Density-Based Clustering For Road Accident Data Analysis

S. Nagendra Babu[1], Dr.J. Jebamalar Tamilselvi[2]
[1]R & D Center, Bharathiar University, Coimbatore, India
[2]Jaya Engineering College, Thiruninravur, Chennai, India

Abstract- *Now a day, traffic safety is a major concern for transportation governing agencies as well as ordinary citizens. In order to give safe driving suggestions, careful data analysis of roadway traffic data is critical to find out variables that are closely related to fatal accidents. In this study, we find factors behind road traffic accidents using data mining algorithms including DBSCAN cluster and Parallel Frequent mining algorithm. We first divide the accident locations into k groups based on their accident frequency counts using DBSCAN clustering algorithm. Next, Then parallel frequent mining algorithm is applied on these to reveal the correlation between different attributes in the accident data and understand the characteristics of these locations and further analyzing them to identify various factors that affect road accidents at those locations. UKDA data set is used and implementation is carried by using Weka tool. The key objective of data analysis is to identify the main problems in the field of road safety. The efficiency of accident prevention depends significantly on the reliability of the collected and estimated data and the appropriateness of the used methods. The results reveal that the combination of DBSCAN clustering and parallel frequent mining explore the accidents data recorded by the police information system, and discover patterns and predict future behaviors and effective decisions to be taken to reduce accidents.*

Keywords- Accident analysis, DBSCAN, FP Growth, UKDA, Weka.

I. INTRODUCTION

Road safety is an important aspect of urban and extra urban transportation systems, particularly due to the high social costs it involves. The resources devoted to the road safely each year are largely smaller than the real needs; users spend more than 60 billion of ECU for accident refunds by means of the insurance companies, while the amount devoted to the prevention of accidents is very lower. The situation is different from Country to Country, both for the differences in the regulations in force and for the different sensibility of users to the safety problem. Different studies carried out in this field linked the risk of accidents, the percentage or the number of accidents, the number of fatal accidents (dependent variables) to different factors or explanatory variables (independent variables) such as: age, and/or sex of the driver, expert or inexpert drivers, speed, length of the network, use of the safety belts, meteorological conditions and so on.Every year there are 0.4 million accidents reported in India, which makes India a country with large accident rate. However, as accidents are unpredictable and can occur in any type of situation, there is no guarantee that this trend will sustain in future also. Therefore, the identification of different geographical locations where most of the accidents have occurred and determining the various characteristics related to road accidents at these locations will help to understand the different circumstances of accident occurrence.

Road and traffic accidents are one of the major causes of fatality and disability across the world. Road accident can be considered as an event in which a vehicle collides with other vehicle, person or other objects. A road accident not only provides property damage but it may lead to partial or full disability and sometimes can be fatal for human being. Increasing number of road accidents is not a good sign for the transportation safety. The only solution requires the analysis of traffic accident data to identify different causes of road accidents and taking preventive measures. A variety of research has been done on road accident data from different countries. Various research studies used different techniques to analyze road accident data using statistical techniques and provide fruitful outcomes. Different other studies useddata mining techniques to analyze road accident data and also claim that data mining techniques are more advanced and better than traditional statistical techniques. Although, both the approaches provided good outcome that certainly useful for traffic accident prediction, reveals that heterogeneity in road accident data exists and should be removed prior to the analysis of road accident data. They also suggested that use of suitable clustering techniques prior to the analysis of accident data reduces the heterogeneity from data and can help in revealing hidden information.

The paper is organized as follows: Next, a brief description of the literature review is given and Traffic accident detection in section 3. In section 4, "Proposed Methodology", a framework is proposed to analyze the road

accident data. In section5," Results and Discussion", the results and findings are elaborated and discussed. Finally, we concluded in section 6,"conclusion and suggestion".

II. 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. 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. Therefore, cluster analysis can assist the segmentation of road accidents.

