Association Rule Mining on Five Years of Motor Vehicle Crashes

Jean Raymond Daher[1], Satish Chilkaka[1], Abdunnaser Younes[2], and Khaled Shaban[2]

[1]*Department of Computer science and Software Engineering· Concordia University, Montreal, Canada.*
[2]*Qatar University, Doha, Qatar*

Abstract. Every year, road accidents kill more than a million people and injure more than 20 million worldwide. This paper aims to offer guidance on road safety and create awareness by pinpointing the major causes of traffic accidents. The study investigates motor vehicle crashes in the Genesee Finger Lakes Region of New York State. Frequency Pattern Growth algorithm is utilized to cultivate knowledge and create association rules to highlight the time and environment settings that cause the most catastrophic crashes. This knowledge can be used to warn drivers about the dangers of accidents, and how the consequences are worse given a specific context. For instance, a discovered rule from the data states that 'most of the crashes occur between 12:00 pm and 6:00pm'; hence, it is suggested to modify existing navigation application to warn drivers about the increase in risk factor.

1 Introduction

Road safety is a major concern all over the world. Approximately 1.3 million people die every year, and 20 to 50 million people are non-fatally injured [1]. In particular, the United States is annually marked with over 37,000 road accident fatalities and 2.5 million injuries

This paper aims to provide suggestions to improve road safety. It outlines the major causes of road accidents by applying data mining algorithms to data collected from past accident records of the Genesee Finger Lakes Region of New York State, where thousands of motor vehicle crashes occur every year. The data was posted in March 2015 by Socrata, a reliable source for analyzing large scales of recorded data [2]. The dataset encompasses 36097 records with 33 attributes. After data preprocessing, the cleaned data is analyzed and mined using the Fp-Growth data mining technique.

The rest of the paper is organized as follows: Section 2 discusses related works. Data pre-processing techniques are discussed in Section 3. Section 4 presents and discusses the results obtained. Finally, Section 5 concludes the paper and outlines potential future work.

2 Literature review

Vehicle accidents have been studied extensively. Xu *et al.* [3] Used Classification and Regression Tree (CART) and found that crash occurrences depend on the roadway characteristics, traffic and environmental factors. However, the dataset is now obsolete and covers three months only.

Fogue *et al.* [4] proposed the Automatic Accident Notification and Assistance System to estimate the severity of traffic accidents and concluded that vehicle speed is the primary factor in head-on collisions. Their findings could be used to improve emergency services; however, the authors did not offer explanations for the causes of crashes.

Krishnaveni *et al.* [5] compared Naïve Bayes, AdaBoostM1, PART, J48 and Random Forest Classifier based on injury severity. They found that weather, vehicle kinematics, vehicle class and driver's age are responsible for crashes.

Nayak *et al.* [6] used text mining to identify the causes of about 20000 motor accidents in Queensland, Austria, collected in 2004-2005. The authors asserted that most crashes are rear-end collisions at intersections and due to speeding. No other important crash factors were considered.

Beshah and Hill [7] studied 18,288 traffic accidents and achieved comparable accuracy using Decision Tree, Naive Bayes and K-Nearest Neighbors.

Zhang and Fan [8] studied data spanning 20 years in Saskatchewan, Canada, using decision trees, and showed that accidents are likely to happen to infirm drivers violating traffic rules, intoxicated drivers, and inexperienced drivers during poor weather conditions. However, vehicle factors were not considered in the study.

Jianfenget *al.* [9] analyzed over twenty thousand accidents in China, seeking a relation between the driver's, vehicle, environment and road conditions and the occurrence of crashes.

Using factor analysis, Haixia and Zhihong [10] identified the most important crash factors in the categories of road and driver conditions. However, their dataset was very small (372 crashes).

Emerson *et al.* [11] used decision trees and regression trees to establish a benchmark for surface skid resistance versus crash probabilities, for all the crashes.

Ramani and Shanthi [12] suggested designing education campaigns for parents and generating awareness to limit the number of fatalities.

Using Classification and Regression Trees (CART), Tree Net and Random Forest Beshah et al. [13] concluded that human behavior is a major factor in road accidents but did not provide any suggestions to improve road safety.

Using multinomial logistic regression, Pakgohar et al. [14] concluded that 98% of crashes happened due to human factor, 70% due to the environmental factor and 32% due to the vehicle factor. However, the environmental factors were not clearly identified.

