Collecting and Analyzing the data of road Accident

Priyanka G. Nikhade
Department of Computer Science and Engineering Rashtrasant Tukdoji Maharaj Nagpur University , Nagpur, India

***Abstract*- Traffic safety is one of the main priority all over the world. It is used to identify the causes and factors of road accidents where different approaches have been considered and has main aim to reduce the traffic accidents. This paper summarizes the work of applying data mining technologies to link recorded road characteristics to accident severity. The relationship between fatal rate and other attributes including collision manner, weather, surface condition, light condition, and drunk driver were investigated. It also predicts future behaviors and it also takes effective decisions to reduce accidents.**

***Index Terms*- Data mining, Classification, Decision tree, Naïve Bayes, Clustering, k-means Association, A-priori algorithm.**

I. INTRODUCTION

Traffic control system is the area, where critical data about the society is recorded and kept to be used safely. Using this data, we can identify the risk factors for vehicle accidents, injuries and fatalities and to make preventive measures to save the life of people. The severity of injuries causes an impact on the society. The main objective of the research was to find the applicability of data mining techniques in developing a model to support road traffic accident severity analysis in preventing and controlling vehicle accidents. It leads to death and injuries of various levels. A traffic crash happens due certain reasons like smashes of two vehicles on road, walking person, animal, or any other natural obstacles. It could result in injury, property damage, and death.

However, the accident is inevitable, given the definition, "an incident is defined as the sudden unintended release of or exposure to a hazardous substance that results in or might reasonably have resulted in, deaths, injuries, significant property or environmental damage, evacuation or sheltering in place".

To assist managers in strategic decisions, huge volumes of data are stored.Since the size of the database in terms of space and time quickly increases, analyse and extract useful information from them without the use of advanced data analysis tools has become a big challenge.

Data mining is a computational technique to deal with large and complex data set and these data sets can be of normal, nominal and mixed. It is quite easy to use in variety of domain belong to science and management.

Data mining is the extraction of hidden predictive information from large databases. It is a powerful new technology with great potential in their data warehouses. In recent years, researchers have been utilizing real-life data in studying various aspects of traffic accidents. So, measures are to be be taken to reduce accidents. It is important that the measures should be based on scientific and objective surveys of the causes of accidents and severity of injuries.

The basic hypothesis of this research is that accidents are not randomly scattered along the road network, and that drivers are not involved in accidents at random. There are complex circumstantial relationships between several characteristics (driver, road, car, etc.) and the accident occurrence.

II.LITERATURE REVIEW

Numerous studies have addressed the various aspects with most focusing on predicting or establishing the critical factors influencing injury severity.

Shetty et. al.,[1] used association rule mining to analyse the previous data and obtain the patterns between road accidents. The two criterion used for association ule mining are support and confidence. Apriori algorithm is one of the techniques to implement association rule mining. In the proposed system, we use apriori algorithm to predict the

patterns of road accidents by analyzing previous road accidents data.

Moradkhani [2] have been analyzing data to extract patterns using data mining algorithms. Data mining is implied, previously unknown and potentially useful data (rules and patterns) extraction process in databases.
Determining previous information, data cleaning and initial processing, constructing data warehouse, selecting target data set, finding used features and determining new features, visualizing data, selecting data mining operations, pattern extraction, evaluating and interpreting results and deleting inadequate patterns and finally and interpreting results of data and concluding from valuable information.

Cioca and Ivascu [3] identifies the causes of accidents, road safety performance indicators and risk indicators.
Analysing and evaluating the data lead to obtaining a framework for the improvement of the road safety system and reducing accidents, which is included in this research.

Jayasudha [4] analysed the traffic accident using data mining technique that could possibly reduce the fatality rate. Using a road safety database enables to reduce the fatality by implementing road safety programs at local and national levels. Those database schemes which describes the road accident via roadway condition, person involved, and other data would be useful for case evaluation, collecting additional evidences, settlement decision and subornation. The International Road Traffic and Accident Database (IRTAD), GLOBESAFE, website for ARC networks are the best resources to collect accident data. Using web data, a self-organizing map for pattern analysis was generated. It could classify information and provide warning as an audio or video. It was also identified that accident rates highest in intersections then other portion of road.
Chang and Chen [5] conducted data mining research focusing on building tree-based models to analyse freeway accident frequency. Using the 2001-2002 accident data of National Freeway 1 in Taiwan, the authors developed classification and regression tree (CART) and negative binomial regression models to establish the empirical relationship between traffic accidents and highway geometric variables, traffic characteristics, and environmental factors.

