JCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-9, Sept. 2018 E-ISSN: 2347-2693

DM Algorithms Based Clustering for Road Accident Data Analysis

Shaik Subhani

Dept. of IT, SNIST(Autonomous), Hyderabad, India

**Corresponding Author: shaiksubhani@sreenidhi.edu.in*

Available online at: www.ijcseonline.org

Accepted: 10/Sept/2018, Published: 30/Sept/2018

Abstract- Road accidents due to traffic are ever more being acknowledged as major problem for transportation agencies as well as general people. A substantial unexpected result of transportation systems is road accidents with injuries and loss of lives. In this scenario purpose safe driving, specific study of road traffic data is severe to find out elements that are connected to mortal accidents. In this research paper, we determine factors behind road traffic accidents issues solving by data mining algorithms together with data mining algorithms like Density-based spatial clustering of applications with noise and Parallel Frequent mining. We primarily separate the accident locations into k clusters depend on their accident frequency with Density-based spatial clustering of applications with noise algorithm. Next, parallel frequent mining algorithm is apply on these clusters to disclose the association between dissimilar attributes in the traffic accident data for realize the features of these places and analyzing in advance them to spot different factors that affect the road accidents in different locations. The main objective of accident data is to recognize the key issues in the area of road safety. The efficiency of prevention accidents based on consistency of the composed and predictable road accident data using with appropriate methods. Road accident dataset is used and implementation is carried by using Weka tool. The outcomes expose that the combination of Density-based spatial clustering of applications with noise and parallel frequent mining explores the accidents data with patterns and expect future attitude and efficient accord to be taken to decrease accidents.

Keywords- Accident analysis, Density-based spatial clustering of applications with noise, Road accident dataset, parallel frequent mining, Weka

I. INTRODUCTION

Key issue of the road traffic accidents are a concern for transportation leading traders as well as common people. Road accidents are damage the public life with multi level of injuries [1]. The number of factors that influence these incidents like Environmental conditions, motorway design, and type of accident, driver characteristics, and vehicle attributes [4]. The key objective of accident data analysis recognizes the major parameters associated to road traffic accidents [2]. However, various natures of accident data generate the task analysis is tough context. The key problem in this accident data analysis disturbs the human life. Thus heterogeneity have to be measured during data analysis [3], a few correlation between the data may remain out of sight. Although, researchers used partition of the data to decrease this heterogeneity using few measures such as professional knowledge, but there is no security that will guide to a best possible partition which consists of similar type of clusters of road accidents. Data partition has been used broadly to overcome this dissimilarity of the accident data [1]. In order to provide safe driving instructions, cautious road traffic of statistics is critical to discover variables that are related to mortal accidents. Data analysis has the ability to recognize the various logics behind road accidents [5]. In this paper, we are building data mining methods to make out high-frequency accident places and additional data to identify the different factors that influence road accidents at different locations. We initially split the accident places into m clusters with the support of accident frequency via Density-based spatial clustering of applications with noise clustering algorithm [6]. The frequent pattern mining algorithm is imposed on these for expose the connection between dissimilar attributes of accident data with dissimilar places. Hence, our major accent will be the understanding of the results. The key idea of this research inspect the responsibility of human, vehicle, and infrastructure-with correlated factors in accident sternness by applying data mining learning techniques on road accident information [7]. The rest of paper is prepared as follows: Section 2 states brief description road accident problems are analyzed in facts. In section 3, propose road accident data framework. Section 4 presents a comparison on road accident analysis techniques. Conclusion of the study is presented in section 5.

II. ANALYSIS OF ROAD ACCIDENTS

A. Reasons for road accidents

Different reasons for road accidents are: 1. Road Users - lack of care, High speed rash driving, abuse of traffic rules, sleep, fatigue and alcohol etc. 2. Vehicle - Defects such as brakes failure, steering system of vehicle, tire burst, and lighting system. 3. Road Condition - Skidding surface of roads, pot holes on roads. 4. Road design - imperfect geometric design of roads, insufficient breadth of roads, awkward curve design, improper traffic maintenance and poor lighting. 5. Environmental factors -critical weather conditions like smoke, snow, mist and heavy rainfall which bound the normal visibility and makes driving is not safe [14].

