A Review and Forestalling Road Accidents using Machine Learning

Sudhir Chitnis
dept. of Computer Science
Vishwakarma College of Arts, Commerce and Science
Pune,India
surusudhir21@gmail.com

Dr. Prasad Gokhale
dept. of Computer Science
Vishwakarma University
Pune, India

Dr.Neha Deshpande
dept. of Electronics Science
A. Garware College
Pune, India

Dr. Arvind Shaligram
dept. of Electronics Science
S. P. Pune University
Pune, India

Vaishali Kale
dept. of Computer Science
Vishwakarma College of Arts, Commerce and Science
Pune,India

Swati Patil
dept. of Computer Science
Vishwakarma College of Arts, Commerce and Science
Pune,India

Manali Jain
dept. of Computer Science
Vishwakarma College of Arts, Commerce and Science
Pune,India

Khushboo Oswal
dept. of Computer Science
Vishwakarma College of Arts, Commerce and Science
Pune,India

Harshada Raut
dept. of Computer Science
Vishwakarma College of Arts, Commerce and Science
Pune,India

Abstract— Traffic situation all over the world is very chaotic now-a-days which leads to road accidents. The reasons of accidents are different from one location to another. The frequencies of accidents also differ from time to time in the same location. As a result, road traffic accident data are most of the times heterogeneous in nature. However, data analysis plays a major role in recognizing the causes of increasing road accidents. The reason of analysis of such accident is performed on a subset of data which can produce many hidden relationships. Data mining techniques such as Clustering and Bayesian theorem are extensively used in the analysis of road accident data. As a result, this study proposes a framework which can take an overview of various clustering techniques used for road traffic accident analysis. Bayesian method tests heterogeneous data and can make multiple decisions. The research work carried out by different researchers based on road traffic accidents using various approaches has been discussed. This article consists of collections of methods in different scenarios with the aim to resolve the road traffic accident. The methods proposed in this study are looking like useful in some ways to decrease the number of fatalities. And the suggested approach will be helpful in different accident scenarios where the road accident and non-fatality cases are leading to fatality of life. It will confer a better approach to different parts of the country.

Keywords—Data Analysis, Data Mining, Clustering, hybrid k-mode, Bayesian theorem, Rule Mining, Machine Learning.

I. INTRODUCTION:

Road traffic accidents are becoming a silent disaster hence road safety is a national concern. Every year thousands of road accidents happen in India, which results in serious injuries and losses. The most common reason of road accidents is narrow and improper planning of road development, illegal parking, growing population and inadequacy of traffic police. Approximately 1.35 million people die each year as a result of road traffic crashes. Projections indicate that road traffic fatalities will be the fifth leading cause of death by the year 2030 unless urgent action is taken to address the issue [1]. For that the new agenda adopted globally is to half the number of deaths and injuries caused due to road traffic crashes. Due to easily available loan, there is increase in vehicle trade. The average number of vehicles in India is growing at the rate of 10.16% annually, over the last few years [2]. Consequently, traffic conditions in India are getting worse day-by-day creating unnecessary conflicts and exchange of arguments resulting in congestion and collision. The injuries due to road accidents become significant economic losses to individual, their families and to nation as a whole**.** Road traffic crashes cost most countries 3% of their gross domestic product [3].

With the aim of reducing injury of an accident, road traffic accident analysis which will give more precision to the result and prediction, and now becomes an area of active research. The research will be helpful in various accident scenarios, as it is creating massive destruction as shown in Table I.