S. Krishnaveni, [6]work with some of classification models to predict the injuries happened in traffic accident in Nigeria's and compared Naive Bayes Bayesian classifier. This research is employed on the artificial neural networks based approach while the decision trees data analysis can be used to works on reduction of massacre on the highways. The data was classified in continuous and categorical data where continuous data analysed using artificial neural networks technique and the categorical data, using decision trees technique. The results reveal that decision tree approach outperformed the ANN with a lower error rate and higher accuracy rate. This research based on three most important causes of accident due to tyre burst, loss of control and over speeding.

Li, Shrestha, Hu [7] first calculated several statistics from the dataset to show the basic characteristics of the fatal accidents. We then applied association rule mining, clustering, and Na¨ıve Bayse classification to find relationships among the attributes and the patterns.

Sachin et, al., [8], proposed a framework for Dehradun, India road accident (11,574) happened during 2009 and 2014 by using K-modes clustering technique and association rule mining. The analysis of result using combination of these technique conclude that the result will be more effective if no segmentation has been performed prior to generate association rules .

Kumar & Toshniwan [9] proposed K-modes clustering technique as a preliminary task for segmentation, association rule mining are used to identify the various circumstances that are associated with the occurrence of an accident for both the entire data set (EDS) and the clusters identified by K-modes clustering algorithm. The findings of cluster based analysis and entire data set analysis are then compared. The results reveal that the combination of k mode clustering and association rule mining is very inspiring as it produces important information that would remain hidden if no segmentation has been performed prior to generate association rules. Further

a trend analysis have also been performed for each clusters and EDS accidents which finds different trends in different cluster whereas a positive trend is shown by EDS. Trend analysis also shows that prior segmentation of accident data is very important before analysis.

III.EXISTING METHODOLOGIES

The basis of the research is that accidents are not randomly scattered along the road network, and that drivers are not involved in accidents at random. There are various and complicated circumstantial relationships between several characteristics (driver, road, car, other people,etc.) and the accident occurrence. As such, one cannot improve safety without successfully relating accident frequency and severity to the causative variables. We will attempt to extend the area by generating additional attributes and focusing on the contribution of road-related factors to accident severity. This will help to identify the parts of a road that are risky, thus supporting traffic accident data analysis in decision making processes.

Data Collection and Preparation-

Collect, analyse and understand the content and structure of available data, is one of the most important things that need special attention. About this study, the data resources used in case of an accident is an accident report that day should be filled by traffic police officers. This information includes complete details about the accident. Good understanding of data can be lead to better success in achieving the goal of data mining.

Classification-

A classification is an ordered set of related categories used to group the data according to its similarities. It consists of codes and descriptors and allows survey responses to be put into meaningful categories in order to produce useful data. A classification is a useful tool for anyone developing statistical surveys. It is a framework which both simplifies the topic being studied and makes it easy to categorize all data or responses received. Decision tree analysis is one of the main techniques used in Data Mining

Clustering-

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data.

A loose definition of clustering could be “the process of organizing objects into groups whose members are similar in some way”. A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. K-nodes is one of the method which is used to cluster the analyzed road datasets.

Several clusters are made based on the attributes and are used to analyses the accidents and based on the factors that are

The traffic is low or high

Time of accident whether in the morning, evening, afternoon or night

Age of the person or number of people who are injured In which month accident occurred etc.

Association-

Descriptive or predictive mining applied on previous road accidents data in combination Other important information as weather, speed limit or road conditions creates an interesting alternative with potentially helpful and useful outcome for all involved stakeholders. It is used to analyse previous data and obtain the patterns with road accidents. The two criterion used for this method are support and confidence. Apriori algorithm is one of them.

