PREDICITION OF ROAD ACCIDENTS BY USING DATA MINING TECHNIQUES

VELPULA SUNDARARATNAM[1], M HARIKA[2], M JOYNISSY[3], P SWETHA4

ASSISTANTPROFESSOR[1], UG SCHOLAR[2, 3&4]

DEPARTMENT OF CSE, MALLA REDDY ENGINEERING COLLEGE FOR WOMEN,MAISAMMAGUDA, DHULAPALLY KOMPALLY, MEDCHAL RD, M, SECUNDERABAD, TELANGANA 500100

ABSTRACT Road traffic accident are considered as major public health concern. In order to give safe driving suggestions, careful analysis of road traffic data is critical to find out variables that are closely related to fatal accidents. In this paper we apply Probability analysis and data mining algorithms on FARS Fatal Accident dataset as an attempt to address this problem. The relationship between fatal rate and other attributes including collision manner, weather, surface condition, light condition, and drunk driver were investigated. Classification model was built by Naive Bayes classifier, and clusters were formed by simple K-means clustering algorithm. Certain safety driving suggestions were made based on probability, classification model, and clusters obtained.

Keywords Roadway fatal accidents, classification, clustering, FARS.

1. INTRODUCTION

There are is lot of vehicles driving on the road every day, and traffic accidents could happen at any time. Some of them accident involves fatality, means people die in that accident. As human being, we all want to avoid accident and stay safe. To find out how to drive safer, data mining technique could be applied on the traffic accident dataset to find out some valuable information, thus give driving suggestion. Data mining uses many different techniques and algorithms to discover the relationship in large amount of data. It is considered one of the most important tool in information technology in the previous decades [2]. Association rule mining algorithm is a popular methodology to identify the significant relations between the data stored in large database and also plays a very important role in frequent item set mining [1]. A classical association rule mining method is the Apriori algorithm who main task is to find frequent item sets, which is the method we use to analyze the road traffic data. Classification in data mining aims at constructing a model (classifier) from a training data set that can be used to classify records of unknown class labels. The Na¨ıve

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. We Used FARS dataset for our System.

2. REVIEW LITRETURE Jayasudha [4] analyzed 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 scheme which describes the road accident via road condition, person involved and other data would be useful for case evaluation, collecting additional evidences, settlement decision and subrogation. 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 [4]. Solaiman et. al. [8] describes various ways accident data could be collected, placed in a centralized database server and visualized the accident. Data could be collected via different sources and the more the number of sources the better the result. This is because the data could be validate with respect to one another few could be discarded thus helping to clean up the data. Different parameters such as junction type, collision type, location, month, time of occurrence, vehicle type could be visualized in a certain time strap to see the how those parameters change and behave with respect to time. Based on those attributes one could also classify the type of accident. Using map API the system could be made more flexible such that it could find the safest and dangerous roads [8]. Partition based clustering and density based clustering were performed by Kumar [6] to group similar accidents together. It's based on a categorical nature of most of the data K-modes algorithm was used. To find the correlation among various sets of attributes association rule mining was performed. First the data set is classified into 6 clusters and each of them are studied to predict some patterns. Among the various rules that are generated those which seemed interesting were considered based on support count and confidence. The experiments showed that the accidents were dependent of location and most of the accident occurred in populated areas such as markets, hospitals, local

colonies. Type of vehicle was also a factor to determine the nature of accident; two wheelers met with an accident the most in intersections and involved two or more victims. Blind turn on road was the most crucial action responsible for those accidents and main duration of accidents were on morning time a.m. to 6 a.m. on hills and 8 p.m. to 4 a.m. on other roads [6]. Krishnaveni and Hemalatha [5] worked with some classification models to predict a severity of injury that occurred during traffic accidents. Naive Bayes Bayesian classifier, AdaBoostM1 Meta classifier, PART Rule classifier, J48 Decision Tree classifier, and Random Forest Tree classifier are compared for classifying the type of injury severity of various traffic accidents. The final result shows that the Random Forest out performs the other four algorithms [5]

EXISTING SYSTEM

- Williams et al. [5] have found through their studies that the age and experience of a driver also play a major role in the occurrence of accidents. Suganya, E. and S. Vijayarani [6] in their paper have analysed the road accidents in India and compared the performance of different classification algorithms such as linear regression, logistic regression, decision tree, SVM, Naïve Bayes, KNN, Random Forest and gradient boosting algorithm using accuracy, error rate and execution time as a measure of performance. They have found the performance of KNN to be better than that of the others.
- Sarkar et al. [7] have done a comparative study on the type of roads that are prominent in accidents. While exploring the other components associated with accidents, they have found that the occurrence of accidents in highways is more common than in a normal road similar to [4]. Stewart et al. [8] have utilized original data in building a neural network model to predict accidents. They found that this model was able to give quicker results than those being used in the models built on Indian roads.
- Zheng et al. [9] have studied the range of injuries that come forth in a motor vehicle accident and have also analyzed the emotions of the drivers involved in the accidents that could

have been a causal factor. Arun Prasath N and Muthusamy.