Table I

Disaster (Natural and Man-made)	***Death***	***Injured***
Bhopal gas tragedy, India 2-3 Dec.1984	20,000	5,30,000
Latur (Killari) earthquake, India, 30 Sep.1993	9000	20,000
Orissa super cyclone, India, 29-30, Oct.1999	20,000	NA
World Trade Centre (9/11), USA, 11 Sep. 2001	3,000+	NA

Bhuj (kuchch) earthquake, India, 26 Jan. 2001	13,800	166,800
Asian Tsunami, many countries, 26 Dec.2004	245,000	1.0 million
Sichuan earthquake, China, 12 May.2008	90,000	375,000
Road accidents in (India), 2007	115,000per year	More than 0.5 million pe

Fig 1: Road Accidents- A Silent Disasters [4]

The data obtained from analysis of road accidents is of heterogeneous type therefore, a technique is used to sort the data from large data set which identifies the pattern i.e. Data mining technique. The fundamental operation of data mining is to separate objects into group of clusters. And, for road accident scenario, data mining technique is easy to use and reliable. It analyzes traffic accident severity problem and find factors behind them.

We need a technology that updates itself with the present data that is Machine Learning which is an application of Artificial Intelligence. Machine Learning is a category of algorithm that allows software application to become more accurate in predicting outcomes without being explicitly programmed [9]. It is divided into two micro-areas: 1. Supervised: It means inferring a function from labeled training data used to map a new set of data. Example: Classification and Regression, 2. Unsupervised: It means inferring a function to describe hidden structures from unlabeled data. Example: Clustering.

Clustering is collection of items that belongs to similar groups. The Clustering is divided into two i.e. Hierarchical and Partitional as shown in Fig.2. Partitional Clustering is a division of set of data objects into non-overlapping clusters. In Partitional Clustering, K-means partition data into K distinct clusters based on distance from the centroids of a subset. It has been used in road accidents but is not an ideal method for categorical data. Therefore, K-modes clustering is used over K-means algorithm. K-modes algorithms extends k-means paradigm to cluster categorical data. The performance of K-means and K-modes is enhanced by using hybrid K-modes algorithm. The advantage of using hybrid K-modes cluster is to reduce the clustering time and execution time. For certainty field of data, hybrid K-modes is ideal.

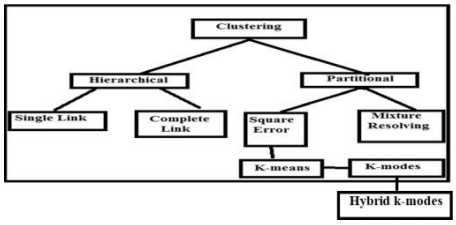

Fig. 2: Types of Clustering

Bayesian networks can represent and solve decision problems under uncertainty. Whereas for the uncertainty field, Bayesian method is used as it is one of the most effective theoretical models for knowledge representation and reasoning. In Bayesian method, data mining technique is applied which is supported by Bayesian Probabilistic Networks in order to represent non-linearity between risk indicating and model response variables, as well as different types of uncertainties which might be present in the development of the specific models. [5]

II. LITERATURE REVIEW

The paper entitled" A data mining framework to analyze road accident data", by Sachin Kumar and et.al in the year 2015 focused on accident data analysis associated with road traffic accident. The data set consists of more than 11,000 road accident records from 2009 to 2015 i.e. 6 years, in Dehradun District of Uttarakhand State. As per the researcher, road accident data makes the analysis task difficult as it is heterogeneous in nature. Data segmentation has been used widely to overcome this heterogeneity of the accident data [6].

In this paper, Clustering analysis is proposed framework that uses K-modes clustering technique in combination with Latent Class Clustering (LCC) followed by association rule mining. This data can be grouped into different homogenous segments. It helps in removing heterogeneity to some extent in the road accident data. Thus, association rule mining is further applied to cluster as well as on entire data set (EDS) so as to generate more appropriate rules. They included and worked on 11 attributes. Using Apriori algorithm accident prone circumstances can be identified and trend analysis can be performed for each cluster and on EDS.

We are working on the three main attributes which are forbidden by the researcher those are: vehicle condition, driver profile and road conditions.

