Analysis of Road Traffic Fatal Accidents Using Data Mining Techniques

Liling Li, Sharad Shrestha, Gongzhu Hu

Department of Computer Science
Central Michigan University, USA
(li8l, shres1s, hu1g)@cmich.edu

***Abstract*—Roadway traffic safety is a major concern for transportation governing agencies as well as ordinary citizens. In order to give safe driving suggestions, careful analysis of roadway traffic data is critical to find out variables that are closely related to fatal accidents. In this paper we apply statistics analysis and data mining algorithms on the 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. Association rules were discovered by Apriori algorithm, classification model was built by Naïve Bayes classifier, and clusters were formed by simple K-means clustering algorithm. Certain safety driving suggestions were made based on statistics, association rules, classification model, and clusters obtained.**

***Index Terms*—Roadway fatal accidents, association, classification, clustering**

I. Introduction

There are a lot of vehicles driving on the roadway every day, and traffic accidents could happen at any time any where. Some 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 itemset mining [1]. A classical association rule mining method is the Apriori algorithm who main task is to find frequent itemsets, which is the method we use to analyze the roadway 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.

We used the FARS dataset for our study. The Fatal Accidents Dataset contains all fatal accidents on public roads in 2007 reported to the National Highway Transportation Safety Administration (NHTSA) [9].

The dataset is downloaded from California Polytechnic State University and all data originally came from FARS (Fatal Accident Reporting System). The dataset contains 37,248 records and 55 attributes.

The data description can be found in the document *FARS Analytic Reference Guide, 1975 to 2007* [11].

II. Related Works

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 roadway 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].

978-1-5090-5756-6/17/$31.00 ©2017 IEEE
SERA 2017, June 7-9, 2017, London, UK

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 kilometers traveled, 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 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. 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 4.00 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 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 in Dubai. Two sets of class association rules were generated using the two algorithms and summarized to get the most interesting rules using technical measures. Exact results demonstrated that the class association rules created by Apriori algorithm were more viable than those created by Predictive Apriori algorithm. More relationship between accident factors and accident severity level were investigated while applying Apriori algorithm [2].

III. Methodology

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

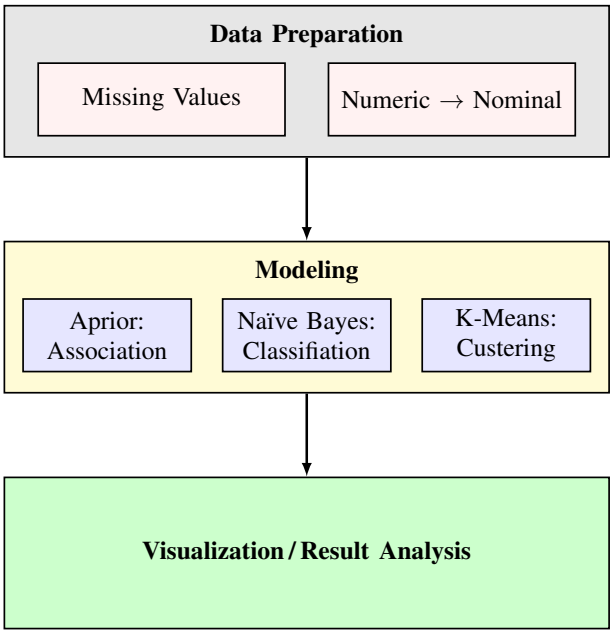


Fig. 1. Work flow

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 $ARRIVAL\_TIME = 60 \times (ARR\_HOUR - HOUR) + ARR\_MIN - MINUTE$. All records with missing values on these time-related attributes are removed, and the early morning hours after 12:00 midnight are added by 24 to make it computationally easier.

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 Naïve Bayse 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 a data analytic tool Weka to perform these analysis.

IV. Experimental Results

A. Statistics results

The number of fatal accident in each months are shown in Fig 2. The most fatal accidents happened in July and the least in February.