B. Road accidental injuries and deaths

There has been advance in road accidental deaths in India over the last few years. Road accidental deaths have increased 9 times, from 15,500 in 1980 to 148,400 in 2018. In comparison to 2004, injuries in 2018 are superior by 55,000 and 89,000, respectively. Table 1 represent all about information. From 2004 to 2013, fatalities have increased of 6% rate per year while the population of the country has larger than before at rate of 1.5% per year. Consequently, road accidental deaths per one lack people, has greater than before 8.8 in 2004 and 12.2 in 2018. In India fatality risk is very high level compared with developed countries. This type of risk in India is high than in the United Kingdom, Sweden. Road accidental deaths occurred due to one lack vehicles, as of 87.5 in 1980 to 8.6 in 2018, it is still quite high compare with developed countries.

Table 1. 1980-2018 Road Accident Statistical data

Year	**Road accidents**	**Road accidental deaths**	**Accident risk**	**Road accidental injuries**	**Fatality risk**	**Fatality rate**	**Accident severity index**
1980	114.1	14.5	21.6	70.1	2.7	87.5	12.7
1990	153.2	24.6	23.1	109.1	3.7	54.4	16.1
2000	282.6	54.1	34.4	244.1	6.6	28.2	19.1
2003	308.3	80.1	30.8	340.2	8.0	16.6	26.0
2006	336.4	84.4	31.5	382.9	7.9	12.6	25.1
2009	430.6	133.9	36.3	470.6	11.3	10.5	31.1
2012	443.0	137.4	36.1	469.9	11.2	8.6	31.0
2015	446.1	138.1	36.0	472.1	11.4	9.1	32.1
2018	451.6	138.4	36.3	471.4	11.1	8.7	32.5

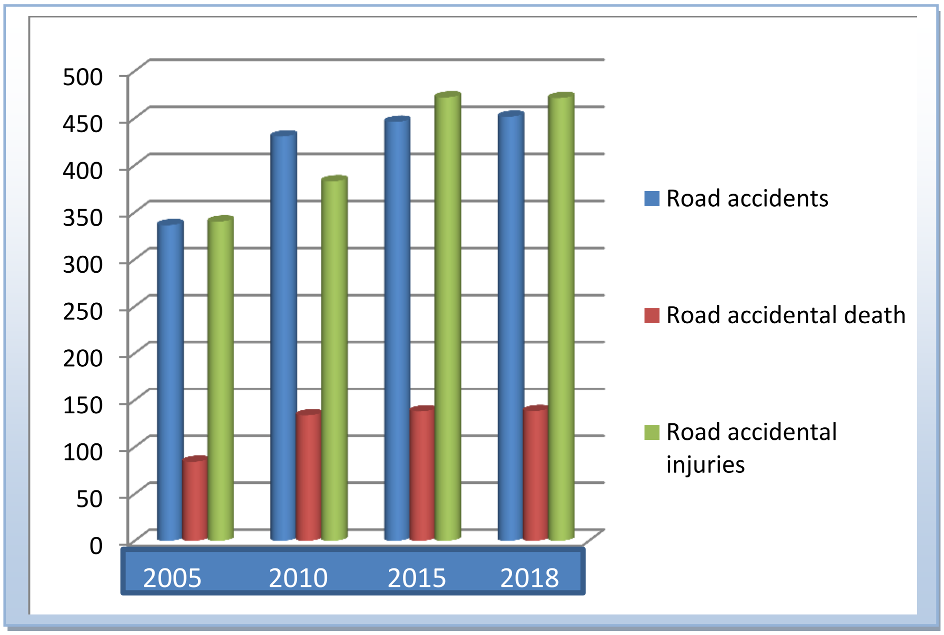

Figure 1. Graphical Representation of 1980-2018 Road Accident Statistical data

C. Road accidental distribution based on Age, Time and sex

The road accident distribution clearly shows that the most creative age group, 20-45 years, is the flat to road accident fatality in India. Age group of 20-40 years comprises 23% of Indian population, faces roughly 38% of total road accidents. During the previous 10 years from 2005 to 2018, number of fatalities faced by this age cluster has also improved significantly. The middle age (30-40) group 13% of the total population, but fatality faces 22%. So age group 30-59 years, the inexpensively energetic age group, is the most susceptible population cluster in India. Half of the road accidents are faced by this group of people which counts for less than 1/3 of the entire population. Sex wise allocation of injuries and road accidental deaths in India for the year 2004 and 2018 presents that the males for 86.2% of all fatalities 81.1% of injuries in 2018. Past 11 years, total number of fatalities by males has improved by 69.3%.