One important feature of this method is the identification and treatment of accident prone locations commonly called black spots; black spots are not the only cause of accidents on the highway. Also, various organizations such as Police High Way Patrol, Vehicle Inspection Officer (VIO), Federal Road Safety Commission (FRSC) among others are charged with the responsibility of maintaining safety thereby reducing road accidents in the project “Analysis of data mining techniques for traffic accident severity problem:A review”.

However, lack of good forecasting techniques has been a major hindrance to these organizations in achieving their objectives.

IV.PROPOSED METHODOLOGIES

To analyze the data, the approach we take for our study follows the data analysis steps, as follows:

A. Data Collection and Preparation -
Collect, analyze and understand the content and structure of available data, is one of the most important things that need special attention. Good understanding of data can be lead to better success in achieving the goal of data mining.
Data preparation was performed before each model construction. All records with missing value in the chosen attributes were removed. All numerical values were converted to nominal value according to the data dictionary in attached user guide.
To make sure that our data preparation is valid, we have checked the correctness of attribute selection. There are several attribute selection techniques to find a minimum set of attributes so that the resulting probability distribution of the data classes is as close as possible to the original distribution of all attributes.
.

B. Data Pre-processing-
Data pre-processing helps to remove noise, missing values, and inconsistencies. Missing values are replaced with NULL so, each attribute data is discretized to make it appropriate for further analysis. Table 1 presents the data before and after transformation.
It is one of the crucial step or process in data mining techniques, which consists of several types of f preprocessing tasks like handling missing values, minimizing noises, dimensionality reductions, attribute aggregations, feature creation, discretization and binarization, attribute transformation, sampling and feature selection which mainly are guided by the data mining goal at hand.
Considering the whole objective of the experiment, the pre-processing task for this research can be consider as light weight pre-processing. The main reasons were the tool's capability of handling data quality issues like missing data and the need to expose the actual data as it is.

C. Training Dataset Description/ Attribute Selection
The next task of the experiment was to identify attributes or features related to the goal of the machine learning task which will obviously be evaluated by the machine learning process through attribute selection. The test dataset is used to validate the rules obtained from trained classifier using new data. Table I provides the list of attributes and their description.

D. Clustering Algorithm
The objective of clustering algorithm is to divide the data into different clusters or groups such that the objects within a group are like each other whereas objects in other clusters are different from each other. Hierarchical clustering technique (e.g. Ward method, single linkage, complete linkage, etc.), K means and latent class clustering (LCC) have been used in road accident analysis. Another clustering technique is K-modes clustering which is an enhanced version of K means algorithm. K-medoids and expectation maximization algorithms are used for clustering, and the following clusters are formed.

E. Classification Algorithm
A major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data. Hence, machine learning is closely related to fields such as artificial intelligence, adaptive control, statistics, data mining, pattern recognition, probability theory and theoretical computer.

1) Decision Tree-
A Decision Tree Forest (DTF) is an ensemble (collection) of decision trees, which the combination of predictions contributes to the overall prediction for the forest. Its ability to handle thousands of input variables without variable deletion. A decision tree forest grows a number of independent trees in parallel, and those trees do not interact until after all of them have been built. Decision tree forest models often have a degree of accuracy that cannot be obtained using a large, single-tree model.

i) CART-
Classification and Regression Trees model consists of a hierarchy of univariate binary decisions. CART operates by choosing the best variable for splitting the data into two groups at the root node, partitioning the data into two disjoint branches in such a way that the class labels in each branch are as homogeneous as possible, and then splitting is recursively applied to each branch, and so forth.

2) Naïve Bayes-

A Naive Bayesian classifier is a simple probabilistic classifier based on applying Bayesian theorem (from Bayesian statistics) with strong (naive) independence assumptions.

training can be done by evaluating a closed-form expression,[1]:718 which takes linear time, rather than by expensive iterative approximation as used for many other types Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihoodof classifiers.

Algorithm-

Step 1: Scan the dataset (storage servers)

Step 2: Calculate the probability of each attribute value. [n, n_c, m, p]

Step3: Apply the formulae P(attributevalue(ai)/subjectvaluevj)=(nc+ mp)/(n+m)

Where: n = the number of training examples for which v = vj nc = number of examples for which v = vj and a = ai

p = a priori estimate for P(aijvj) m = the equivalent sample size

Step 4: Multiply the probabilities by p

Step 5: Compare the values and classify the attribute values to one of the predefined set of class.

