# *Analyzing Road Accident Data using Machine Learning Paradigms*

**Priyanka A. Nandurge**
*Department of Computer Science and Engineering*
*Rajarambapu Institute of Technology,*
*Islampur India*

priyankanandurge@gmail.com

**Nagaraj V. Dharwadkar**
*Department of Computer Science and Engineering*
*Rajarambapu Institute of Technology,*
*Islampur India*

nagaraj.dharwadkar@ritindia.edu

*Abstract*- **To determine the main factors associated with road traffic accidents is one of main objectives of accident data analysis. Due heterogeneity nature of road accident data makes analysis tricky. To overcome heterogeneity of data partitioning is used. The proposed method uses k-means clustering method as the main task of segmentation of road accident data. Further, association rule mining is applied to discover the situations related with the occurrence of the whole data set and the occurrence of clusters recognized by the k-means clustering algorithm. The combined result of k-means clustering and association rule produces major information.**

***Index Terms- Data mining, Road Accident data analysis, k-means clustering, Association rules mining.***

## I. INTRODUCTION

Road accident data is the basic measure of safety without which the scale and nature of road safety problems cannot be established with certainty. Therefore, the existence of a reliable accident database is a key factor in the management of road safety. The country's accident data collection system is still inconsistent and irregular because there is neither a uniform data collection format nor a robust system of reliable and meaningful retrieval of regular and systematic data.

Accurate and comprehensive accident records are the basis for accident analysis. The effective use of accident records depends on three factors, namely, the accuracy of the data, record retention and data analysis. The need for a high standard of incident reporting is a major prerequisite for the use of accident records to develop road safety measures. If the original incident report itself is poor, then the analysis and use of the results will be poor. Inaccurate and incomplete accident data make the results fuzzy, misleading, and not very fruitful.

Road accidents are unsure and irregular events and their analysis needs to be aware of the factors that affect them. Road accidents are defined by a set of attributes that are often different. The main difficulty of accident data analysis is its heterogeneity. Therefore, heterogeneity must be measured through the analysis of the data, or else a number of relationships among the data may stay hidden. Segmentation is used to reduce data heterogeneity using a number of measures such as expert knowledge, but there is no guarantee that this will result in the best segmentation of the group including road accidents. Cluster analysis can helps to segment road accidents data.

## II. RELATED WORK

Sachin Kumar et al. [1], used data mining techniques to identify locations where high frequency accidents are occurred and then analyze them to identify the factors that have an effect on road accidents at that locations. The first task is to divide the accident location into k groups using the k-means clustering algorithm based on road accident frequency counts. Then, association rule mining algorithm applied in order to find out the relationship between distinct attributes which are in accident data set and according to that know the characteristics of locations.

S. Shanthi et al. [2] proposed data mining classification technology based on gender classification, in which RndTree and C4.S use AdaBoost Meta classifier to provide high-precision results. From the Critical Analysis Reporting Environment (CARE) system provided by the Fatal Analysis Reporting System (FARS) used by the training data set.

Tessa K. Anderson et al. [3] proposed a method of identifying high-density accident hotspots, which creates a clustering technique that determines that stochastic indices are more likely to exist in some clusters, and can therefore be compared in time and space. The kernel density estimation tool enables the visualization and manipulation of density-based events as a whole, which in turn is used to create the basic spatial unit of the hotspot clustering method.

Miao Chong et al. [4] The severity of damage occurring during a traffic accident is replicated using the performance of various machine learning paradigms, such as neural networks trained using hybrid learning methods, support vector machines, decision trees, and concurrent mixed models involving decision trees and neural networks. The experimental results show that the hybrid

decision tree neural network method is better than the single method in machine learning paradigms.

