Assisting Frequent Cause of Accidents on Roads using Data Mining Techniques

Srinivasa Rao D[1], Gowtham V[2], Tanuja T[3], Revanth CH[4], Venkateswara Rao B[5]
Department of Computer Science and Engineering
Lakireddy Balireddy College Of Engineering, India

***Abstract* -** Accidents on the road are one of the most important factors that cause the death between the public and financial loss of public and private economy. Safety on road is a term that deals with the planning and introducing certain strategy to escape the accidents causing on the road. The analysis of the road traffic can help to find out the most frequent ways of causing an accident and will help us to avoid the accident before it cause. The heterogeneity of accidents on road data is one of the biggest challenges in road safety analysis. From this case, we use Association rules were discovered by Apriori algorithm, classification model was built by Naive Bayes classifier, and clusters were formed by simple K-means clustering algorithm new road accident data from the states of USA. We choose our dataset from FARS fatal accident dataset to address the problem. This is all done to provide a precaution to the public for their safety.

***Keywords* -** Roadway fatal accidents, association, classification, clustering

I. INTRODUCTION

There humongous vehicles travelling on the road daily, and accidents due to traffic could happen at most of the times. Few accidents may involve fatality, means people might cause to death in that accident. As a responsible human being, we all need to avoid accidents by keeping ourselves safe. To find how to do safe driving, data mining technique could be applied on the traffic accident dataset to find out some useful information, thus give driving suggestion.

Data mining use various techniques and algorithms to find the relationship in large amount of dataset. It is observed that one of the most important tool in information technology in the last ten years. 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 itemset mining. A classical association rule mining method is the Apriori algorithm who main task is to find frequent item sets, with the help of this dataset we will start pre-processing the data given, after the data is pre-processed the algorithms on the dataset are applied we perform algorithms like apriori, k-means and clustering. After these were done we get the results of the dataset accordingly. Like we get the graphs of the required dataset. After these all were done we analyze the output generated and predict the cause of an accident at particular state.

II. RELATED WORKS

Jayasudha 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. That database scheme which describes the road accident via roadway condition, person involved and other data would be useful for case evaluation, collecting additional evidences, settlement data would be useful for case evaluation, collecting additional evidences, settlement

The effect on road speed on accident in the state of Washington was investigated by Eric [7]. Some researches claim that those states which increased speed limit from 55mph to 65mph after 1974 had the fatality rate go up by 27% compared to increase in 10% in the states that did not increase the speed limit. It is claimed that as the effect of change in maximum speed is varies between urban and rural areas. After 1987 accident in rural areas increased while urban areas stayed relatively constant but clash rate in urban intersection is twice as high as in rural intersection. Accident is dependent on area (urban/rural), type of street (intersection, highway) [7].

It is assumed that the fatality rate of an accident might be reduced with the introduction of an express emergency system. Reducing the time of delivery for emergency medical services (EMS) accident victims could be treated in time saving their lives. Accident notification time, the difference between crash and EMS notification time is the most crucial. Trauma is time dependent disease and trauma victims could be saved if treated on time. Trauma victims could stabilized if treated soon. The first 10 minutes is called golden hour. The time was 12.3 minutes for the one who died and 8.4 minutes for the one who survived. The fatality rate was also much higher in rural areas because of unavailability of rapid EMS response in those areas. There are also several other factors affecting the fatality such as vehicle kilometres travelled, alcohol consumption, driver age distribution, accident notification time, personal income per capita and so on [3].

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 validating 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. Based on the 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 4.00a.m.to6a.m.onhillsand8p.m.to4a.m.on other roads [6].

Krishnaveni and Hemalatha [5] worked with some classification models to predict the 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 outperforms the other four algorithms [5].

Amira, Pareek, and Araar [2] applied association rules mining algorithm on a dataset about traffic accidents which was gathered from Dubai Traffic Office, UAE. After information preprocessing, Apriori and Predictive Apriori association rules algorithms were applied to the dataset to investigate the connection between recorded accidents and factors to accident severity.

III. METHODOLOGY

The approach we took for our study follows the traditional data analysis steps, as shown in The following Fig .1.

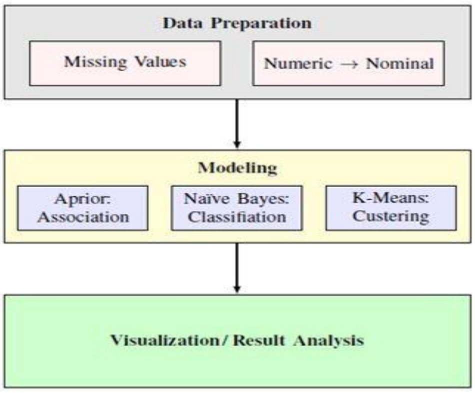


Figure 1

A. Data Preparation - Data preparation was performed before each model construction. All records with missing value (usually represented by 99 in the dataset) in the chosen attributes were removed. All numerical values were converted to nominal value according to the data dictionary in attached user guide. Fatal rate were calculated and binned to two categories: high and low.

