Research Institute Dehradun, Uttarakhand to provide data for our research.

**Compliance with ethical standards**

**Conflict of interest** None of the authors have any competing interest.

**References**

Abdel-Aty MA, Radwan AE (2000) Modeling traffic accident occur- rence and involvement. Accid Anal Prev 32(5):633-642

Akaike H (1987) Factor analysis and AIC. Psychome 52:317–332

Barai S (2003) Data mining application in transportation engineering.

Transport 18:216–223

Chaturvedi A, Green P, Carroll J (2001) k-Modes clustering. J Classif

18:35–55

Chen W, Jovanis P (2002) Method of identifying factors contributing to driver-injury severity in traffic crashes. Transp Res Rec 1717

Depaire B, Wets G, Vanhoof K (2008) Traffic accident segmen- tation by means oflatent class clustering. Accid Anal Prev

40(4):1257–1266

Fraley C, Raftery AE (1998) How many clusters? Which clustering method? Answers via model-based cluster analysis. Comput J

41:578–588

Geurts K, Wets G, Brijs T, Vanhoof K (2003) Profiling of high fre- quency accident locations by use of association rules. Transp Res Rec 1840

Han J, Kamber M (2001) Data mining: concepts and techniques. Mor- gan Kaufmann, New York

Han J, Pei H, Yin Y (2000) Mining frequent patterns without candi- date generation. In: Proceedings of the conference on the man- agement of data (SIGMOD’00, Dallas, TX). ACM Press, New York

[http://www.emri.in Accessed 15 Dec 2015](http://www.emri.in)

Islam S, Mannering F (2006) Driver aging and its effect on male and female single-vehicle accident injuries: some additional evi- dence. Accid Anal Prev 37(2):267–276

Joshua SC, Garber NJ (1990) Estimating truck accident rate and involvements using linear and poisson regression models. Trans Plan Tech 15(1):41–58

Karlaftis M, Tarko A (1998) Heterogeneity considerations in accident modeling. Accid Anal Prev 30(4):425–433

Kumar S, Toshniwal D (2015) A data mining framework to ana- lyze road accident data. J Big Data 2(1):1–18. doi:[10.1186/ s40537-015-0035-y](http://dx.doi.org/10.1186/s40537-015-0035-y)

Kumar S, Toshniwal D (2016) A data mining approach to characterize road accident locations. J Mod Transp 24(1):62–72

Ona JD, Lopez G, Mujalli R, Calvo FJ (2013) Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks. Accid Anal Prev 51:1-10

Raftery AE (1986) A note on Bayes factors for log-linear contin- gency table models with vague prior information. J R Stat Soc B 48:249–250

Sasidharan L, Wu KF, Menendez M (2015) Exploring the application of latent class cluster analysis for investigating pedestrian crash injury severities in Switzerland. Accid Anal Prev 85:219–228

Savolainen P, Mannering F (2007) Probabilistic models of motor- cyclists’ injury severities in single- and multi-vehicle crashes. Accid Anal Prev 39(5):955–963

Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc B

63:411–423

Ulfarsson GF, Mannering FL (2004) Difference in male and female injury severities in sport-utility vehicle, minivan, pickup and pas- senger car accidents. Accid Anal Prev 36(2):135–147

Vermunt JK, Magidson J (2002) Latent class cluster analysis. In: Hagenaars JA, McCutcheon AL (eds) Advances in latent class analysis. Cambridge University Press, Cambridge

1 3