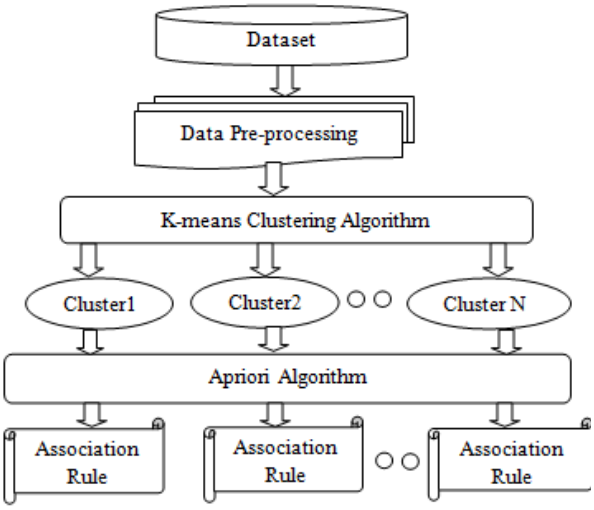
Tibebe Beshah et al.[5] An adaptive regression tree was developed at the Addis Ababa Urban Transport Office to establish a decision support system for the analysis of road traffic accidents. The study focused on the severity of the injury resulting from the use of actual data obtained from the Addis Ababa Department of Transportation. Empirical results show that the developed model can reasonably and accurately classify accidents.

Tibebe Beshah et al. [6] main goal is exploring and predicting the impact of road users on the potential risk of injury to machine learning. Classification and adaptive regression trees (CART) and stochastic forest methods. In order to determine the relevant models and illustrate the performance of road safety technology, road accident data collected from the Addis Ababa Transport Bureau were analyzed in many ways. Experimental results show that the model can classify accidents as promising accuracy.

Sachin Kumar et al.[7], A framework for the preliminary allocation of 11,574 road traffic accidents on the Dehradun road system was proposed in 2009-2014 (including two) using K-mode clustering as a preliminary task. In addition, a association rule mining is used to discover the situations related with the occurrence of the Entire data set and the occurrence of clusters recognized by the k-modes clustering algorithm And then compare cluster analysis with the analysis of the entire data set.

## III. METHODOLOGY

### A. *Proposed Model*




### B. *K-means clustering*

**Fig 1: Flow diagram of Analysis of Road accident data.**

Data mining techniques are categorized as supervised and unsupervised. Clustering is one of the unsupervised data mining techniques. Grouping the data objects into clusters is the important task of clustering. So that objects which present in same group are more similar than the objects in other. Mostly k-means clustering algorithm is used for numerical data. K-mean clustering groups the data items into k clusters. The choice of the appropriate clustering algorithm depends on the type and nature of the data.

*K-Means algorithm*

Input: {D, k} // D- Dataset consists of n data objects, k - Number of clusters.
Output: k clusters
Algorithm steps:
1: K objects are randomly selected from D as the initial cluster center.
2: Each data object is assigned to its closest cluster.
3: Update cluster mean i.e. calculate the mean of the objects in every cluster.
4: Repeat until no data object changes its cluster membership or meets any other convergence criteria [7].

### C. *Estimating number of clusters*

The main difficulty of clustering algorithm is to estimate the number of clusters. In k-means clustering, value of k must be given by the user which is one of the limitations of this algorithm. If the value of k is incorrect then it may lead to incorrect clustering results. The solution to this problem is to use gap statistics [11] to obtain the optimal value of k.

The gap statistics [11] compare the sums of intra-cluster variations with different k values to their expected values at zero-reference distribution of the data. The sum within the group variation within a given k-cluster is the sum within the sum of the sums of squares ($W_k$). For the observed data and the reference data, the total intra-group dissimilarities were calculated using different k values. The gap statistics for a given k is given as follows:

$$\mathrm{Gap}_n(K) = E_n^*\{\log(W_K)\} - \log(W_K) \quad (1)$$

Where $E_n^*$ denotes the expected value of the sample of size n from the reference distribution. The estimation

of the optimal clustering $\widehat{k}$ will be to maximize the value of $\mathrm{Gap_n(K)}$. This means that the clustering structure is distributed evenly away from the points.

