"Analysis of Road Accidents using Apriori, Naive-Bayes and K-Means"

Neha Patil, Prof. Deepesh Jagadale

Abstract – Road accidents are the main cause of death as well as serious injuries in the world. India is among the emerging countries where the rate at which traffic accidents occur is more than the critical limit. As a human being, everyone wants to avoid traffic accidents and stay safe. In order to stay safe, careful analysis of road-way traffic accident data is important to find out factors that are related to fatal, grievous injury, minor injuries, and non-injury. The relationship between critical rate and other attributes include combining weather conditions, road type, sunlight conditions, speed limit, drunk driver and so on are considered. Here, data mining algorithms are applied on critical accident dataset to address this problem and predict the accident severity. Apriori Algorithm is used for finding an association between attributes. Naive based approach is used for classifying how attributes are conditionally independent. K-means are used to form clusters and analyze them based on attributes. Comparison based on parameters is done to prove the efficiency of the various road accident detection techniques and approaches. The comparison result shows the best road accident detection method. By using these statistics, government/private agencies can take decisions in developing new roads and taking additional safety measures for the general public and awakening a sense of responsibility of road users.

I. INTRODUCTION

1.1 Fundamentals

There are a lot of vehicles driving on the roadway every day and traffic accidents could happen at any time any where. Some accidents involve fatality, which means people die in that accident. As a human being, we all want to avoid accidents and stay safe. To find out how to drive safer, data mining techniques could be applied on the traffic accident dataset to find out some valuable information, thus give driving suggestions. Data mining uses many different techniques and algorithms to discover relationships in large amounts of data. It is considered one of the most important tools in information technology in the previous decades. Association rule mining algorithm is a popular methodology to identify the significant relations between the data stored in large databases and also plays a very important role in frequent itemset mining. A classical association rule mining method is the Apriori algorithm whose main task is to find frequent itemsets, which is the method we use to analyze the road-way traffic data. Classification in data mining methodology aims at constructing a model (classifier) from a training data set that can be used to classify records of unknown class labels. The Naïve Bayes technique is one of the very basic probability-based methods for classification that is based on the Bayes' hypothesis with the presumption of independence between each pair of variables.

- Neha Patil *is currently pursuing masters degree program in Msc.IT in Mumbai University ,India,9372304391. E-mail: pneha5055@gmail.com.*
- *Deepesh Jagadale is currently head of IT department in PHCACS, Rasayani in Mumbai University, India, 9028609874. E-Mail: djagdale@mes.ac.in.*

1.2 Objectives

The general objective is to investigate the role of road-related factors in accident severity using predictive models.
The primary objectives of this study are:-

- To identify risk factors related to road accident fatality.
- To Predict accident severity using different data mining techniques.
- To compare standard classification models for this task.

Understanding how the risk factors are related to the occurrence of an accident is useful because risk factors potentially play a vital role in the road safety and may help in taking appropriate measures in reducing vehicular accident fatalities and injuries.

1.3 Scope

This study will help the government, the general public, the Federal Road Safety Commission and other agencies concerned with safety on our roads in the following ways:-

- It will help the Federal Road Safety Commission and other authorities concerned with similar assignments to assess their performance over the years.
- It will help the Federal Road Safety Commission and other institutions concerned organizing sensitization workshops on seminars programmes for road users ascertain the positive impact of such workshops or seminars being organized.
- It will awaken the sense of responsibility of road users and government.

1.4 Outline

In this paper, the study of different domain techniques is presented. The different techniques such as Apriori, Naive-Bayes and K-Means are used. The comparative study of various techniques mentioned

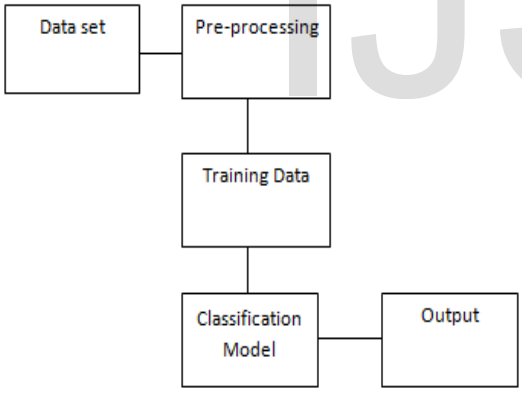
above is presented in this paper. The hybrid approach is proposed along with performance measures