III. MATERIALS AND METHODS

The primary thing of analysis is data pre- processing is the primary step for remove noise from given input. In second phase attribute selection done by Density-based spatial clustering of applications with noise algorithm. parallel frequent mining algorithm is apply on these clusters to disclose the association between dissimilar attributes in traffic accident data for realize the features of these places and analyzing in advance those to spot different factors that affect the road accidents. Finally visualize the patterns of performance evaluation.

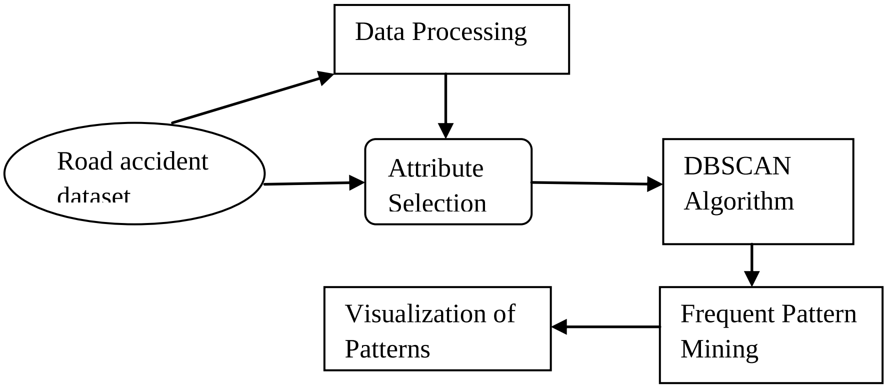

Figure 2. Proposed system architecture

Cluster analysis split the data components into different groups in a way that maximizes the homogeneity of components within the different clusters. This technique is known as an unsupervised learning algorithm as the accurate number of clusters and their shapes are unknown. Generally, cluster analysis is a procedure of repetitively maximizing the intra cluster components. These similarity-based clustering methods calculate similarity using a specific distance function and measures for components with qualitative. The well known among similarity-based method is density-based approach [15].

Aim: Road accidents analysis using data mining algorithm called DBSCAN

Input: D, data set of road accidents; K, no of clusters; M, mean for each cluster

Algorithm Density-based spatial clustering of applications with noise (D, epsilon, min_points):

```
    C = 0
   for each unvisited point P in dataset
              mark P as visited
              sphere_points = regionQuery(P, epsilon)
    if sizeof(sphere_points) < min_points
              ignore P        else
            C = next cluster
  expandCluster(P, sphere_points, C, epsilon, min_points)

  expandCluster(P, sphere_points, C, epsilon, min_points):
            add P to cluster C
 for each point P’ in sphere_points           if P’ is not visited
              mark P’ as visited
              sphere_points’ = regionQuery(P’, epsilon)
      if sizeof(sphere_points’) >= min_points
               sphere_points = sphere_points joined with sphere_points’
 if P’ is not yet member of any cluster
               add P’ to cluster C,     regionQuery (P, epsilon):
return all points within the n-dimensional sphere centered at P with radius epsilon (including P).
```

Output: k cluster groups

Figure 3. Density-based spatial clustering of applications with noise

A. Density-Based Road Accident analysis

Density-based spatial clustering of applications with noise is a density-based clustering algorithm and it is designed to overcome large data sets with noise and is capable of determining different sizes and shapes. Density-based means that cluster are connected points where the density of points is equal to or more than a threshold. If the density is less than the threshold, the data are considered as noise. When a data set is given, Density-based spatial clustering of applications with noise divides it into segments of clusters and a set of noise points.[15] The density threshold condition is that there should be at least Min Pts number of points in ε-neighborhood. Clusters contain core points and boundary points. A core point is a point that meets the density condition, and a boundary point is a point that does not meet the density condition but is close enough to one or more core point’s ε-neighborhood. Points that are not core points or boundary points are considered as noise. Below is the pseudo code, prepared as functions for road accident data analysis. The function of regionQuery( ) proceeds the points within the n-dimensional sphere. The function expandCluster ( ) returns for every points in the sphere, the Density-based spatial clustering of applications with noise algorithm is presented below in figure 4.