The paper entitled "Modeling of road traffic fatalities in India: Accident Analysis and Prevention", by Rahul Goel published in the journal Elsevier in March2018, focused on six different modes of transport for injury modeling. The dataset includes all the road accidents records from 2010-2012.

The researcher mainly focuses on the 8 passenger modes in India i.e. walk, cycle, car, two-wheelers, and Intermediate Public transport modes (IPT) such as three-wheeled auto rickshaws, bus and train. A Poisson-log normal mixture regression model is developed to explore the relationship of road deaths of all road users with commute travel distance by different modes on road. The researcher established a relation between the road death and travel modes. Researcher includes four regression models which vary from minimal to maximal controlled.Model1 includes only commute distance variables, Model 2 adds diesel consumption,Model3 has additional adjustment of length of National Highways and Model4 also includes population density and percent urban population. This estimates of the Bayesian modeling in the form of posterior distribution of all parameters -Coefficient as well as error terms. Means and standard deviations are the major form of representation of results of these distributions. K-means clustering method is use to classify states into group of clusters with smaller distribution of mode share. State classification was into 5 clusters. The researcher considered eight attributes. The researcher shows that high risk of road accidents is caused by 2 wheelers and 4 wheelers whereas cycle, walk and IPT are at lower risk. The researchers suggested about bringing the change in mode from 2W to cycle or walk. Comparison shown of only two modes at a time is unacceptable. Hence, in complex scenarios such approach would not be recommended.

III. METHODOLOGY USED

a) K-modes Clustering:

K-mode clustering method is an enhanced version of k-means algorithm. It can be applied for categorical data. Distance measure and the clustering process is the major expansion in k-mode algorithm. The formula used for measuring distance and its procedure is explained below:

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

$$d(X,Y)=\sum_{i=0}^{N} \delta(Xi,Yi)$$

$$\text{where, } \delta(Xi,Yi) = \begin{cases} 0, X \\ 1, Y \end{cases}$$

- K-mode clump procedure:

  To form K-clusters from the data set D K-mode algorithmic program performs the following steps:

  1. Generate k clusters by arbitrarily selecting data objects and choose k initial cluster center.
  2. Observe the gap between each object.
  3. Assign data object to the cluster whose distance to the cluster center is minimum.
  4. Update the k cluster base on allocation of data objects and calculate k latest modes of every one cluster.
  5. Repeat step 2 to 3 awaiting no data object has changed cluster relationship otherwise some additional predefined criterion is fulfill.

b) Hybrid K-modes:

- Form good quality cluster
- Don't have to calculate distance repeatedly, which makes it less time consuming as well as easy to implement.
- Execution and clustering is less-time consuming

c) Bayesian Method:

Prediction, diagnosis, classification, and other tasks can be achieved by using learning and statistical inference functions of Bayes theorem. It is a directed acyclic network topology consisting of node set and directed edge, and each node denotes one variable state, while directed edge denotes the dependence between variables [7].

$$P(X) = \prod_{i=1}^{n} (Xi|pai)$$

A set of variables $X = \{X_1, X_2, \ldots, X_n\}$ of Bayesian network consists of the following components[37] S is a network structure which denotes the conditional independent assertion in variable set X, P is a set of local probability distribution associated with each variable, X_i denotes the variable node, and pa_i denotes the father node of X_i in S.

S and P defines the joint probability distribution of X. S is a directed acyclic graph (DAG), and each node in S corresponds to a variable in X. the default arc between nodes of S conditional independence. Use 'P' to denote the local probability distribution in (1), namely, the product term $p(X_i \mid pa_i)$ $(i = 1,2,\ldots,n)$; then the binary group (S,P) denotes the joint probability distribution p(X).