Lee et al. indicated that statistical models were a good choice in the past to analyze road accidents to identify the correlation between accident and other traffic and geometric factors. However, Chen and Jovanis determined that analyzing large dimensional datasets using traditional statistical techniques may result in certain problems such as sparse data in large contingency tables. Also, statistical models have their own model specific assumptions and violation of these can lead to some erroneous results. Due to these limitations of statistical techniques, data mining techniques are being used to analyze road accidents. Data mining is a set of techniques to extract novel, implicit and hidden information from large data. Barai discussed that there are various applications of data mining in transportation engineering such as road roughness analysis, pavement analysis and road accident analysis. Various data mining techniques such as association rule mining, classification and clustering are widely used for the analysis of road accidents.
Accident cases in India are usually recorded by police officer of the region in which the accident has occurred. Also, the area covered by a police station is limited and they keep record of accidents that have occurred in their regions only. Ponnaluri discussed that the report prepared by police only contains the basic information that are not much useful for the research purpose. He suggests that data collection method used by police needs a lot of improvement. However, Indian researchers used these data and analyzed it for some highway portions using statistical methods. Data mining can be described as a novel technique to extract hidden and previously unknown information from the large amount of data. Several data mining techniques such as clustering, classification and association rule mining are widely used in the road accident analysis by researchers of other countries. Geurts et al. used association rule mining technique to understand the various circumstances that occur at high-frequency accident locations on Belgium road networks. Tesema et al. [26] used adaptive regression tree model to build a decision support system for the road accidents in Ethiopia. Abellan et al. developed various decision trees to extract different decision rules for different trees to analyze two-lane rural highway data of Spain. They found that bad light conditions and safety barriers badly affect the crash severity. Depaire et al. [28] used clustering technique to analyze road accident data of Belgium and suggest that cluster-based analysis of road accident data can extract better information rather analyzing data without clustering. Kashani et al.used classification and regression tree (CART) to analyze road accidents data of Iran and found that not using seat belt, improper overtaking and speeding badly affect the severity of accidents. Kwon et al. used Naïve Bayes and decision tree classification algorithm to analyze factor dependencies related to road safety. Severity of accident is directly concerned with the victim involved in accidents, and its analysis only targets the type of severity and shows the circumstances that affects the injury severity of accidents. Sometime accidents are also concerned with certain locations characteristics, which makes them to occur frequently at these locations. Hence identification of these locations where accident frequencies are high and further analyzing them is very much beneficial to identify the factors that affect the accident frequency at these locations.

Cluster analysis which is an important data mining technique can be used as a preliminary task to achieve various goals. Karlaftis and Tarko 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. Ma and Kockelman 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 and negative binomial (NB) models 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.Chang and Chen analyzed national freeway-1 data from Taiwan using CART and negative binomial regression model. Abellan et al. analyzed two lane rural highway data of Granada, Spain using decision rules extracted from decision tree method. Depaireet al. applied latent class clustering on two road user traffic accident data from 1997to 1999 of Belgium which divides the accident

data into seven clusters. Rovsek et al.analyzed crash data from 2005 to 2009 of Slovenia with classification and regression tree(CART) algorithm. Kashani et al. uses CART to analyze crash records obtained from information and technology department of the Iran traffic police from 2006 to 2008.

III. TRAFFIC ACCIDENT DETECTION

Passenger safety is one of the essential prospects in an ITS. Studies demonstrate the use of active safety equipment, which complements the conventional passive ones. Advanced driver assistance systems and collision avoidance systems are a result of such studies. The main focus in these systems is to supply reliable information about traffic incidents around the vehicle to road users, to use systems like adaptive cruise control and collision avoidance. Inter vehicle spacing, relative speed, lane change tracking, and inter vehicle time gaps are microscopic traffic variables that have been utilized for anomaly detection. In a vehicular-based networking and computing grid, it was reported that location information of each vehicle can solely be used to find out if the vehicle is in a queue and propagate information to neighboring vehicles. Another recent implementation named wireless local danger warning (WILLWARN) utilizes on-board equipment to measure microscopic variables (e.g., wheel speed, reduced friction) to discover possible dangers. However, the information is mainly disseminated to other vehicles as hazard warnings. Some of the recent anomaly detection systems include Vehicle Infrastructure Integration with Support Vector Machine (VII-SVM), Vehicle Infrastructure Integration with Artificial Neural Network (VII-ANN), and Notification Of Traffic In CidEnts (NOTICE). They use lane-changing characteristics and speed profiles of each vehicle. However, to obtain such fine-grained information, these systems need a specific road side infrastructure that is equipped with sensors and wireless transceivers mounted appropriately on each road section and/or on each lane.