The idea behind Density-based spatial clustering of applications with noise and its developments is the notion that points are assigned to the similar group if they are density-reachable from every other cluster. To know this model, we will go through the definitions used in Density-based spatial clustering of applications with noise and associated algorithms. Clustering starts with dataset E containing a set of point’s p ∈ E. Density-based spatial clustering of applications with noise estimates the density around a point using the concept of ϵ-neighborhood.

1. ϵ -Neighborhood. The ϵ -neighborhood, N ϵ(a), of a data point p is the set of points within a specified radius ϵ around p.

2. M ϵ(a) = $\{ b \mid d(a,b) < \epsilon \}$ where d is some distance measure and $\epsilon \in R^{+.}$ Note that the point p is always in its own ϵ -neighborhood, i.e., a ∈ M ϵ(a) always holds.

Following this definition, the size of the neighborhood |M ϵ (a)| can be seen as not normalized kernel density estimate around p using a uniform kernel and a bandwidth of ϵ. Density-based spatial clustering of applications with noise uses minPts, detect dense areas for classify the points in a dataset into core, border, or noise points of the cluster.

1. Point classes, A point a ∈ E is classified as a center point if M ϵ(a) has large density, i.e., |M ϵ(a)| minPts where minPts ∈ Z^+ is a user defined density threshold,
2. Directly density-reachable, A point b ∈ E is density-reachable from a point a ∈ E with respect to ϵ and minPts if, and only if,
i.|M ϵ(a)| ≥ minPts, and
ii.b ∈ M ϵ(a).
That is, p is a core point and q is in its ϵ -neighborhood.
3.Density-reachable, A point p is reach to density from q if their exist in E in sequence of points (a_1, a_2, ..., a_n) with b = a_1 and a = a_n such that a_i+1 directly density reach from a_i ∀ i 2 {1, 2, ..., n − 1}.
4. Connected density, A point a ∈ E is connected density to a point b ∈ E if there is a point o ∈ E, both a and b is density-reachable from o.

B.Parallel Frequent Association Mining

Association rule mining is an extremely popular data mining method that extracts attractive and hidden relations between dissimilar attributes in a huge dataset. Association rule mining generates different rules that illustrate the underlying patterns in the dataset. The FP-growth algorithm using for the issue of discovery frequent patterns recursively add the suffix. This algorithm uses minimum frequent items as a suffix; it is well selection for the process reduce the search cost and extracts the frequent patterns. The FP growth algorithm is shown below in figure 4.

```
Algorithm FP- growth (FPT, S, P)
// FPT – Tree on Frequent Items
// S-Minimum Support and P-Current Item set Suffix.
Begin
   1.   If FPT is a single path do
   2.   For every C of nodes in path do
          a.  Inform all patterns C ∪ P;
                                          Else
          b.  For every item i in FPT do
              Begin
                         i.   Produce pattern Pi = set i ∪ P;
                        ii.   Inform pattern Pi as frequent;
               End
   3.    Use pointer to extract condition prefix paths for item one;
   4.   Construct conditional Frequent Pattern Tree FPTi from condition
   5.   From prefix paths after eliminating infrequent items;
   6.   If (FPTi ≠ ∅) FP- growth (FPTi, Pi, S)
              End
   7.   End
```