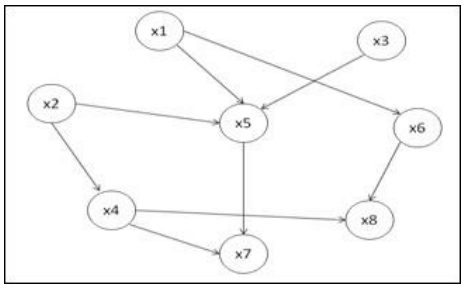

Fig. 3: Directed Acyclic Graph

Clustering is done using K-modes, hybrid k-modes and Bayesian and then Rule mining is applied on basis of:

- Severity of accident: Accident type is of critical(major)/ non-critical(minor)
- Time of Day: The accident took place at what time of day i.e.T1[0-4]….T6[20-24]
- Road Type: it happened on a dry/wet road
- Driver profile: If it took place due to driver's inattention or was drunk

IV. OBJECTIVES

- To focus on taking precautions before an accident happens.
- To analyze different aspects those are responsible for road accidents.
- To use K-Mode and Bayesian method for accurate severity detection.
- To identify Black spot (accident-prone areas), considering the location, hour of the day and type of accident (Table II) by following the data provided by NHAI.
- To identify hazardous locations that will decrease the risk factor of road accidents.
- To identify the probable source of crisis based on the accident.

V. DATABASE

We have considered the recent data given by the National Highway Authority of India (NHAI) that is from the year 2012-2018. Categorical data of all major cities in India is been taken.

TABLE II

Fatal Accident based on "Type of Accident":

Nov-15 to Aug-16			**Feb-17 to May-17**		
Nature of Accident	***Number of Accidents***	***%of Fatal Accidents to the Total fatal Accidents***	***Nature of Accident***	***Number of Accidents***	***% of Fatal Accidents to the Total fatal Accidents***
Rear End	42	73.3%	Rear End	7	50.0%

Collision			Collision		
Others	7	12.3%	Others	5	35.7%
Head-on Collision	5	8.8%	Over turning, Skidding	1	7.1%
Over Turning	2	3.5%	Head-on, side swipe, right turn collision	0	0.0%
Skidding	1	1.8%	-	-	-
Collision Brush/ side swipe	0	0.0%	-	-	-
Right Turn Collision	0	0.0%	-	-	-
Left Turn Collision	0	0.0%	-	-	-
Total Accidents	**57**	**100%**	**Total Accidents**	**14**	**100%**

Fig. 4: Accident Severity by Type of accident

In the above Table II, the type of accident is considered where maximum of 42 and 7 fatal accidents (73.3% and 50% of the total fatal accidents) have occurred due to Rear End Collision. Set of rules are formed which can find out in which circumstances rear-end collision can be encountered so as to decrease in death rate .This can done by giving training to the data provided by NHAI. On the basis of the rule mining, location that has rear end collision can be acknowledged as black spot area.