- Punithavalli [10] have conducted an extensive survey on the different techniques used in road accident detection over the years, the approaches implemented in them and discusses their merits and de-merits.
- George Yannis et al. [11], in their paper, have discussed about the current practices used in the development of accident prediction models on an international level. Detailed information on various models have been collected with the help of questionnaires and they have made use of this data to identify which could be the most useful model that can be applied for accident prediction.
- Anand, J. V [12] has developed a method to determine the effect of different variables in the detection and prediction of atmospheric deterioration all over the world. Fuzzy C means clustering, R-studio, and the ARIMA frame work have been made use of in creating this method. A similar approach can also be tried in evaluating the impact of various factors on road accidents. Analyzing the original cause of accidents is important because this will tell us the impact factor and contribution of each attribute towards road accidents. Tiwari et al. [13] have made use of self-organizing maps, K-mode clustering techniques, Support Vector Machines, Naïve Bayes and Decision tree to classify the data from road accidents based on the type of road users.

DISADVANTAGES

1) The system doesn't have facility to train and test on large number of numbers.

2) The system doesn't measure an accurate road accident due to poor classification models.

PROPOSED SYSTEM

In the proposed system, the system has built an application that is capable of predicting the possibility of occurrence of accidents based on available road accident data. Data pre-processing is done on this road accident data to obtain a dataset. The data preprocessing step includes cleaning to remove the null and garbage values, and normalization of the data, followed by feature selection, where only relevant

features from the original dataset are selected to be included in the final dataset. The dataset is then subjected to different data mining techniques. Clustering is performed on this dataset. The clusters are then subjected to other algorithms like Support Vector Machines (SVM) and Apriori. Since the data being used for the study has an unknown distribution and we need to sort out the frequent and infrequent items in the dataset, the former (SVM) is used to predict the probable risk of accidents while the latter (Apriori) is applied to perform rule mining, that is, to generate a frequent item set based on given support and confidence values. Rules have been set considering different combinations of factors which have caused accidents of varying nature and severity in different road types and weather conditions. For the frequently occurring item sets, the chosen support and confidence values imply the higher probability of the particular combination of attributes in leading to an accident. For example, based on the rule mining done, the probable risk of an accident occurring even during fine weather in a junction on account of over-speeding is high and could prove to be fatal based on the training dataset. SVM classification has been used to characterize each accident event into a high or a low risk category. Various data mining techniques and exploratory visualization techniques are applied on the accident dataset to get the interpreted results..

ADVANTAGES

1) These optimized models can be efficiently utilized by the government to reduce road accidents and to implement policies for road safety.

2) The overall model has helped to give an understanding of the combinations of factors that have proven fatal in accident scenarios.

RESULT ANALYSIS The percentage of fatal accidents is depend on four variables: SPEED_LMT (speed limit), LIGT_CONDITION (light condition), WEATHER_CONDITION (weather condition) and SURFACE_COND (road surface condition).

Collision Type: The percentage of fatal accidents happened on different collision types. In Front-toFront (Head-on Collision), the percentage of people and fatals involved are much higher than the percentage of accident number, which reveals that head-on collision has higher fatal rate in a fatal accident.

Speed Limit: The percentage of fatal accidents happened at different speed limit.

Light Condition: The percentage of fatal accidents happened on different light condition. most fatal accidents happen in day light condition because much more roadway traffic happens in day time other than at night.

Weather Condition: The percentage of fatal accident happened on different weather. Most fatal accidents happened at clear/cloud weather. This is understandable because clear/cloud is the most usual case of weather condition. ¬ Surface Condition: Its The percentage of fatal accident happened on different roadway surface condition. Most fatal accidents happened on dry surface. 1

CONCLUSION As seen in statistics, Linear Regression, and the classification, the environmental factors like road 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. From the clustering result we could see that some states/regions have higher fatal rate, while some others lower. We may pay more attention when driving within those risky regions. Through the task performed, we realized that data seems never to be enough to make a strong decision. If more data, like non-fatal accident data, weather data, mileage data, and so on, are available, more test could be performed thus more suggestion could be made from the data.

REFERENCES.

[1] Amira A El Tayeb, Vikas Pareek, and Abdelaziz Araar. Applying association rules mining algorithms for traffic accidents in dubai. International Journal of Soft Computing and Engineering, September 2015.

[2] William M Evanco. The potential impact of rural mayday systems on vehicular crash fatalities. Accident Analysis & Prevention, 31(5):455–462, September 1999.

[3] K Jayasudha and C Chandrasekar. An overview of data mining in road traffic and accident analysis. Journal of Computer Applications, 2(4):32– 37,2009.

[4] S. Krishnaveni and M. Hemalatha. A perspective analysis of traffic accident using data mining techniques. International Journal of Computer Applications, 23(7):40–48, June 2011

[5] Sachin Kumar and Durga Toshniwal. Analysing road accident data using association rule mining. In Proceedings of International Conference on Computing,

Communication and Security,pages1–6, 2015.

[6] Eric M Ossiander and Peter Cummings. Freeway speed limits and traffic fatalities in washington state. Accident Analysis & Prevention, 34(1):13– 18,

[7] KMA Solaiman, Md Mustafizur Rahman, and Nashid Shahriar. Avra Bangladesh collection, analysis & visualization of road accident data in Bangladesh. In Proceedings of International Conference on Informatics, Electronics & Vision, pages 1–6. IEEE,