II. Analysis of Road Accidents :

Overview :

In order to predict the pattern of new road accidents, an association and classification data mining techniques are used that is, Apriori, Naïve Bayes classifier and Kmeans, which are highly scalable. Even if we are working on a data set with millions of records with some attributes, this classifier can yield best results. There are models that assign class labels to problem instances, which are represented as vectors of feature values, and the class labels are drawn from some finite set. The data is collected from police stations which are restricted to an area. The below figure represents the architecture diagram for predicting the road accidents where a data repository is created based on the data collected from different police stations. Based on this uploaded data, the system predicts the patterns between road accidents.

Existing System Architecture:



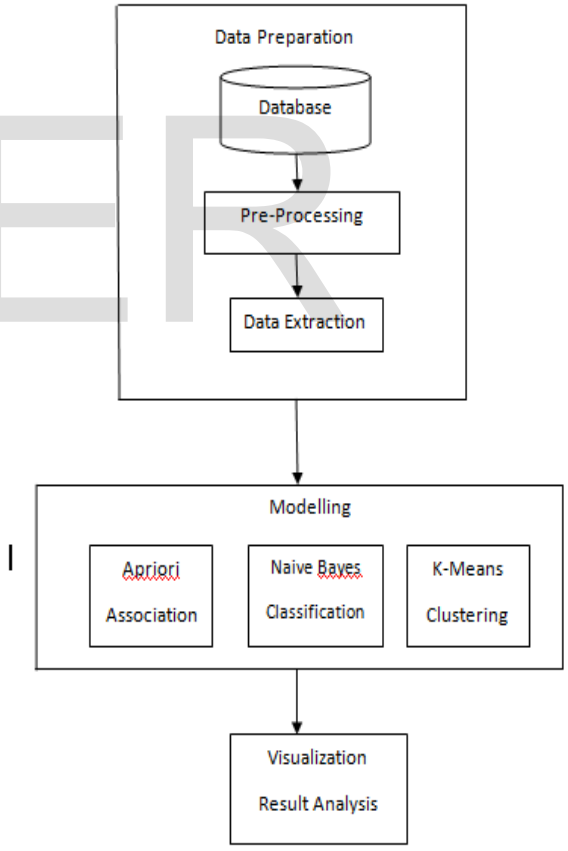
Existing System Architecture

1. **Data Set** : The datasets for training the classifier is provided or obtained from a source. The dataset consists of around thousands of samples of traffic attributes like Time of accident, area, speed, season, lighting conditions, etc. The data collected is not in the format required to analyze it, so the data must be preprocessed.It is important to eliminate missing values from the dataset since there were no values that could negatively impact the real performance.

2. **Data Pre-Processing** : Data preprocessing is one of the most significant functions in information mining. Data preprocessing primarily involves removing noise, handling missing values, removing irrelevant characteristics to make the information ready for assessment. In this step, the goal is to pre-process the accident information in order to make it suitable for assessment.

3. **Classification Modelling** : This section describes the methods selected classification methods for accident severity prediction. Then, the output is displayed, for a given set of attributes which predicts whether the accident would be critical or noncritical. Then, to obtain association the same data is subject to classification using Decision Tree. Accidents can be analysed Using this data. Naive Bayesian classifier based on Bayes rule is used to get the severity. Decision tree classifier is another classifier, which gives good results for accident severity calculation. Finally K-Nearest Neighbor (KNN) classifier is employed for severity calculation.The accuracy of the algorithms are compared and it is found that KNN performs better than the other two algorithms employed.

Proposed System Architecture:



Proposed System Architecture

1. Data Preparation : Data preparation is performed before each model construction. All records with missing value (usually represented by 99 in the dataset) in the chosen attributes are removed. All numerical values are converted to nominal value according to the data dictionary in the attached

user guide. Fatal rate is calculated and binned into two categories: High and Low.