Several variables are calculated from other independent variables.

Here are two examples:

- **Fatal Rate -** This variable denotes the percentage of fatality in a fatal accident computed as FATAL RATE = FATALS/PERSONS, where FATALS is the number of fatalities and PERSONS is the number of persons involved in the accident. It is also referred as “rate” in the analysis.
- **Arrival Time -** This variable is the arrival time of emergency staff in minutes, calculated as ARRIV AL TIME = 60 × (ARR HOUR − HOUR) +ARR MIN −MINUTE. All records with missing values on these time-related attributes are removed, and 24 to make it computationally easier add the early morning hours after 12:00 midnight.

B. Modeling - We first calculated several statistics from the dataset to show the basic characteristics of the fatal accidents. We then applied association rule mining, clustering, and Naive Bayes classification to find relationships among the attributes and the patterns.

C. Result Analysis - The results of our analysis include association rules among the variables, clustering of states in the USA on their populations and number of fatal accidents, and classification of the regions as being high or low risk of fatal accident. We used Java along with eclipse to perform these analyses.

IV. EXPERIMENTAL RESULTS

The experimental results are as follows they are given in the tabular format, table format and bar graphs. One the data sent to the code is preprocessed the algorithms on the data set are applied and once the results are ready to display there are several formats to display the dataset and the data is to be preprocessed once the system undergoes the algorithms.

The Results are as follows:

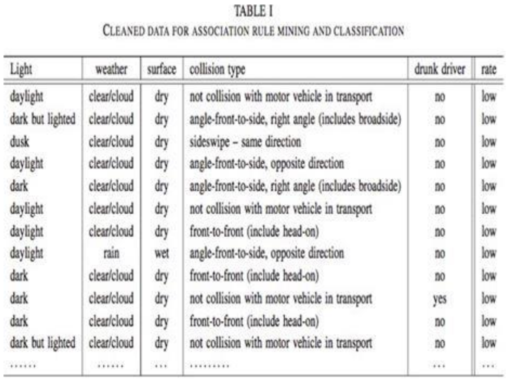


TABLE I

CLEANED DATA FOR ASSOCIATION RULE MINING AND CLASSIFICATION

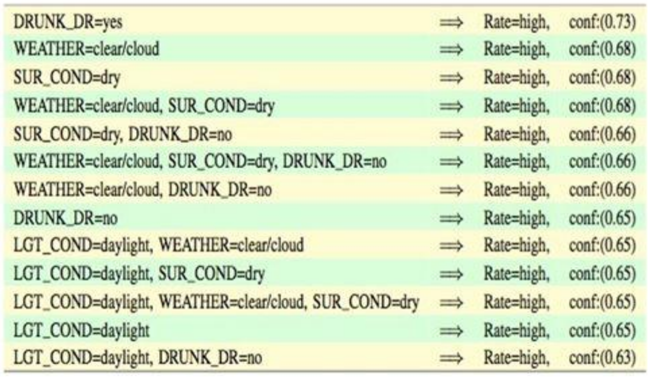
Light	weather	surface	collision type	drunk driver	rate
daylight	clear/cloud	dry	not collision with motor vehicle in transport	no	low
dark but lighted	clear/cloud	dry	angle-front-to-side, right angle (includes broadside)	no	low
dusk	clear/cloud	dry	sideswipe – same direction	no	low
daylight	clear/cloud	dry	angle-front-to-side, opposite direction	no	low
dark	clear/cloud	dry	angle-front-to-side, right angle (includes broadside)	no	low
daylight	clear/cloud	dry	not collision with motor vehicle in transport	no	low
daylight	clear/cloud	dry	front-to-front (include head-on)	no	low
daylight	rain	wet	angle-front-to-side, opposite direction	no	low
dark	clear/cloud	dry	front-to-front (include head-on)	no	low
dark	clear/cloud	dry	not collision with motor vehicle in transport	yes	low
dark	clear/cloud	dry	front-to-front (include head-on)	no	low
dark but lighted	clear/cloud	dry	not collision with motor vehicle in transport	no	low
……	……	…	………	…	…

This is the cleaned data table format that is obtained after the preprocessing of the data on the preprocessed data set we apply the algorithms like association rule mining and classification. These can be done only after the data is preprocessed. With out cleaning or preprocessing the data no algorithms are to be applied. If the data is not preprocessed there are many chances for the wrong outputs due to false data. The above table shows us the complete data that is preprocessed. There are six different attributes that are considered in the given table above.

The six different attributes that are considered in the above table are Light, Weather, surface, collision type, drunk driver and rate. Based on these six attributes we decide how many chances there are there for occurring and accident. At last finally we predict the chances of occurring an accident whether the chance are high or low.