F. Association Rule Mining-

Association rule mining is used to analyse the previous data and obtain the patterns between road accidents. The two- criterion used for association rule mining are support and confidence. Apriori algorithm is one of the techniques to implement association rule mining. In the proposed system, we use apriori algorithm to predict the patterns of road accidents by analysing previous road accidents data.

• A-priori Algorithm- Algorithm-

Step 1: Scan the data set and find the support(s) of each item. Step 2: Generate L1 (Frequent one item set). Use Lk-1, join Lk- 1 to generate the set of candidate k - item set.

Step 3: Scan the candidate k item set and generate the support of each candidate k – item set.

Step 4: Add to frequent item set, until C=Null Set.

Step 5: For each item in the frequent item set generate all non- empty subsets.

Step 6: For each non-empty subset determine the confidence. If confidence is greater than or equal to this specified confidence. Then add to Strong Association Rule.

G. Weka Toolkit-

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. Since WEKA's explorer generally chooses reasonable defaults, the J48 decision tree algorithm was performed using its default parameters: a confidence interval of 0.25, pruning allowed, and a minimum number of objects for a leaf of 3. Training and testing were done using ten-fold cross-validation.

VI.PROPOSED WORK

A. Flow-Chart:

It consists of the data-sets through which all the data of the accidents of various regions are collected.Then comes login page into which if we have an account already existed then sign in or else create an account and then login and collect the databases present into it or else want to enter the new databases, enter it and then click on submit and click on pre-processing button which will pre-process and give an error if not submitted properly, then click on cluster button which cluster it into cities or regions wise and take a decision if the road is safe to drive or not.

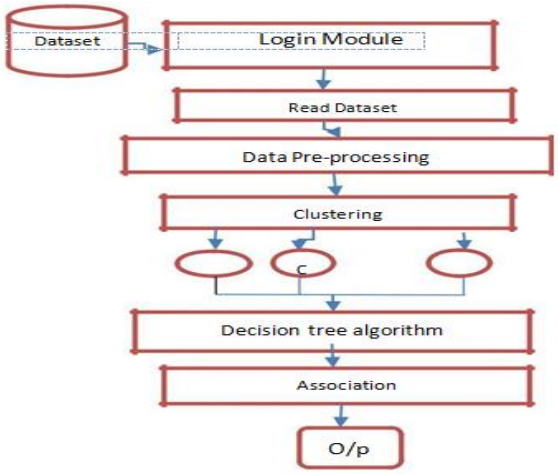

Fig 2- Flow Chart of the proposed work

And associate button will give the analysis about speed limit, weather condition when it is easy to drive or not safe to drive on the road and will give an output into the form of result.

B. Data Flow Diagram-

User will first login or create an account. Then login process will take place, datasets can be read or entered any new datasets into the particular region and pre-process it by clicking on pre-process and then clustering takes place and forms clusters according to the attributes of the regions.

Classification takes place through decision tree in which CART process takes place internally according to which gives the decision about the road that is safe to drive or not or accidents takes place

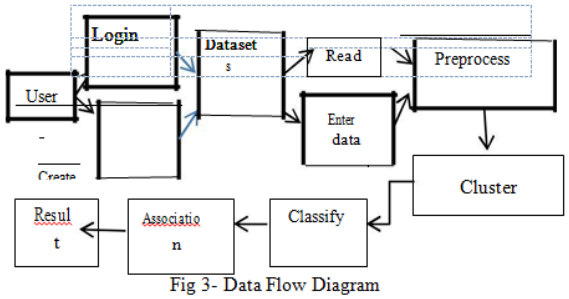

Fig 3- Data Flow Diagram

To avoid them, association takes place which gives confidence and support about the road conditions, weather conditions, speed limit and many more.

Then finally result analysis takes place which gives the analysis of the result by giving output.

VII. OUTCOME POSSIBLE RESULTS

The results of our analysis include association rules among the variables, clustering of states on their populations and number of fatal accidents, and classification of the regions as being high or low risk of fatal accident. We used a data analytic tool Weka to perform these analyses.