T. Dipo Akomolafe, A. Olutayo [15] used several classifiers and concluded that tire burst, broken shaft and speeding are the most culpable collision factors. However, other factors that contribute to crashes were not considered.

In retrospect to the findings of the examined literature, we suggest the use of a larger dataset that considers more factors and spans longer periods of time to yield more comprehensive, more reliable and more conclusive collision factors.

3 Data pre processing

The accident data records have two different categories of attributes: qualitative and quantitative [16]. However, most of the attributes are qualitative, which include nominal, binary, and ordinal attributes. The quantitative attributes include interval-scaled attributes, and do not have continuous attributes.

Only 21 attributes of the dataset are considered in this study; the remaining 10 attributes are removed because they have little significance. The dataset contains 450 undefined values, which are filled using random attribute filling [16]. Missing attribute values are chosen randomly between their minimum and maximum limits. Driver age outliers that are less than 15 or more than 95 are ignored. The total number of removed records constitutes only 3.2% of our dataset; hence, the data remains cohesive.

Finally, the five years of accident data records is transformed into numerical records for plotting purposes.

Table 1. Investigated attributes

Attribute	Description	Domain (possible values)
Collision Type	The other party involved in the collision, e.g. another vehicle, a train, a structure.	34
Weather Condition	The weather condition during the collision, e.g., clear weather, cloudy, rainy.	9
Road Surface	The condition of the road on which collision takes place, e.g., dry, muddy, icy.	8
Lights	The road illumination during the collision. e.g., dark, daylight, dawn,	6
Damage	The seriousness of the collision, e.g., fatal, injury, property damage,	6
Traffic Control	Traffic control signs in the accident location, e.g., stop sign, flashing signal, school zone.	19
Collision Location	How the collision occurs, e.g., head on, rear end, sides swipe.	11
Vehicle 1 type	These two attributes describe the vehicles involved in the collision, e.g., a bus, a car, a bicycle.	8
Vehicle 2 type		
Factor Driver 1	The two attributes describe factors or reasons that contributed to the collision, e.g., a defective brakes, alcohol, distraction.	55
Factor Driver 2		
Intersection	Whether one or two roads intersecting.	2
Day of the week	The day of the week on which the collision happens	7
Time	The time rounded to hours at which the collision happens	24
Month	The month in which the collision happens	12
Year	The year in which the crash happens	5

4 Data mining process

The data mining process is a combination of classification of the attributes and their corresponding values and an association mining rule using Fp-growth algorithm [16]. Approximately 2000 rules are generated from the dataset. These rules are then inspected for significance. The first step is eliminating rules that have a support less than 40%. Then, rules that have a confidence less than 70% are eliminated. The confidence is set to 70% to ensure reliability of rules. However, the support is only 40% because some important rules are only present in a small part of the dataset. Rules with a lift measure equal to one are eliminated, as it indicates a lack of correlation between the attributes. From there, almost 700 rules remain and the manual interpretation starts for each rule. The rules contribute to refining the knowledge discovered, thus improving the meaningfulness of those results.

4.1 Preliminary analysis

Rapid Miner is used to evaluate the distribution of the dataset values. Other tools, such as Microsoft Excel are used to count the specific values of interest.

The total number of crashes studied is 34945. The most recorded crashes are with other motor vehicles: 21918 crashes. Most of the crashes are serious and record property damage and injury: 24174 crashes.

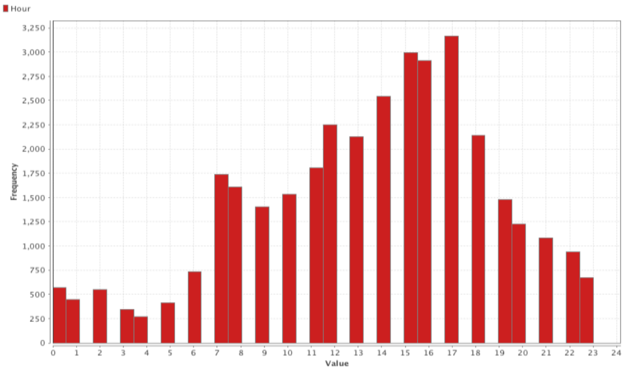


Figure 1. Effect of the hour of the day on crashes distribution

The effect of the hour of the day on the number of crashes is shown in Fig. 1. It can be seen that crashes are considerably more during daytime. More specifically, the number of crashes is the highest in the afternoon between 12:00 and 6:00 pm; in fact, nearly half the accidents (18139) occur during this period, with the peak at 5:00pm (3168 crashes)