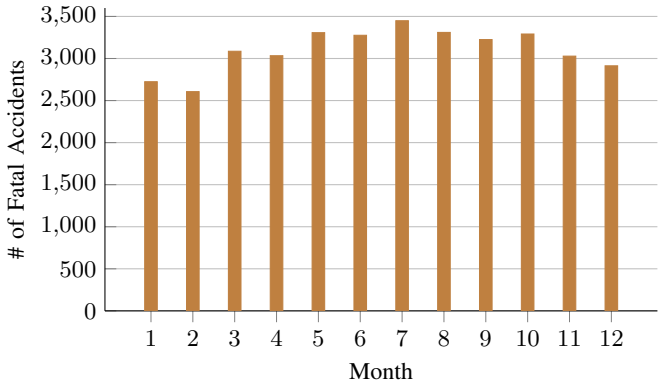
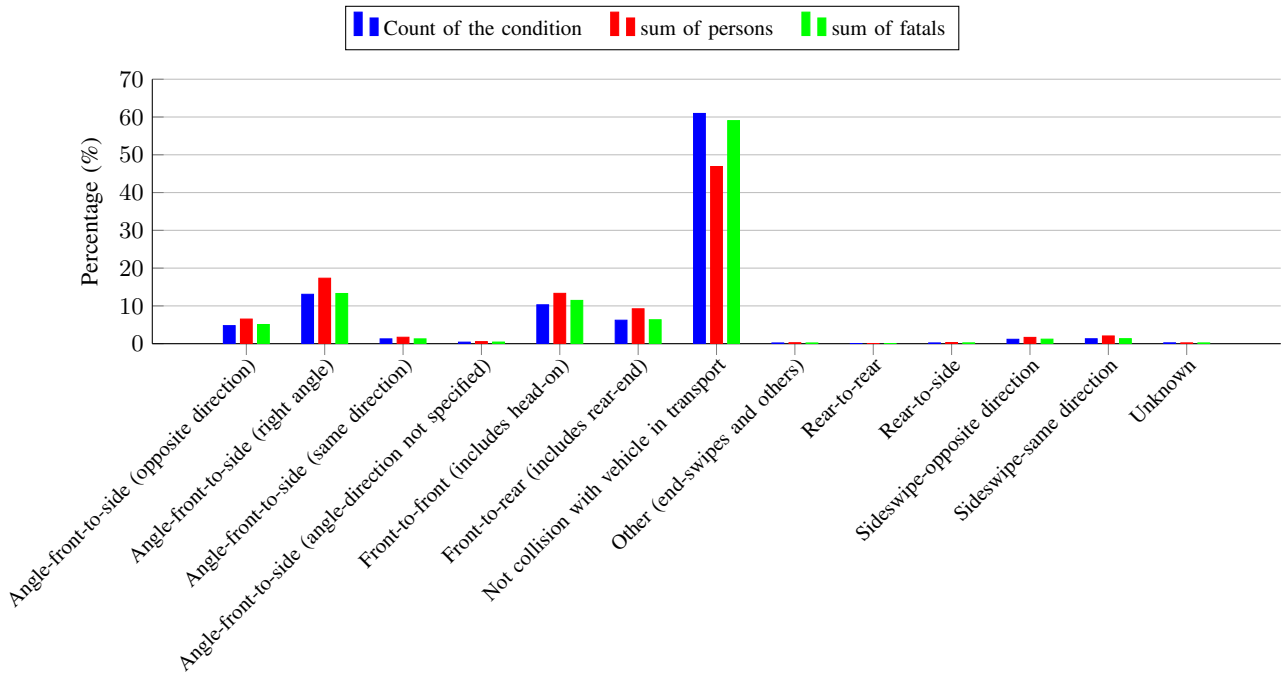


Fig. 2. Number of fatal accident per month

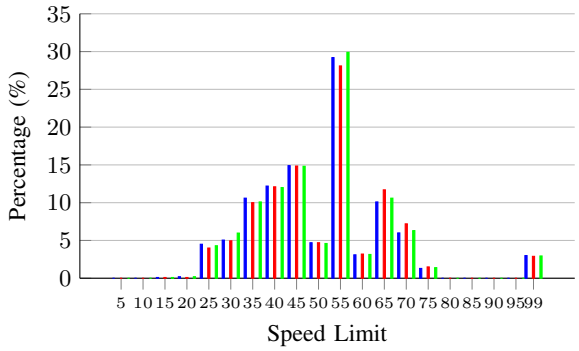
Fig. 3 shows the percentage of fatal accidents in four other variables: SP_LIMIT (speed limit), LGT_COND (light condition), WEATHER (weather condition), and SUR_COND (road surface condition).