As seen in statistics, association rule mining, and the classification, the environmental factors like roadway surface, weather, and light condition do not strongly affect the fatal rate, while the human factors like being drunk or not, and the collision type, have stronger affect on the fatal rate.

Accuracy Measures-

In our study we used precision and recall as accuracy measures to evaluate the accuracies of classifiers. The high values of precision and recall denote high accuracy.

VIII.CONCLUSION AND FUTURE SCOPE

In this study it is tried to choose the interesting and superior rules to provide a lot of valuable information for policies to provide better safety policies. This article can be a step towards providing useful information for highway engineers and transportation designers to design safer roads. Percentage distribution of accidents on various criteria, speed limit and injury severity, distribution of accidents by time of accidents and deceased age, distribution of accidents by month and weather during the accident, distribution of accidents by lightness and speed limit, distribution of accidents by accident type (human factors), distribution of accidents by day of accident and deceased age, distribution of accidents by deceased emotions, distribution of accidents by hospital reported and ambulance used is also made.

Current system is manual where government sector makes use of ledger data and analyses the data manually, based on the analysis they will take the precautionary measures to reduce the number of accidents. Proposed system uses road accidents data to mine frequent patterns and crucial factors causing diverse types of accident. It discovers the associations among road accidents using A-priori algorithm. It also predicts the common accidents that may cause for new roads with the help of Naïve

Bayes algorithm.

In this paper, we proposed a framework for analyzing accident patterns for different types of accidents on the road which makes use of K modes clustering and association rule mining algorithm. The paper also discussing about various data mining techniques which is proved supporting to resolve traffic accident severity problem and conclude which one could be optimal technique in road traffic accident scenario. The brief survey will also help us to find better mining technique in this kind of problem.

Future work is to make analysis on road accidents' dataset by considering more features and clusters and to use deep learning techniques to better cluster the records.

REFERENCES

[1] Poojitha Shetty, Sachin P C, Supreeth V Kashyap, Venkatesh Madi, "Analysis of road accidents using data mining techniques", Volume: 04 Issue: 04 | Apr -2017

[2] Road Accidents in India Issues & Dimensions, Ministry of Road Transport & Highways Government of India (2014)

[3] K. Geetha and C. Vaishnavi, "Analysis on Traffic Accident Injury Level Using Classification", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 2, February 2015, ISSN: 2277 128X

[4] Farzaneh Moradkhani, Somayya Ebrahimkhani, Bahram Sadeghi Begham,"Road Accident Data Analysis: A Data Mining Approach ",Article in Indian Journal of Scientific Research May 2014

[5] Lucian-Ionel Cioca, ID and Larisa Ivascu, "Risk Indicators and Road Accident Analysis for the Period 2012–2016"

[6] K. Jayasudha & Dr. C. Chandrasekar,"An Overview of Data mining in Road Traffic and Accident Analysis",Journal of Computer Applications, Vol – II, No.4, Oct – Dec 2009

[7] Ghazi G. Al-Khateeb,"Analysis of Accident Data and Evaluation of Leading Causes for Traffic Accidents in Jordan" , Jordan Journal of Civil Engineering, Volume 4, No. 2, 2010

[8] Liling Li, Sharad Shrestha, Gongzhu Hu,"Analysis of Road Traffic Fatal Accidents Using Data Mining Techniques", SERA 2017, June 7-9, 2017, London, UK

[9] Luis Martín, Leticia Baena, Laura Garach, Griselda López and Juan de Oña,"Using data mining techniques to road safety improvement in Spanish roads", Luis Martín et al. / Procedia - Social and Behavioral Sciences 160 (2014) 607 – 614

[10] Meenu Gupta, Vijender Kumar Solanki, Vijay Kumar Singh,"Analysis of Datamining Technique for Traffic Accident Severity Problem: A Review",Proceedings of the Second International Conference on Research in Intelligent and Computing in Engineering pp. 197–199 DOI: 10.15439/2017R121 ACSIS, Vol. 10 ISSN 2300- 5963

[11] Sachin Kumar and Durga Toshniwal,"A data mining framework to analyze road accident data", Kumar and Toshniwal Journal of Big Data (2015) 2:26 DOI 10.1186/s40537-015-0035-y