TABLE II

THIRTEEN ASSOCIATION RULES WITH HIGHEST CONFERENCE DISCOVERED BY APRIORI ALGORITHM



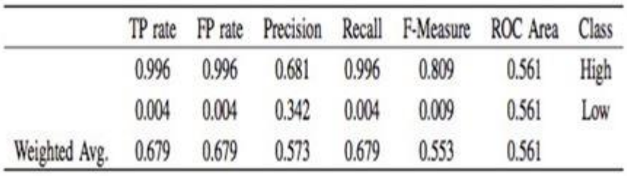
DRUNK_DR=yes	⟹	Rate=high,	conf:(0.73)
WEATHER=clear/cloud	⟹	Rate=high,	conf:(0.68)
SUR_COND=dry	⟹	Rate=high,	conf:(0.68)
WEATHER=clear/cloud, SUR_COND=dry	⟹	Rate=high,	conf:(0.68)
SUR_COND=dry, DRUNK_DR=no	⟹	Rate=high,	conf:(0.66)
WEATHER=clear/cloud, SUR_COND=dry, DRUNK_DR=no	⟹	Rate=high,	conf:(0.66)
WEATHER=clear/cloud, DRUNK_DR=no	⟹	Rate=high,	conf:(0.66)
DRUNK_DR=no	⟹	Rate=high,	conf:(0.65)
LGT_COND=daylight, WEATHER=clear/cloud	⟹	Rate=high,	conf:(0.65)
LGT_COND=daylight, SUR_COND=dry	⟹	Rate=high,	conf:(0.65)
LGT_COND=daylight, WEATHER=clear/cloud, SUR_COND=dry	⟹	Rate=high,	conf:(0.65)
LGT_COND=daylight	⟹	Rate=high,	conf:(0.65)
LGT_COND=daylight, DRUNK_DR=no	⟹	Rate=high,	conf:(0.63)

After applying the data preprocessing algorithms we make an algorithm called association rule mining with the help of this algorithm we find out the mostly occurring data sets in a given data and we can easily have a chances to predict an incident based on that data. If we observe the above table there are different cases in which there is a chance of occurring and accident the data above tells us that those are the few set of condition in which the accidents are mostly occurred. So from this we can estimate a condition and act accordingly.

After the applying of association rule mining we apply an algorithm called apriori algorithm.

TABLE III

RESULTS OF THE NAÏVE BAYES CLASSIFICATION



	TP rate	FP rate	Precision	Recall	F-Measure	ROC Area	Class
	0.996	0.996	0.681	0.996	0.809	0.561	High
	0.004	0.004	0.342	0.004	0.009	0.561	Low
Weighted Avg.	0.679	0.679	0.573	0.679	0.553	0.561	

The above table (iii) displays us the result of Precision, Recall, F-measure, ROC Area and Class. From all the above values the prediction of an accident can be easily predicted so that there are huge chance to avoid accidents on the road. This can be applied to any data set of a particular location and all that matters is the attributes that we look forward to consider the naïve bayes classification. In which the prediction values of the above given dataset. The predicted values are as follows. The algorithm gives the values of TP rate, FP rate,

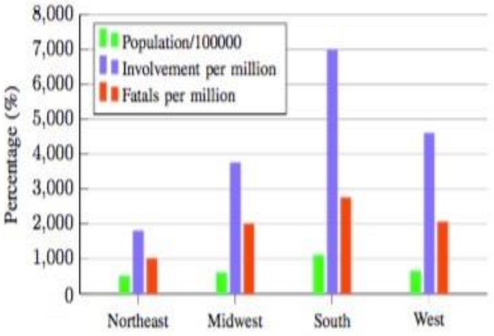


Fig. 7. Fatal accident in different regions

The above bar graph show the percentage of accidents that took place indifferent regions. We consider the factors like Population, Involvement per million and fatal per million this could to really useful to access an accident before it occurs. And can be easily predicted at particular locations where there is a chance to take place of an accident.

V. CONCLUSION

From the above experiments all conducted we can easily predict an accident before it occurs in particular location because we take all the data sets based on the locations and considering few factors like collision type, speed, whether the driver is drunk or not. Based on these proportions we accurately judge the chances of an accident that might take place.

And additionally we represent the data obtained in a pictorial and graphical form, which helps the predictors to make an easy analysis based on them, and help to give an easy prediction.

VI. REFERENCES

[1]. Amiira A El Tayeb, Vikass 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. Krishnavenii and M. Hemalathaa. A perspective analysis of traffic accident using data mining techniques. International Journal of Computer Applications, 23(7): 40–48, June 2011.

[5]. Sachiin Kumar and Durga Toushniwal. Analysing road accident data using association rule mining. In Proceedings of International Conference on Computing, Communication and Security, pages 1–6, 2015.