Figure 4. FP Growth algorithm

Using Bayes rule, we can find the probability of label given the observation of a frequent pattern FP_i as:

$$P\,(K/FP_i) = \frac{P\left(\frac{FP_i}{K}\right) * P(K)}{P(FP_i)} \qquad (1)$$

P(K) is the probability of the label which is assumed constant, given by NK/T where NK is the number of images of the class K, and T is the total number of images across all classes . We guess equal number of training data items for all classes i.e, $NK_i = NK_j$. the above assumption P (FP_i | K) can be rewritten as N_{FP}^{K} / N_K. The probability of observing a frequent pattern P (FP_i) is NFP_i / T i.e., the number of data items on which FP_i fired regardless of the label, separated by the total number of images [19]. Substituting all of these in the above equation we have,

$$P\,(K / FP_i) = \frac{P\left(\frac{FP_i}{K}\right) * P(K)}{P(FP_i)} = N_{FP_i}^{K} \text{ X } \frac{N^K}{T} \text{ X } \frac{T}{N_{FP_i}}$$

Therefore $$P(K/FP_i) = \frac{N_{FP_i}^{K}}{N_{FP_i}} \qquad (2)$$

The above outcome displays that for testing a label, the operator sets should be ordered according to the ascending order of $\frac{N_{FP_i}^{K}}{N_{FP_i}}$. This is for an operator set to get a better score in this phase, either the frequency of observing the operator set FP_i for the particular label is high or that the probability of the operator set FP_i firing for other classes is less.

Data pre- processing is the primary step for remove noise from given dataset. Next level attributes selection done by Density-based spatial clustering of applications with noise algorithm. It can be constructing as a groups based on attributes. parallel frequent mining algorithm is apply on these clusters to disclose the association between dissimilar attributes in traffic accident data for realize the features of these places and analyzing in advance those to spot different factors that affect the road accidents. Finally visualize the patterns of performance evaluation.

IV. DATASET DESCRIPTION

The road accident dataset consists of 11,574 road accidents from 2012 to 2016 for last 6 years period. After preprocessing of the data, 11 variables were recognized for the research with satisfactory. The dataset comprised of accident features time, type of accident, and number of injured victims age, gender, road type, and area around accident location [1]. Data add number of people, vehicles involved, road surface, location and weather conditions. The Easting's and Nothings are generated at the roadside where the accident occurred. Attributes of dataset include reference number Northing and Easting number of vehicles, Accident Date, Time (24× 7) 1st road class, Road Surface, Lighting Conditions, Weather Conditions with reference number Grid Ref: Easting Grid Ref: Northing Number of vehicles Accident Date Time 21G0539, 427798, 426248, 5, 16/01/2015, 1205; 21G0539, 427798, 426248, 5, 16/01/2015, 1205; 21G1108, 431142, 430087, 1, 16/01/2015, 1732; 21H0565, 434602, 436699>, 1, 17/01/2015, 930; 21H0638, 434254, 434318, 2, 17/01/2015, 1315; 21H0638, 434254, 434318, 2, 17/01/2015, 1315; .

V. RESULTS AND DISCUSSION

A variety of data mining methods, algorithms and tools are proposed for road traffic accident data analysis accident location tracking, prediction and identification of different contributory factors that affect the accident cruelty levels. Liu et al. [17] they have been construct statistical design using stepwise regression analysis method for guessing incident duration. The result analysis displays that over 85% of differences in occurrence duration can be predicted by the eight factors implicated in the regression model. Density-based spatial clustering of applications with noise and frequent pattern mining algorithms are used for clustering, and the following clusters are constructed. Cluster 1 represents the traffic clusters in such a way accidents occur because of high traffic. Cluster 2 represents the time of accident cluster in which accidents happen during day and night time. Cluster 3 represents the age of the drivers cluster. Cluster 4 presents the accident occurred every month. Cluster 5 states the weather condition at the time of accident. Cluster 6 is the lightening condition issue on the roads. Cluster 7 describes about type of accident the road condition. Cluster 8 describes the speed limit of vehicles at the time of accident. F-measure is used for cluster analysis because it throughput node-based analysis using the following equations.

$$\text{Precision} = TP/\ (TP + FP) \quad (3)$$

$$\text{Recall} = TP/(TP + FN) \quad (4)$$

$$\text{F-Measure} = (1+ \propto)\ /\ ((1 + \text{precision}) + (\propto/\text{Recall})$$

Where $\propto = 1$ (5)

Cluster based analysis findings and road accident dataset analysis are compared. The outcome reveal that the mixture of Density-based spatial clustering of applications with noise clustering and frequent pattern mining is extremely inspirational as it generates important data that would remain hidden, if no partition has been performed prior to produce frequent item sets. Weka is data mining software that uses a collection of machine learning algorithms. These algorithms can be applied directly to the data. The following table shows data mining algorithms, Comparison for road accident analysis of different methodologies, classifiers and their result.