Besides all these studies that focused on analyzing road accident data and identifying factors that affects severity of road accident, trend analysis of road accident data can also be useful to understand the nature of road accidents in certain locations. Time series data consists of a set of data points or values which have been measured on a certain fixed interval of time. Time series data is very important and useful to understand the nature of trend in different application such as detecting weather trend and forecasting stock market trend over a period of years. This is the motivating factor of this study. In this study, we have distributed 1 year road accident counts into 12 slots. Each slot is representing the total number of road accident that has occurred in 2 h of slots. More specifically, we have divided 24 h in 12 slots with 2 h in each slot and in time series data, slot1 is representing the total number of road accidents occurred in between 00:00 a.m. and 2:00 a.m. in 1 year period. So, we have a total of 60 counts for 5 year duration in our time series data. In order to analyze this data, we are using hierarchical clustering on all 26 time series data. The problem with hierarchical clustering of time series data is that it is quite difficult and unusual to manually decide the distance metric to be used with clustering algorithm. The wrong selection of distance metric certainly results in bad clusters. Our approach is fairly deal with this problem. Therefore, our method can be applied prior to clustering of data to find the best suitable distance metric for clustering. Clustering on time series data and trend analysis of each cluster shows that all the time series objects in each cluster having similar patterns. Hence, in order to perform trend analysis of time series data from several locations, our approach is suitable to apply prior to start trend analysis of road accidents. The results proves that our approach is capable to put all locations that have similar accident patterns in one cluster, that will definitely ease the difficulty in handling road accident time series data of different locations together.

IV. PROPOSED METHODOLOGY

A. Density-Based Clustering Technique

DBSCAN [9] is a density-based clustering algorithm. High-density regions represent clusters while regions with less density of points represent noise or anomalies. This algorithm is designed to overcome large data sets with noise and is capable of determining different sizes and shapes. DBSCAN is a density-based spatial clustering algorithm and 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, DBSCAN divides it into segments of clusters and a set of noise points. The density threshold condition is that there should be at least MinPts 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.

DBSCAN Algorithm

Below is the pseudo code, organized as functions for our purpose. I have picked it up from Wikipedia since it is in sync with what I have explained. The function regionQuery () returns the points within the n-dimensional sphere. The

function expandCluster () expands the cluster for each of the points in the sphere.

```
DBSCAN (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).
```

B. Parallel Frequent Association Mining

Association rule mining 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 FP-growth algorithm transforms the problem of finding lengthy frequent patterns to penetrating for smaller ones recursively and then combines the suffix. It uses the smallest frequent items as a suffix, giving fine selection. The procedure substantially minimizes the search costs and straight extracts the frequent patterns. The FP growth algorithm is presented in Figure 2 [49].

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

Figure 2 frequent Pattern growth algorithms

There are several clustering algorithms 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. DB Scan clustering techniques, after that we can use FP growth algorithm of association rule mining for processing the clusters.