1) Collision Type: The percentage of fatal accidents happened on different collision types in comparison of people and fatals involved are shown in Fig 3(a). Surprisingly, the most fatal accidents are not in collision with motor vehicle in transportation. In Front-to-Front (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.

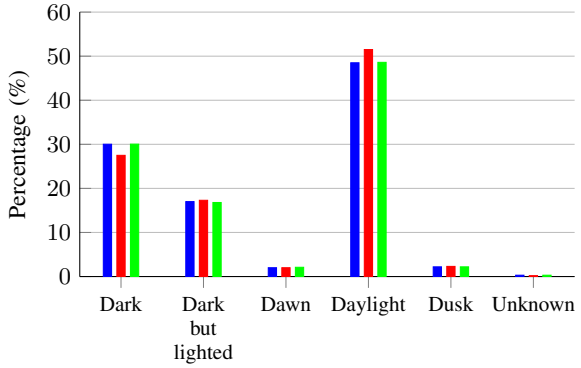
2) Speed Limit: The percentage of fatal accidents happened at different speed limit in comparison of people involved and fatals involved are shown in Fig 3(b). Most of fatal accidents happened at speed limit 55 mph. The value "99" infers the missing value on attribute SP_LIMIT.



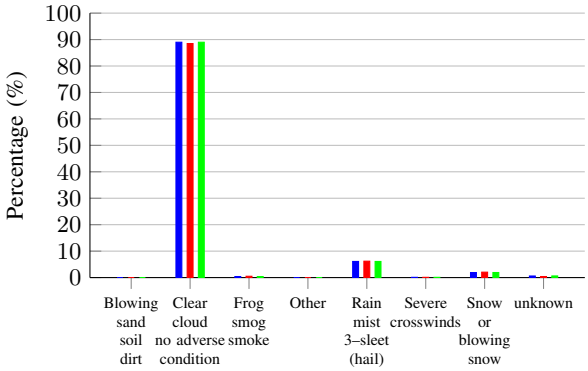
(a) Fatal accidents on different collision types



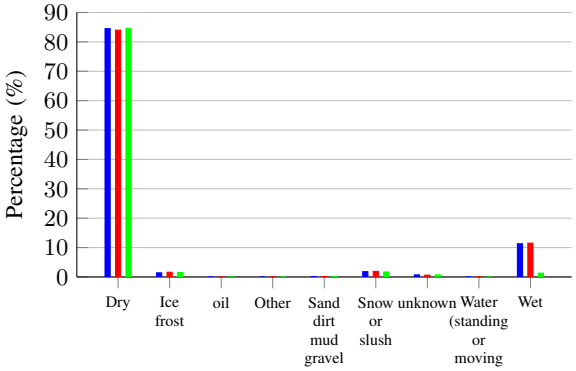
(b) Fatal accidents on different speed limits