2. Modeling : We first calculate several statistics from the dataset to show the basic characteristics of the fatal accidents. We then apply Apriori, Naïve-Bayes and K-Means to find relationships among the attributes and the patterns.

3. Result Analysis : 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 use a data analytic tool Weka to perform these analyses.

III. Implementation:

Algorithms:

1. Apriori Algorithm:

Descriptive or predictive mining applied on previous road accidents data in combination with other important information such as weather, speed limit or road conditions creates an interesting alternative with potentially useful and helpful outcome for all involved stakeholders. Association rule mining is used to analyze the previous data and obtain the patterns between road accidents. The two criteria 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 analyzing previous road accidents data.

The steps for the Apriori Algorithm:-

- Scan the data set and find the support(s) of each item.
- Generate L1 (Frequent one item set). Use Lk-1, join Lk-1 to generate the set of candidate k-item set.
- Scan the candidate k item set and generate the support of each candidate k – item set.
- Add to the frequent item set, until C=Null Set.
- For each item in the frequent item set generate all non empty subsets.
- For each non empty subset determine the confidence. If confidence is greater than or equal to this specified confidence .
- Then add to the Strong Association Rule.

2. Naive-Bayes Classifier:

Naive Bayes is a probabilistic classifier based on Bayes theorem. It assumes variables are independent of each other. The algorithm is easy to build and works well with huge data sets. It
has been used because it makes use of small training data to estimate the parameters important for classification.
Bayes Theorem states the following:-

1. P(c|x) = (|) ( )/ ( )
2. P(c|x) = P(1|c)*P(2|c)*...P(|c)*P(c)

Where P(c|x) is the posterior probability of class (target) given predictor (attribute). P(c) is the
prior probability of class. P(x|c) is the likelihood which is the probability of predictor given class.
P(x) is the prior probability of a predictor.

The steps for the Naive-Bayes Classifier:-

- Calculate the preceding probabilities for each attribute class.
- Calculate the likelihood of proof going into the denominator.
- Calculate the probability of evidence going into the numerator.
- Use the Bayers rule to calculate the probability of a specific attribute.

3. K-Means Clustering:

K-means is one of the simplest unsupervised learning algorithms that solve the well known
clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k-centers, one for each cluster. These centers should be placed in a cunning way because different locations cause different results. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early age group is done. At this point we need to re-calculate k new centroids as bary center of the clusters resulting from the previous step. After we have these k new centroids , a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more.

Finally, this algorithm aims at minimizing an objective function know as squared error function given by where:-

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\left\| x_i - v_j \right\| \right)^2$$

- '||xi - vj||' is the Euclidean distance between xi and vj.
- 'ci' is the number of data points in ith cluster.
- 'c' is the number of cluster centers.

The steps for the K-Means Clustering:-
- Let X = {x1,x2,x3,....., xn} be the set of data points and V = {v1,v2,.....,vc} be the set of centers.
- Randomly select 'c' cluster centers.
- Calculate the distance between each data point and cluster centers.
- Assign the data point to the cluster center whose distance from the cluster
The center is the minimum of all the cluster centers.

● Recalculate the new cluster center using:

where, ‘ci’ represents the number of data points in the cluster.

- Recalculate the distance between each data point and new obtained cluster centers.
- If no data point was reassigned then stop, otherwise repeat from step 3.

IV. Applications:

1. Improving Road Safety

The significance of this research lies in its development of new insights related to road accidents and these insights will provide valuable help in developing methods to improve road safety, particularly in the phase of choosing the appropriate means and budget allocations of resources. Considering the size of the accident data set, applying data mining techniques to model data records can help to reveal how the drivers’ behavior and roadway and weather conditions are causally connected with different injury severities. This can help decision makers to formulate
better traffic safety control policies, label roads with necessary signs informing drivers and pedestrians of accident risks, and design better roads.

2. Improving Vehicle Safety

Every year, vehicle manufacturers research, design, and develop strategies to help make vehicles better and safer.