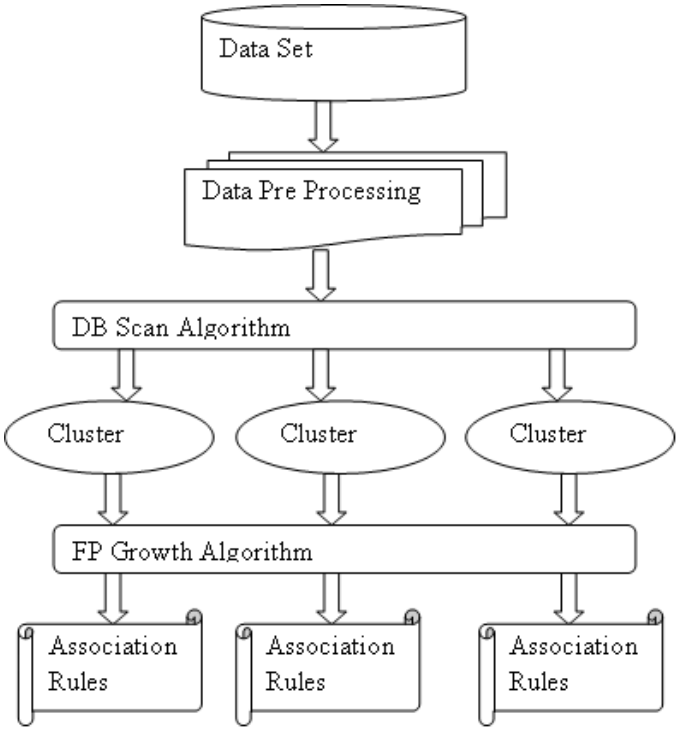

Figure 3. Framework of Proposed Method

V. RESULTS AND DISCUSSION

Cluster analysis

The basic requirement for cluster analysis is to determine the number of clusters to beformed by clustering algorithm. To achieve the solution for this, we used several informationcriteria such as Akaike Information Criteria (AIC) Bayesian InformationCriterion (BIC) and Consistent AIC (CAIC)]. We generated 20 models for 1cluster to 15 clusters. Figure illustrates the evolution of BIC, AIC and CAIC for the 20 models generated. It shows that there is a reduction in the values of AIC, BIC and CAICwith an increase in the number of clusters. Based on the Figure, we select the model with 6 clusters as there is no improvement after this. Ourselection also follows the approach used by previous studies.After getting number of clusters to be made, we used FP growth algorithm using Rstatistical software to segment the accident data set. After getting appropriate segmentationof the data set, our next task is the characterization of each cluster.

Cluster 1 (C1), It consists of 79 % of two wheeler accidents which are distributed on intersections nearmarkets, hospitals, local colonies across highways and non-highway roads. Those accidentswhich 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 hillarea, forest area or agriculture land area. In this cluster 74 % of accidents involved morethan two injuries and 26 % accidents involved 1 injury and rest involved more than injuries.Cluster 3 (C3), It consists of all accidents which were due to vehicle falling down from height. Around85 % of these cases are critical where ARA was hill. Rest of the accidents of this category belongs to non-critical injury. About 78 % of these accidents involved more than twoinjuries and rest accidents were two injuries involved.Cluster 4 (C4), It consists of accidents involving multiple vehicle accidents and divider hit/fixed objecthit cases. The accidents that are mostly happened in night time on highways are criticalaccidents whereas accidents at other locations such as market, colonies at night time arenon-critical in this cluster. In the same way clusters can be generate the performance based on Fp growth algorithm.

VI. CONCLUSION

Road and traffic accidents (RTA) are one of the important problems in India. Now a day, traffic safety is a major concern for transportation governing agencies as well as ordinary citizens. In order to give safe driving suggestions, careful data analysis of roadway traffic data is critical to find out variables that are closely related to fatal accidents. Various data mining techniques such as association rule mining, classification and clustering are widely used for the analysis of road accidents. Different studies carried out in this field linked the risk of accidents, the percentage or the number of accidents, the number of fatal accidents (dependent variables) to different factors or explanatory variables (independent variables) such as: age, and/or sex of the driver, expert or inexpert drivers, speed, length of the network, use of the safety belts, meteorological conditions and so on. In this study, we find factors behind road traffic accidents using data mining algorithms including DBSCAN cluster and Parallel Frequent mining algorithm. First, DBCSAN clustering technique is used as a preliminary task for segmentation of road accidents. Next, Parallel Frequent mining are used to identify the various circumstances that are associated with the occurrence of an accident for the clusters identified by DBSCAN clustering algorithm. UKDA data set is used and implementation is carried by using Weka tool. The key objective of data analysis is to identify the main problems in the field of road safety. The efficiency of accident prevention depends significantly on the reliability of the collected and estimated data and the appropriateness of the used methods. The results reveal that the combination of DBSCAN clustering and parallel frequent mining explore the accidents data recorded by the police information system, and discover patterns and predict future behaviors and effective decisions to be taken to reduce accidents.