(c) Fatal accidents on different light conditions

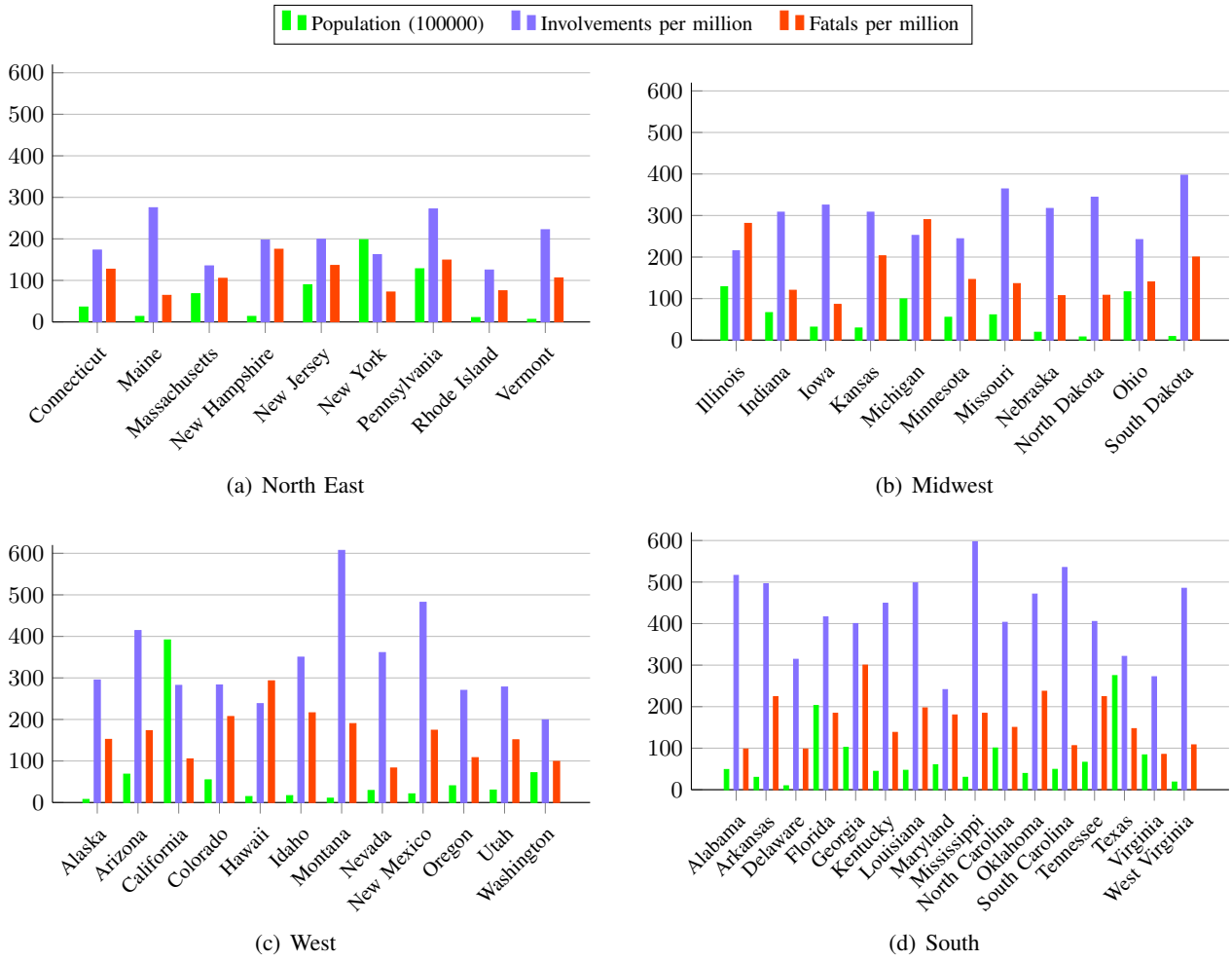


(d) Fatal accidents on different weather conditions



(e) Fatal accidents on different surface conditions

Fig. 3. Percentage of fatal accidents on four attributes



(a) North East (b) Midwest

(c) West (d) South

Fig. 4. fatal accidents in states

3) Light Condition: The percentage of fatal accidents happened on different light condition in comparison of people and fatals involved are shown in Fig 3(c). Unsurprisingly, most fatal accidents happen in day light condition because much more roadway traffic happens in day time other than at night.

4) Weather Condition: The percentage of fatal accident happened on different weather is shown in Fig 3(d), in comparison with percentage of people and fatals involved. Most fatal accidents happened at clear/cloud weather. This is understandable because clear/cloud is the most usual case of weather condition.

5) Roadway Surface Condition: The percentage of fatal accident happened on different roadway surface condition is shown in Fig 3(e) . Most fatal accidents happened on dry surface. This is understandable because the most usual case of road condition is that the road surface is dry.