The standard deviation $(Sd_k)$ of $\log(W_k^*)$ is also computed in order to describe the standard error ( $s_k$ ) of the recreation as follow:

$$s_k = Sd_k \times \sqrt{1 + 1/B} \quad (2)$$

Finally, the most favorable number of clusters K is chosen as the minimum k such that:

$$\text{Gap (k)} \geq \text{Gap (k+1)} - s_{k+1} \quad (3)$$

The minimum value of k is selected in such a way that the gap statistics are within one standard deviation of the gap at k + 1 [11].

### D. Association rule

Association rule mining is often used to create set of rules that define the fundamental patterns in a data set. The correlation of the two attributes of an accident data is determined by the frequency at which they occur in the data set. Rule A → B shows A occurs first, then B also occurs.

Pseudo code for Apriori algorithm
$L_k$ = {Frequent item set with size k}
$C_k$ = {Candidate item set with size k}
$L_1$ = {Frequent 1 item sets}
K=1;
While ($L_k$-1≠ϕ) then
$C_{k+1}$= candidates generated from $L_k$
For each transaction t ϵ D do
Increment the count of candidate in $C_{k+1}$ that also contained in t
$L_k$+1= candidate in Ck+1 with minimum support
K=K+1;
Return $U_k L_k$. [7]

In association rule mining, to evaluate the quality of rules various interesting measures are used. These interesting measures for association rule A → B are as follows [1]:

#### 1) Support

The support of rules A → B defines the frequency at which A and B appear in each other in the data set, and the Eq. (4) can be used. Support is also known as frequency constraints. A frequent item set is defined as a set of items that satisfy some of the support thresholds. Frequent item sets are also used to generate association rules based on other metrics.

$$\text{Support} = P(A \cap B) \quad (4)$$

#### 2) Confidence

The fraction of the appearance of A and B against A is the confidence of the rule A → B and can be used as an equation. (5). If the confidence value is higher than the rule A → B, the probability of occurrence of B is higher. Sometimes, only the value of confidence is insufficient to evaluate the descriptive benefits of the rule.

$$\text{Confidence} = P(A \mid B) = \frac{P(A \cap B)}{P(A)} \quad (5)$$

#### 3) Lift

The occurrence rates of A and B as expected is the Lift of rule A → B. In other words, Lift is ratio of confidence of rule and expected confidence. Expected confidence is the appearance of A and B and the appearance of B. The value of boost range from 0 to ∞. Increasing the value greater than 1 makes the rule helpful for predicting the consequences in a future dataset A and B. Elevation is measured co-occurrence only, and as well as is symmetrical with respect to A and B. The lift can calculated with Eq. (6).

$$\text{Lift} = \frac{P(A \cap B)}{(P(A)\,P(B))} \quad (6)$$

#### 4) Leverage

The leverage measurements for rules A → B appear in the dataset together for the difference between A and B, and if A and B are statistically dependent on expectations. It is possible to use Eq.(6). Leverage values range from [-0.25 to +0.25]. If the value of leverage is 0 then it indicates that the variables are statistically independent. If the variables occur more frequently together, it will increase to +1, and if a single variable appears more frequently, it will decrease toward -1.

$$\text{Leverage} = P(A \cap B) - P(A)\,P(B) \quad (7)$$

#### 5) Conviction

Conviction is an additional measure that takes up some confidence and lift. Rule A → B compares the probability of occurrence of A with the probability of not having B if they are related to the true frequency of the occurrence of A without B. The conclusion is asymmetric, that is conviction (A → B) ≠ conviction (B → A). The conviction value is in the range [0.5, ∞]. A value far from 1 represents an interesting rule. Beliefs, preconditions and consequential support are all taken into account. Conviction calculated using Eq. (7)

$$\text{Conviction} = \frac{P(A) \times P(B)}{P(A \cap \bar{B})} \quad (8)$$

## E. RESULT AND DISCUSSION

### A. Cluster analysis

To find out the number of cluster to be produced by clustering algorithm is the basic necessity of cluster analysis. Once getting the number of clusters after that use k-means clustering algorithm using R statistical software to segment the accident data sets. After the proper partitioning of the data set is obtained, the subsequently task is the classification of each cluster. An efficient analysis of each and every cluster gives the cluster-based accident variables. The cluster with brief description is given below:

Cluster 1 (C1)
In this cluster consists of accidents distributed on slight curve near or inside village, near a factory of industrial area, and near recreation place or cinema across non-highway roads(other road). Those accidents occurred on curve only has only one injury.