Table 2. Data mining Algorithm for Comparison of Road Accident Analysis

Application	Methodology	Classifiers	Result	Efficiency
Traffic accidents analysis based on road users	K-modes Clustering	Support Vector Machine	77.99 %	Low
A prospective traffic accident analysis	PART algorithm	Random Forest Tree	86.66%	High Low
Classification of vehicle crash structure in road accidents	CS-MC4 algorithm	Naive Bayes Classifier	82.59%	Medium
Traffic accidents in Dubai	Apriori Algorithm	Association rules	80.63%	Low
Detection of key factors for traffic injury harshness	CART Algorithm	Rule Induction	74.49 %	High

predicting cause of accident places on highways	ID3 Algorithm	Decision Tree	79.70%	Low
Traffic incident duration calculation based on ANN	ANN Algorithm	Artificial Neural Network	87.35%	Low
Road accidents in Korea	DT Algorithm	Artificial Neural Network	87.2%	High
A data mining framework for road accident data analysis	K-Modes Algorithm	Association Rule Mining	79.5%	High
Gender-specific classification of road accident patterns	C4.5 Algorithm	Random Tree	85.02%	Low
Imbalanced traffic accidents datasets	Naive Bayesian Algorithm	Bayesian networks classifiers	.80.2%	Low
Identifying accident-prone locations	fuzzy K-NN Algorithm	Bayesian networks classifiers	80.6%	Medium
Accident Data analysis	Density-based spatial clustering of applications with noise Algorithm	FP-Growth	99.6%	High

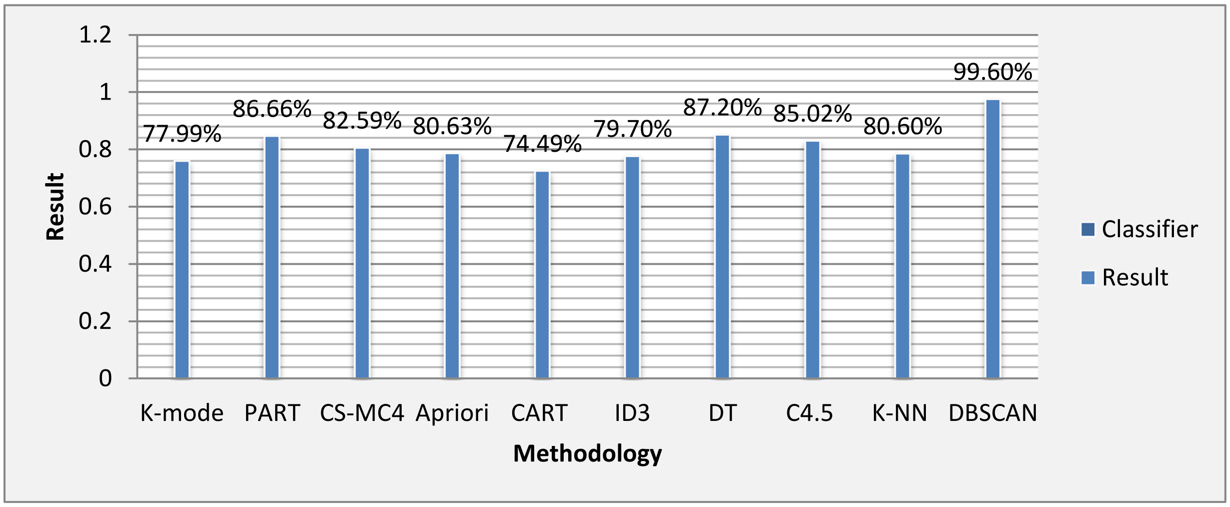

Figure 5. Assessment of Data mining Algorithms

Figure 5 shows the graphical representation of table 2 values. The following table statistical results prove the Density-based spatial clustering of applications with noise with combination of FP growth generates better results compare to other methods. In this combination of methodology datasets with altering densities are tricky. So they can be working aggressively up to datasets are not alter.