Evanco [3] pointed out that there are "Golden Ten Minutes" for trauma in traffic accident. To find out the possible relation between emergence service (EMS) arrival time (in minute) and fatal rate, the correlation is performed by R, as shown in Fig 5.

Most fatals happened in short time and there is no significant relation between the EMS arrival time and fatal rate (correlation=0.1132231). The minimum arrival time is 0 minute, which either because the EMS was at scene or data entry error. The average time take for EMS arrival is 18.27 minutes, the median is 10 minutes, and the maximum time is over 18 hours.

We also need to mention that all the data is about fatal accident, so no matter how long would it take for EMS to arrive, there would always be fatals. Also, there is no variable recording at what time the death happened, and a lot of records are missing value at time, so very limit information could be inferred from the time relevant attributes.

Similar statistics also performed on other attributes

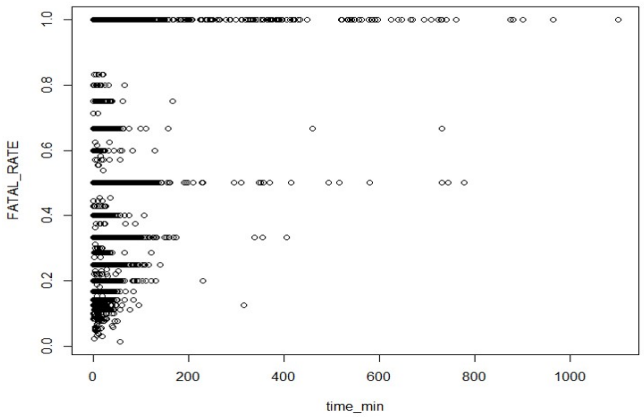


Fig. 5. Correlation between EMS arrival time and fatal rate

and the results are not significant so are not included here.

After performing basic statistics and related work research, five attributes (collision type, weather, surface condition, light condition, drunk driver) are considered and selected as affecting fatal rate.

B. Association Rule Mining

Before applying the algorithms, the tuples with missing value in chosen attributes were removed, the numerical values were converted to nominal values according to data dictionary in the user guide [11]. The clean data was stored in CSV format and ready to be analyzed by the data analyzing tool Weka.

The clean data for association rule mining and classification contains 36,789 tuples, 5 condition attributes, and 1 decision attribute. A small partial sample of the dataset is shown in Table I. All values were converted to nominal values.

After applying Apriori algorithm with minimum *support* = 0.4 and minimum *confidence* = 0.6 in Weka, association rules with fatal rate at the right side as decision were generated. The best 13 rules are shown in Table II.

We could see that fatal accidents involving drunk driver have higher fatal rate, which means drunk drivers are much more dangerous than others. Also the clear/cloud weather condition with day light has high fatal rate, this reveals that not only the accident percentage is higher, as shown in basic statistics, but also the fatal rate are high (with confidence level = 0.65).

C. Classification

Naïve Bayes classifier was built on the cleaned data. Of the total 36,789 records, 24,994 were correctly classified giving a 67.95% accuracy rate. The various evaluation measures are given in Table III.

The Naive Bayes Classifier shows that the fatal rate does not strongly depend on the given attributes, although they are considered feature in comparison to other attributes in the dataset.

D. Clustering of States

To find out which states are similar to each other considering fatal rate, and which states are safer or more risky to drive, clustering algorithm was performed on the fatal accident dataset.

To perform the clustering, total number of fatality per state was calculated. Also the population data for each states in 2007 was obtained from U.S. Census Bureau [10]. With the fatal accident and the population dataset, fatalities per million people in the state was calculated. This allowed us to compare relative fatal rate in a state regardless of population of state.

The simple K-means algorithm with Euclidean distance as the dissimilarity measure was applied to the data of 48 states (without Wisconsin, Wyoming, and District of Columbia, for some reason) with two variables: population (in 100,000) and number of fatal accidents. The states were grouped into 3 clusters as shown in Fig 6.