Cluster 2 (C2)
In this cluster consists of accidents distributed on straight road, near a religious place, near a office complex, and Residential area across non-highway roads. In this cluster 53% of accidents involved only one injury and 20% of accidents involved more than one injury.

Cluster 3 (C3)
In this cluster accidents distributed on straight road in which 50% of accident involved two injuries. Those accidents occurred near a school or colleges and near office complex.

Cluster 4 (C4)
In this cluster consists 60% of accidents occurred across non highway roads and 25 % of accident across state highway with two injuries, near school or college and in bazaar.

Cluster 5 (C5)
In this cluster consists of accidents distributed on straight road and slight curve. In that 56% of accidents consist of only one injury occurred at pedestrian crossing.

Table1: Attributes firm road accident data

| **Name of Attribute** | **Type** | **Attribute values** |
|---|---|---|
| Time | Nominal | 1,2,3,……24 |
| Day | Nominal | Sun, Mon, Tue, Wed, Thurs, Fri, Sat |
| Month | Nominal | 1,2,3,……...12 |
| Classification | Nominal | Fatal, Serious, Minor |
| of accident | | |
| Drivers Age | Nominal | Below 18, 18-21, 21-30, 30-40, 40-50, 50-60, 60-70, & 70 above |
| Drivers Gender | Binary | Male/Female |
| Number of Injured | Nominal | 0, 1, 2, 3,4 …… |
| Light Condition | Nominal | Daylight, Twilight, |
| Weather Condition | Nominal | 1,2,3…….12 |
| Hit and Run | Binary | Yes/No |
| Location of Area | Binary | Urban/Rural |
| Type of area | Nominal | 1,2,3……..,13 |
| Road Type | Nominal | National/State Highway, Other road |
| Road Feature | Nominal | Straight curve, Slight curve, Sharp curve |

*B. Number of cluster selection using gap statistic*

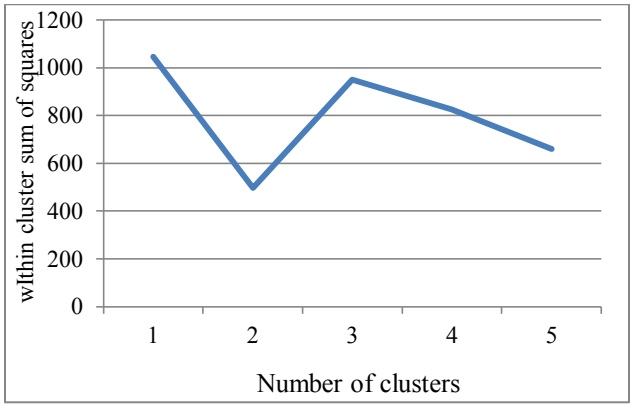



Fig:2(a) Within cluster sum squares function $W_k$ verses number of clusters

Fig 2(a) shows the relation between the error measure Wk (within the cluster discretization) of the k-means clustering algorithm and the number of clusters k used. The error measure Wk directly proportional to the number of clusters, but decreases from some k forwards.


Fig:2(b) observed and Expected log($W_k$)

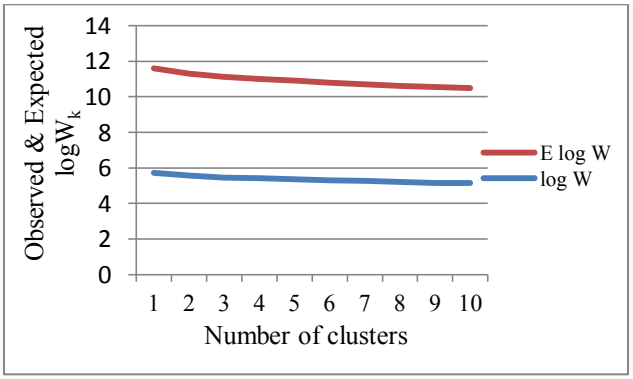



Fig: 2(c) Gap statistic curve

*C. Association rule mining*

Using weka tool Apriori algorithm has applied on every cluster to generate association rules. The minimum support value is 30% to create association rule for every cluster and whole data set. Based on the confidence, lift, leverage and conviction measures these rules are evaluated. The strong rules which considered for analysis that are having the high lift value. The strong rules for every cluster and whole Data Set have shown in following table:

Association rule for cluster1 defines, accidents distributed on slight curve near or inside village, near a factory of industrial area, and near recreation place or cinema across non-highway roads. Those accidents occurred on curve only has only one injury. Rules revealed that most accidents are occurred at Daylight and most are hit and run accidents.

Association rule for cluster2 defines, accidents distributed on straight road, near a religious place, near a office complex, and Residential area across non-highway roads. In this cluster 53% of accidents involved only one injury and 20% of accidents involved more than one injury. Rules revealed that most accidents are occurred in rural area with very hot weather condition.

Association rule for cluster3 defines, accidents distributed on straight road in which 50% of accident involved two injuries. Those accidents occurred near a school or colleges and near office complex. Rules revealed that most accidents are occurred in urban area having age group 21-30.

Association rule for cluster4 defines, 60% of accidents occurred across non highway roads and 25 % of accident across state highway with two injuries, near school or college and in bazaar. Rules revealed that most accident occurred are Daylight condition but not in hit and run accident category.

Association rule for cluster5 defines, accidents distributed on straight road and slight curve. In that 56% of accidents consist of only one injury occurred at pedestrian crossing. Rules revealed that, most accidents are fatal accidents.

Table 2:Association rules

| Rule No. | Rules | Confidence | Lift | Lev. | Conv. |
|---|---|---|---|---|---|
| **C1** | | | | | |
| **1** | {Male, other road, hit & run= Yes, Daylight}→ {injured=1, fine} | 1 | 2.17 | 0.17 | 4.31 |
| **2** | {injured=1,fine}→{Male, other road, hit & run=Yes, Daylight} | 0.67 | 2.17 | 0.17 | 1.66 |
| **3** | {Male, other road, Daylight}→ {injured=1, hit & run=Yes, fine} | 0.89 | 2.10 | 0.16 | 2.60 |
| **4** | {injured=1, hit & run = Yes, Fine} → {Male, other road, Daylight} | 0.73 | 2.10 | 0.16 | 1.8 |
| **5** | {Male, other road, hit & run = Yes} → {injured=1,Fine} | 0.89 | 1.93 | 0.15 | 2.42 |
| **6** | {injured=1, Fine}→ {Male, other road, hit & run= Yes} | 0.67 | 1.93 | 0.15 | 1.57 |
| **7** | {Male, other road, hit & run =Yes}→{injured=1, Daylight, Fine} | 0.89 | 1.93 | 0.15 | 2.42 |
| **8** | {injured=1, Daylight, Fine}→ {Male, other road, hit & run= Yes} | 0.67 | 1.93 | 0.15 | 1.57 |
| **9** | {Male, other road} → {injured=1, hit & run=No, Fine} | 0.8 | 1.89 | 0.14 | 1.92 |
| **10** | {injured=1, hit & run=No, Fine}→ {Male, Other road} | 0.73 | 1.89 | 0.14 | 1.69 |
| **C2** | | | | | |
| **11** | {Female} → {other road, hit & run=No} | 0.83 | 2.5 | 0.2 | 2 |
| **12** | {other road, hit & run=No}→ {Male} | 1 | 2.5 | 0.2 | 3 |
| **13** | {hit & run=Yes}→ {Male, other road, Daylight} | 0.71 | 2.14 | 0.18 | 1.56 |