REFERENCES

[1] Durrant-Whyte H, Bailey T. Simultaneous localization and mapping: Part I. IEEE Robot Autom Mag 2006; 13: 99–110.

[2] Sheu JB. A sequential detection approach to real-time freeway incident detection and characterization. Eur J Oper Res 2004; 157: 471–485.

[3] Chen A, Khorashadi B, Chuah C, Ghosal D, Zhang M. Smoothing vehicular traffic flow using vehicular-based ad hoc networking and computing grid (vgrid). In: Proceedings of the 9th International Conference on Intelligent Transportation Systems (ITSC); 17–20 September 2006; Toronto, Canada. New York, NY, USA: IEEE. pp. 349–354.

[4] Abuelela M, Olariu S, Weigle MC. NOTICE: architecture for the notification of traffic incidents. In: Vehicular Technology Conference; Spring 2008. pp. 3001–3005.

[5] Thajchayapong S, Barria JA. Anomaly detection using microscopic traffic variables on freeway segments. In: Transportation Research Board 89th Annual Meeting; 2010; Washington, DC, USA. pp. 695–704.

[6] Mitropoulos GK, Karanasiou IS, Hinsberger A, Aguado-Agelet F, Wieker H, Hilt H J, Mammar S, Noecker G. Wireless local danger warning: cooperative foresighted driving using intervehicle communication. IEEE T Intell Transp 2010; 11: 539–553.

[7] Barria JA, Thajchayapong S. Detection and classification of traffic anomalies using microscopic traffic variables. IEEE T Intell Transp 2011; 12: 695–704.

[8] Ma Y, Chowdhury M, Sadek A, Jeihani M. Real-time highway traffic condition assessment framework using vehicleinfrastructure integration (VII) with artificial intelligence (AI). IEEE T Intell Transp 2009; 10: 615–627.

[9] Ester M, Kriegel H, J¨org S, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the KDD; 1996. pp. 226–231.

[10] Arlia D, Coppola M. Experiments in parallel clustering with DBSCAN. Management. In: Euro-Par ‘01 Proceedings of the 7th International Euro-Par Conference; 2001; Manchester, UK. pp. 326–331.

[11] Aron M., Biecheler M.-B., Hakkert S., Peytavin J.F. Headways, rear-end collisions and traffic: the case of French motorways, 7’” Mernarkmd Conference on Trajic Safety on Two Continents, 1997.

[12] Broughton J., Markey, K.A. In-car equipment to help drivers avoid accidents, TRL Project Report 198: Transport Research Laboratory, Crowthorne, 1996.

[13] Broughton, J. A study of causation factors in car accidents, Road Safety In Europe Conference, Birmingham, 9-11 September, 1996.

[14] Broughton, J. A new system for recording contributory factors in road accidents, 7’h international Conference on Traf/ic Safety on Two Continents, 1997.

[15] Ernvall T. Risks exposures and accident data, Th lntemational Conference on Traftk Safety on Two Continents, 1997.

[16] MORTH (2014) Road Accidents in India 2013. New Delhi:Ministry of Road Transport and Highways Transport ResearchWing, Government of India, August 2014. http://morth.nic.in/showfile.asp?lid=1465. Accessed 20 May 2015.

[17] Kononov J, Janson BN (2002) Diagnostic methodology for thedetection of safety problems at intersections. Transp Res Rec.doi:10.3141/1784-07.

[18] Lee C, Saccomanno F, Hellinga B (2002) Analysis of crashprecursors on instrumented freeways. Transp Res Rec. doi:10.3141/1784-01.

[19] Chen W, Jovanis P (2000) Method for identifying factors contributingto driver-injury severity in traffic crashes. Transp ResRec. doi: 10.3141/1717-01.

[20] Barai S (2003) Data mining application in transportation engineering.Transport 18:216–223. doi:10.1080/16483840.2003.10414100.