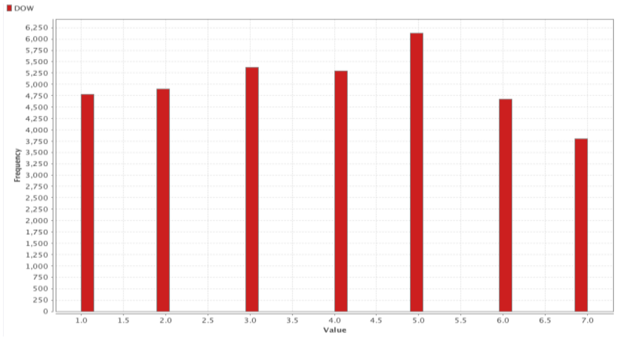


Figure 2. Effect of the day of the week on crashes distribution.

The effect of the day of the week on the number of crashes is shown in Fig. 2. Daily crashes tend to be constant. However, Saturday and Sunday have the lowest number of crashes, which can be explained by the reduction in traffic during these holidays. On the other hand, Fridays have the peak number of crashes, since people tend to stay out late at this day; thus, more traffic volume and the greater possibly of intoxicated drivers result in more accidents during this day.

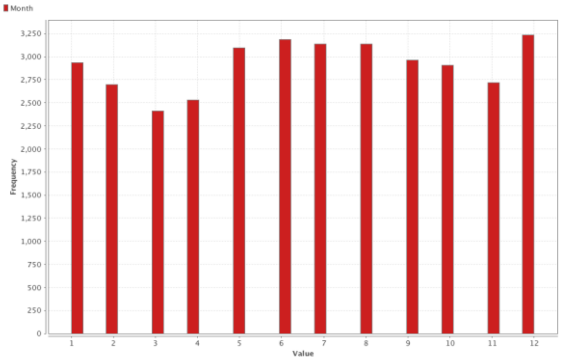


Figure 3. Effect of the month of the year on crashes distribution

In addition, Fig. 3 shows that December records 3337 crashes, the highest number compared to other months. December's records can be attributed to poor weather conditions and also widely celebrated holidays such as Christmas and New Year's Eve during which people consume more alcohol. This factor will be studied later in conjunction with seriousness of crash and weather conditions.

More accidents happen when the road surface is wet (3623, 10%) than when it is snowy (6787, 19.4%); it can be inferred that people are more cautious when it is snowing but ignore the dangers of slippery wet roads. However, most accidents are recorded in dry conditions: 24167 crashes.

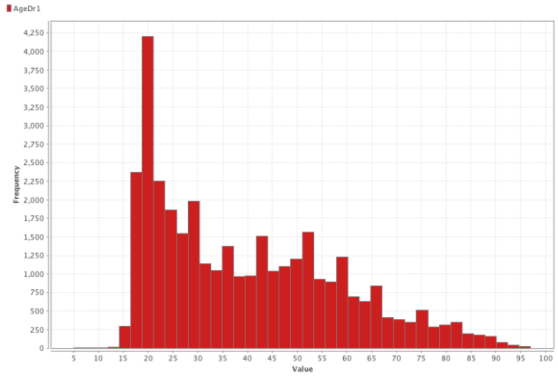


Figure 4. Age of Driver 1 involved in the crash

The effect of the driver's age on the number of crashes is shown in Fig. 4, where it is clear that drivers between the ages of 19 and 23 are the ones accountable for most accidents totaling 6455 accidents, that is, 18% of all crashes.

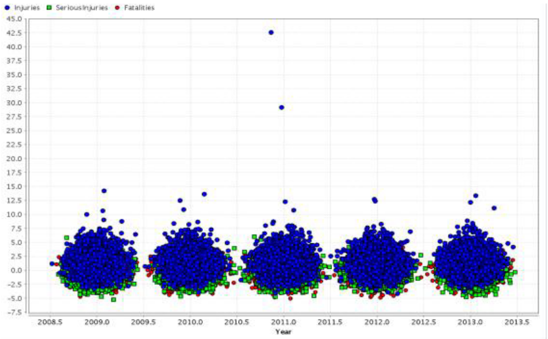
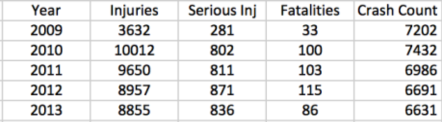


Figure 5. Visualization of occurrence according to severity of crash (injuries/serious injuries/ fatalities)

Fig. 5 shows the crashes over a five-year span along with their severity. Jittering is introduced in the plot, by adding random noise to the data to alleviate the problem of over-plotting similar or neighboring points, and thus better visualize the close values of data. The figure shows that, the crash severity is slightly declining with time. The figure is not conclusive but helps visualize the decreasing trend in the number of crashes and their severity over the years.