| | | | | | |
|---|---|---|---|---|---|
| **14** | {Male, other road, Daylight}→ {hit & run=Yes} | 1.0 | 2.14 | 0.18 | 2.67 |
| **15** | {Male, other road}→ {hit & run= Yes, Daylight} | 0.83 | 2.08 | 0.17 | 1.8 |
| **16** | {hit & run= Yes, Daylight}→ {Male, other road} | 0.83 | 2.08 | 0.17 | 1.8 |
| **17** | {Female, other road}→ {hit & run= No} | 1 | 1.88 | 0.16 | 2.33 |
| **18** | {hit & run =No}→ {Male, other road} | 0.63 | 1.88 | 0.16 | 1.33 |
| **19** | {Fatal, Very cold}→{Rural} | 0.83 | 1.79 | 0.15 | 1.6 |
| **20** | {Rural}→{Fatal, Very cold} | 0.71 | 1.79 | 0.15 | 1.4 |
| **C3** | | | | | |
| **21** | {Female}→{other road} | 1 | 1.67 | 0.12 | 2.4 |
| **22** | {other road}→{ Female} | 0.5 | 1.67 | 0.12 | 1.2 |
| **23** | {other road}→{ hit & run= No, Urban} | 0.5 | 1.67 | 0.12 | 1.2 |
| **24** | {hit & run No, Urban}→{other road} | 1 | 1.67 | 0.12 | 2.4 |
| **25** | { Male, Daylight}→{Fine} | 0.6 | 1.5 | 0.1 | 1.2 |
| **26** | {Fine}→{ Male, Daylight} | 0.75 | 1.5 | 0.1 | 1.33 |
| **27** | {Fatal}→{Urban} | 1 | 1.43 | 0.09 | 1.8 |
| **28** | {Urban}→{Fatal} | 0.43 | 1.43 | 0.09 | 1.09 |
| **29** | {Age=21-30}→{Urban} | 1 | 1.43 | 0.09 | 1.8 |
| **30** | {Urban}→{ Age=21-30 } | 0.43 | 1.43 | 0.09 | 1.09 |
| **C4** | | | | | |
| **31** | {Slight curve) →{Very hot} | 0.75 | 1.67 | 0.12 | 1.47 |
| **32** | {Very hot}→{Slight curve} | 0.67 | 1.67 | 0.12 | 1.35 |
| **33** | {Slight curve} → {Daylight, Very hot} | 0.75 | 1.67 | 0.12 | 1.47 |
| **34** | {Slight Curve, Daylight} → {Very hot} | 0.75 | 1.67 | 0.12 | 1.47 |
| **35** | {Very hot}→{Slight Curve, Daylight} | 0.67 | 1.67 | 0.12 | 1.35 |
| **36** | {Daylight, Very hot}→ {Slight Curve} | 0.67 | 1.67 | 0.12 | 1.35 |
| **37** | {injured=1, hit & run = No}→ {Male, Daylight} | 0.86 | 1.56 | 0.11 | 1.57 |
| **38** | {Male, Daylight}→{injured= 1, hit & run=No} | 0.55 | 1.56 | 0.11 | 1.19 |
| **39** | {injured1, Male}→{hit & run= No, Daylight} | 0.75 | 1.5 | 0.1 | 1.33 |
| **40** | {hit & run=No, Daylight} → {injured=1, Male } | 0.6 | 1.5 | 0.1 | 1.2 |
| **C5** | | | | | |
| **41** | {hit & run= No} →{Fatal, Daylight} | 0.71 | 2.14 | 0.15 | 1.56 |
| **42** | {Fatal, Daylight} → {hit & run= No} | 0.83 | 2.14 | 0.15 | 1.83 |
| **43** | {Fatal, other road}→{Slight curve} | 0.71 | 1.84 | 0.13 | 1.43 |
| **44** | {Slight curve}→{Fatal, other road} | 0.71 | 1.84 | 0.13 | 1.43 |
| **45** | {State highway}→{injured= 2, hit & run=Yes} | 0.71 | 1.84 | 0.13 | 1.43 |
| **46** | {injured=2, hit & run=Yes} → {State highway} | 0.71 | 1.84 | 0.13 | 1.43 |
| **47** | {Slight Curve}→{other road, Daylight} | 0.71 | 1.84 | 0.13 | 1.43 |
| **48** | {other road, Daylight} → {Slight Curve} | 0.71 | 1.84 | 0.13 | 1.43 |
| **49** | {other road}→{Slight curve, Daylight} | 0.5 | 1.8 | 0.12 | 1.2 |
| **50** | {Slight curve, Daylight} → {other road} | 1 | 1.8 | 0.12 | 2.22 |
| **EDS** | | | | | |
| **51** | {injured=1}→{hit & run= No, Daylight} | 0.57 | 1.23 | 0.06 | 1.19 |
| **52** | {hit & run=2, Daylight} → {injured=1} | 0.65 | 1.23 | 0.06 | 1.27 |
| **53** | {hit & run= No}→{injured= 1, Daylight} | 0.53 | 1.14 | 0.04 | 1.1 |
| **54** | {Injured=1, Daylight} → {hit & run= No} | 0.65 | 1.14 | 0.04 | 1.16 |
| **55** | {Male} →{Daylight, rural} | 0.45 | 1.14 | 0.04 | 1.07 |
| **56** | {Daylight, Rural} →{Male} | 0.75 | 1.14 | 0.04 | 1.24 |
| **57** | {Male, Daylight}→{Rural} | 0.57 | 1.13 | 0.03 | 1.1 |
| **58** | {Rural} →{Male, Daylight} | 0.6 | 1.13 | 0.03 | 1.12 |