VI. CONCLUSIONS

Data mining has been confirmed as a consistent method in analyzing road accident data. Several authors used data mining method for analyzing road accident data of dissimilar countries. The data mining methods like association rule mining, clustering and classification are generally used accepted multiple reasons that affect the severe of road accidents. In this situation current safe driving suggestions and careful road traffic data analysis is unsafe to find out factors that are powerfully related to unhelpful accidents. In this research, we situate so many factors behind road accidents, these accidents are analysis by using data mining algorithms like Density-based spatial clustering of applications with noise and Parallel Frequent mining algorithms. We primarily crack the accident locations into k groups based on their frequency of accident outcome by means of Density-based spatial clustering of applications with noise algorithm. Next, parallel frequent mining algorithm is exposing the relationship between different attributes in accident data, when it is applied on groups. Understand the features of these places and additionally analyzing those to identify dissimilar factors affect the road accidents at dissimilar places. The main objectives of road accident data are inspection to detect the major problem in the field of road safety. The efficiency of accident avoidance depends considerably on the reliability of collected and estimated data. Road accident dataset is used and execution is carried by using Weka tool. The result reveal that dataset for road accident and its analysis using Density-based spatial clustering of applications with noise and FP mining algorithm reveal that this procedure can be reused on novel accident data with extra attributes to detect dissimilar factors linked with road accidents.

ACKNOWLEDGEMENT

We would like to express thanks to Department of Information Technology, Sreenidhi Institute of Science and Technology, Hyderabad, India for supporting this research development.

REFERENCES

[1] Sachin K and D Toshniwal,” A data mining framework to analyze road accident data”, Journal of Big Data, 2005.

[2] Savolainen P, Mannering F and Quddus,” The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternative"s”, Accid Anal Prev., 43:1666–76,2011.

[3] Depaire B, Wets G and Vanhoof K ,” Traffic accident segmentation by means of latent class clustering, accident analysis and prevention”, vol. 40, Elsevier, 2008.

[4] Karlaftis M and Tarko A ,”Heterogeneity considerations in accident modeling, Accid Anal Prev, 30(4):425–33,1998.

[5] Ma J and Kockelman K ,” Crash frequency and severity modeling using clustered data from Washington state”, IEEE Intelligent Transportation Systems Conference, Toronto, Canadá,2008.

[6] Jones B and Janssen L ,” Analysis of the frequency and duration of freeway accidents in Seattle, accident analysis and prevention”, Elsevier, vol. 231991.

[7] Miaou SP and Lum H ,” Modeling vehicle accidents and highway geometric design relationships, accident analysis and prevention”, Elsevier, vol. 25,1993.

[8] Morth ,” Road Accidents in India 2013. New Delhi: Ministry of Road Transport and Highways Transport Research Wing”, Government of India,2014.

[9] Kononov J and Janson BN ,” Diagnostic methodology for the detection of safety problems at intersections”, Transportation Research Record: Journal of the Transportation Research Board, Vol. 1784,2014.

[10] Lee C, Saccomanno F and Hellinga B ,” Analysis of crash precursors on instrumented freeways”, Transportation Research Record: Journal of the Transportation Research Board, Vol. 1784,2014.

[11] Chen W and Jovanis P ,” Method for identifying factors contributing to driver-injury severity in traffic crashes, Transportation Research Record: Journal of the Transportation Research Board, Vol. 1784,2014.

[12] S. K. Barai ,” Data mining application in transportation engineering Transport", journal of Transport, Vol XVIII, Issue 5, PP: 216-223,2003.

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

[14] T. Soni Madhulatha ,” an overview on clustering methods", IOSR Journal of Engineering, Vol. 2, pp: 719-725,2012.

[15] PandyaJalpa P. and Morena Rustom D ,” A Survey on Association Rule Mining Algorithms Used in Different Application Areas", International Journal of Advanced Research in Computer Science, Volume 8, No. 5,2017.

[16] A. Garib, and A.E Radwan ,” Estimating Magnitude and Duration of Incident Delays”, Journal of Transportation Engineering, vol. 123, pp. 459-465,1997.