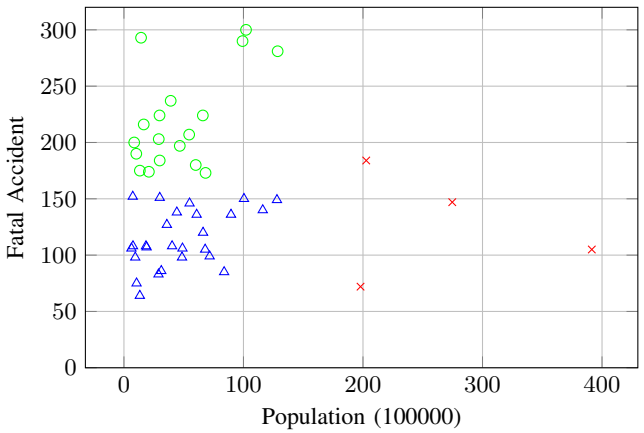


Fig. 6. Clusters of states by fatality of per million people in states

The three clusters are:

- Cluster A (blue): 26 states. Those states in cluster A represents the safe state with relatively lower fatal rate per million people.

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
.....	.....	...	.....	...	...

TABLE II
THIRTEEN ASSOCIATION RULES WITH HIGHEST CONFERENCE DISCOVERED BY APRIORI ALGORITHM

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

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	

- Cluster B (green): 18 states were clustered to cluster B which had relatively higher fatal rate.
- Cluster C (red): 4 states, California, Texas, New York and Florida, formed cluster C. These states have relatively large population and lower fatal rate. They are considered safe driving region and also outliers.

After careful observation it was found that none of the states from mid-west or north-east region lied on cluster A, and almost all the states from south were in cluster B. Only two states from the south were located in region of cluster A, which is considered to be safe. Virginia and Washington DC, those too were in the southern region bordering north east. Georgia had highest fatal rate per million people, whereas Mississippi had highest number of per million people involved in fatal accidents. But Montana from west also had as much people involved in fatal accident as Mississippi.

All four regions, North East, Midwest, West, and South, were also compared to each another to find out what's the difference between them. The comparison is shown in Fig 7.

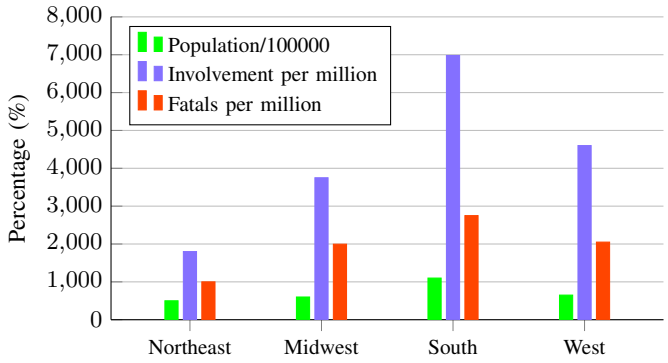


Fig. 7. Fatal accident in different regions

The southern region seemed to have 350% more people involved in an accident and almost 300% higher fatal rate compared to north east. This means that south is much more risky compared to rest regions. North east is the safest region and followings are mid-west and west.

V. Conclusion

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.

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 states/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] Divya Bansal and Lekha Bhambhu. Execution of Apriori algorithm of data mining directed towards tumultuous crimes concerningwomen. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(9), September 2013.

[2] 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.

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

[4] 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.

[5] 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.

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

[7] Eric M Ossiander and Peter Cummings. Freeway speed limits and traffic fatalities in washington state. *Accident Analysis & Prevention*, 34(1):13–18, 2002.

[8] 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, 2013.

[9] Trac Integrated SCM & Project Management. Fatal Accidents Dataset. https://wiki.csc.calpoly.edu/datasets/wiki/HighwayAccidents.

[10] U.S. Census Bureau. Population Estimates. http://www.census.gov/popest/data/historical/2000s/vintage_2007/, 2007.

[11] U.S. Department of Transportation. FARS analytic reference guide, 1975 to 2009. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811352, 2010.