## VI. CONCLUSION

This paper presents a method for analyzing accident patterns of different types of accidents on roads, which uses k to represent clustering and association rules mining algorithms. The study used accidents in the Maharashtra road network in 2015 and 2016. K-means clustering finds five clusters (C1-C5) based on attribute accident type, road type, light condition, and road characteristics. Association rule mining has been applied to every cluster as well as the whole data set to create rules. Strong rules are used for analysis with high lift values. The rules of every cluster disclose situations related to incidents within the cluster.

## REFERENCES

[1] Sachin Kumar, Durga Toshniwal, "A data mining approach to characterize road accident locations", J. Mod. Transport. (2016) 24(1):62–72.

[2] S. Shanthi and Dr. R. Geetha Ramani, "Gender Specific Classification of Road Accident Patterns through Data Mining Techniques", IEEE-International Conference on Advances in Engineering, Science and Management (ICAESM -2012) March 30, 31, 2012.

[3] Tessa K. Anderson, "Kernel density estimation and K-means clustering to profile road accident hotspots", Accident Analysis and Prevention 41 (2009) 359–364.

[4] Miao Chong, Ajith Abraham and Marcin Paprzycki, "Traffic Accident Data Mining Using Machine Learning Paradigms", Oklahoma State University, USA.

[5] Tibebe Beshah Tesema, Ajith Abraham And Crina Grosan, "Rule Mining and Classification Of Road Traffic Accidents Using Adaptive Regression Trees", I. J. of Simulation Vol. 6 No 10 and 11.

[6] Tibebe Beshah Tesema, Ajith Abraham, Dejene Ejigu, "Learning the Classification of Traffic Accident Types", copyright IEEE 2012.

[7] Sachin Kumar, Durga Toshniwal, "A data mining framework to analyze road accident data",
Journal of Big Data (2015).

[8] Tibebe Beshah, Dejene Ejigu, Ajith Abraham, Vaclav Snasel, Pavel Kromer, "Pattern Recognition and Knowledge Discovery from Road Traffic Accident Data in Ethiopia:Implications for improving road safety", 978-1-4673-0126-8/11 © 2011 IEEE.

[9] "Road Accident Recording Forms A-1 and A-4", IRC: 53-1982.

[10] Kama Koti, "Road Accident Recording Forms A-1 and A-4 (second revision)", IRC: 53-2012.

[11 Robert Tibshirani, Guenther Walther and Trevor Hastie, "Estimating the number of clusters in a data set via tha gap statistic", Stanford University, USA.

[12] Sachin Kumar, Durga Toshniwal, " A comparative analysis of heterogeneity in road accident data using data mining techniques", J. Mod. Transport. (2016) 24(1):62–72.