VI. PROPOSED SYSTEM

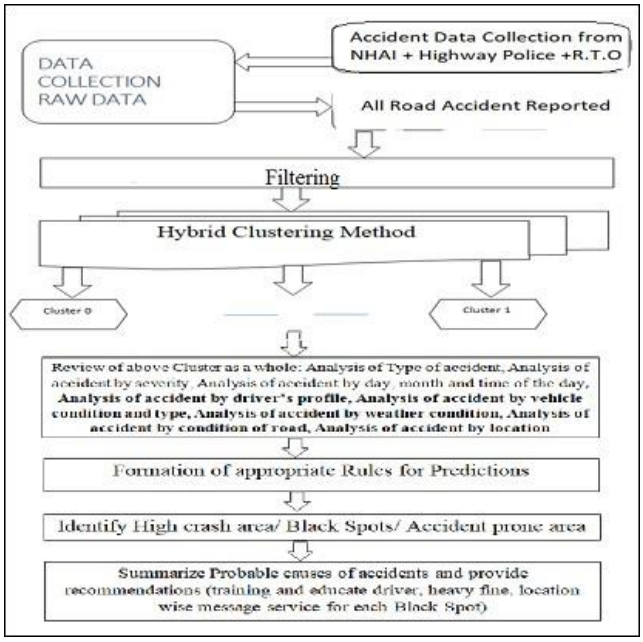

Fig 5: Proposed Flow

Road accidents cannot be totally prevented but by suitable traffic engineering and management the accident rate can be reduced to a certain extent [8]. Hence, the proposed system mainly aims to prevent accidents by taking into consideration all the important factors that affects the increased rate of accidents. Thus, the system collects the data from National Highway Authority of India (NHAI), as well as from highway police and R.T.O. to analyze the causes of road accidents in a proactive manner. Prescriptive clustering analysis is performed on all the reported accidents. This method takes into consideration the following attributes such as type of accident, accident by severity, accident by day, month, time, accident by driver's profile, accident by vehicle condition and type of the vehicle, accident by weather conditions, accident by condition of road, and accident by location for providing better recommendations. On the basis of all the above attributes, some rules are formed which identifies 'Black spots'. The system also summarizes the most probable causes of accidents and provides better recommendations for forestalling road accidents the average ratio of road accidents may diminish.

VII. RESULT AND DISCUSSION

To identify severity of an accident attributes such as: AGE: Child, Young, Adult, Gender: M/ F, TOD: (time of day) T1 [0–4]... T6 [20-24], MON: (Season) Winter/Rainy/ summer, LOR (lighting on road), ROF (roadway feature): Intersection/ Slope/Curve/Normal/Unknown, RTY (road type), LOC (Location), EC (Environmental Conditions): Fog/ rain/temperature are considered.

All the above attributes are combined from various research studies, in addition to these by considering Indian scenario some more vital attributes such as the DP (Driver Profile),VC(Vehicle Condition), RC(Road condition).

We have done rule mining by using these attributes. For example: Rule: If (LOR=DLT) && (LOC=MAR) then (SOA=NC). The rule suggests that light on road is 'day-light' and location is 'market' area where people usually drive slowly then severity of accident will be non-critical. In this way we can form different set of rules for critical and non-critical for better predictions according to the recommendation.

Recommendations such as: Provide message service system to driver, give alert messages on dashboard about black spot, hazardous locations should be exhibited; hilly area should not only have warning boards, but also appropriate rumbling speed and black spot area alert. A traveller should watch two minute video to decrease possible threats before moving on to the destination. Also every driver must follow the given instructions dutifully and if anyone does not obey then there should be heavy penalty or their license should be held for 3-6 months and to get it back one must start the procedure for renewal.

All these efforts are not just to analyze but to spread awareness for forestalling accidents, and to reduce them to a considerable amount.

ACKNOWLEDGMENT

This work was supported by the National Highway Authority of India (NHAI). I am thankful to Principal Dr. Arun Patil for his special guidance and support.

REFERENCES

[1] World Health Organization Report on “INJURIES and VIOLENCE”, pp 1

[2] https://www.nhai.org/roadnetwork.htm.

[3] World Health Organization Report on “Road Traffic Injuries”, 7 December 2018

[4] ROAD ACCIDENT SCENARIO IN INDIA AND ABROAD, Inception report-National Highway Authority of India, pp.4

[5] M.Deublein, M. Schubert , Bryan T.Adey, Jochen Köhler, Michael H., “Faber Accident analysis and prevention "Prediction of road accidents: A Bayesian hierarchical approach" Volume 51, March 2013

[6] Sachin Kumar, Durga Toshniwal "A data mining framework to analyze road accident data " Journal of Big Data, December 2015

[7] Xin Zou and Wen Long Yue "A Bayesian Network Approach to Causation Analysis of Road Accidents Using Netica" Journal of Advanced Transportation Volume 2017, Article ID 2525481

[8] Dr. Tom V. Mathew, Transportation Systems Engineering, NPTEL- Accident Studies, Chapter 42, Page No. 42.1, February 19, 2014

[9] Margaret Rouse, AWS analytics tool help make sense of big data, August 2015.