Adaptive Headlights :

Using information such as steering wheel movement and vehicle speed, adaptive headlights are able to pivot in the direction you're traveling, helping you see the road ahead.

Blind Spot Monitoring/Blind Spot Detection :

When you're driving, vehicles behind or beside you are sometimes hidden in what's called a "blind spot." This can lead to a collision if you try to turn or change lanes. Blind spot monitoring systems visually alert you when a vehicle is traveling in your blind spot.

Front Crash Prevention :

Front crash prevention systems use forward-facing sensors to monitor distance and relative speed between vehicles. If the system senses an impending crash, it will alert you with sound, visual cues or physical sensations such as a vibration of the steering wheel. If you don't respond, some systems make adjustments to lessen the crash impact, or automatically brake the vehicle to prevent it.

Lane Departure Prevention :

A lane departure system, which often uses a camera near the rearview mirror, keeps track of your vehicle's position in a lane.

Park Assist and Backover Protection :

One or both of these systems may soon be required in most new vehicles. They help drivers avoid collisions when parking or reversing, using sensors in cameras to alert you of objects behind your vehicle. Some backover protection systems may automatically brake to avoid collisions.

V. Conclusion

Road accident detection is considered to be the contemporary ever growing process focused primarily to reduce death. Here, this study provides road accident detection techniques by analyzing the novel ideas. The analysis of these methods provides a better understanding of the steps involved in each process in a way of consequently increasing the scope for finding the efficient techniques to achieve maximum accurate performance. The comparison of the techniques used here, that is Apriori, Naive-Bayes and K-Means is carried out in terms of precision and recall. Environmental factors like roadway surface, weather, and light conditions do not strongly affect the fatal rate, while the human factors like being drunk or not, and the collision type, have a stronger effect on the fatality rate. From the clustering result we can see the states/regions which have a higher fatality rate, while some others lower. We should pay more attention when driving within these risky states/regions. Current system is manual where the government sector makes use of this data and analyzes it manually. Based on the analysis, they
will take precautionary measures to reduce the number of accidents.

VI. Reference:

[1] Analysis of Road Traffic Fatal Accidents using Data Mining Techniques (IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA)) Published by Liling Li, Sharad Shrestha and Gongzhu Hu in 2017

[2] Mining Techniques for Analysing Road Accidents (International Journal of Innovative Research in Computer and Communication Engineering, Volume 6) Published by P. D. S. S.Lakshmi Kumari and S. Suresh Kumar in 2018

[3] Analysis of road accidents using data mining techniques (International Research Journal of Engineering and Technology, Volume 4) Published by Poojitha Shetty, Sachin P. C., Supreeth V. Kashyap and Venkatesh Madi in 2017

[4] Analysing Road Accident Criticality using Data Mining (International Journal of Scientific Research in Computer Science, Volume 5) Published by Shahsitha Siddique and Nithin Ramakrishnan in 2019

[5]Survey on Analysis and Prediction of Road Traffic Accident Severity Levels using Data Mining Techniques (International Journal of Current Engineering and Technology, Volume 7) Published by Baye Atnafu and Gagandeep Kaur in 2017

[6]A Review on Road Accident Detection using Data Mining Techniques (International Journal of Advanced Research in Computer Science, Volume 9) Published by Arun Prasath N. and Dr.M.Punithavalli in 2018

[7]Road Accident Analysis System using Data Mining (International Journal on Future Revolution in Computer Science, Volume 4) Published by Chanchal Sharma and Anil Wadhwa in 2016

Apriori Algorithm

https://www.scribd.com/document/381999708/Review-On-Road-Accident-Analytics-Using-Data-Mining-Technique
Naive-Bayes Classifier
https://www.coursehero.com/file/p6mckd4/1421-Naive-Bayes-Classifier-Naive-Bayes-is-a-probabilistic-classifier-Based-on/
K-Means Clustering
https://towardsdatascience.com/applying-machine-learning-to-classify-an-unsupervised-text-document-e7bb6265f52