Table 2. Discrete count of occurrence according to severity of crash (injuries/serious injuries/ fatalities)



Year	Injuries	Serious Inj	Fatalities	Crash Count
2009	3632	281	33	7202
2010	10012	802	100	7432
2011	9650	811	103	6986
2012	8957	871	115	6691
2013	8855	836	86	6631

Table 2 shows that the number of injuries and fatalities decreased between the years 2010 and 2013. However, the year 2009 has the least severity of crashes.

4.2 Discovered association rules

In this section, more knowledge from the data is extracted using an association data mining technique [16]. The data mining process consists of four main steps illustrated in Fig. 6: Crash data retrieval, nominal to binomial conversion, Fp-Growth Algorithm and association rules creation.

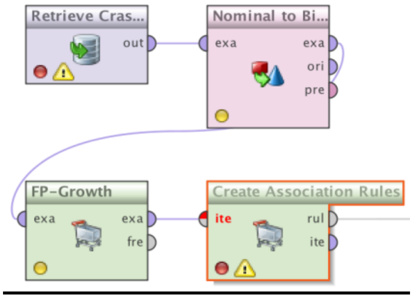


Figure 6. Visualization of occurrence according to severity of crash (injuries/serious injuries/ fatalities)

To generate association rules, the Fp-Growth Algorithm [16] is run on the data fifteen times; every time the properties of the Fp-Growth algorithm and association rule generation are modified to tune the parameters to generate slightly different rules. Each iteration is a search refinement towards more useful data. The challenge appears in interpreting the rules, since many rules are logical but do not contribute towards any knowledge discovery.

The algorithm generates 2247 rules; however, only a few of these rules are useful and non-trivial. Some of the meaningful rules discovered by this study are given below along with their support, confidence, and lift measures:

Rule 1: Crashes in daylight have no fatalities or serious injuries. (Sup:0.6335 Conf:0.8980 Lift: 1.1154)

Rule 2: Crashes in daylight involving two vehicles have no fatalities. (Sup:0.4093 Conf:0.8390 Lift:1.5644)

Rule 3: Crashes in daylight in which the second vehicle involved is a car, van or pickup do not cause serious injuries. (Sup:0.4244 Conf:0.9196 Lift:1.5930)

Rules 1, 2 and 3 indicate that accidents in daylight have little chance of being catastrophic. The lift shows that there is a positive correlation between the lights, and the fatalities and serious injuries. The second and third rules clearly assert that a crash between motor vehicles in daylight will not lead to fatalities or serious injury.

Rule 4: If the road surface is dry then there are no fatalities or serious injuries. (Sup:0.60523 Conf:0.8751 Lift:1.1021)

Rule 5: If the road surface is dry and two motor vehicles are involved in a crash then there are no fatalities. (Sup:0.4020 Conf:0.8389 Lift:1.5641)

Rules 4 and 5 show that the severity of crash decreases when the road surface is dry. The lift measure shows a positive correlation between road surface, fatalities and serious injuries. In Rule 5 it is demonstrated that when a crash occurs between two vehicles and the road surface is dry, the number of deaths is 0. The 1.56 value of lift indicates the strength of this correlation

Rule 6: Crashes in clear weather have no fatalities. (Sup:0.5206 Conf:0.9862 Lift:0.8990)

This rule states that when the weather is clear, no deaths occur with a confidence of 99%. Thus, it is concluded that in daylight, dry road surface and clear weather conditions, crashes are less severe and tend to have no fatalities. A negative correlation between weather and fatalities is present; thus lift <1.

Rule 7: If there are no fatalities and the weather is clear and the second involved vehicle is a car, van or pickup then the road surface is dry. (Sup:0.4102 Conf:0.9362 Lift:1.3538)

This rule shows a positive correlation between the two members. This helps us understand the impact of the conditions on the fatalities. If no one dies in a clear weather crash then the road must be dry.

Rule 8: If no one dies, then no one is seriously injured. (Sup:0.8843 Conf:0.8959 Lift:1.001)

Rule 9: If no one is seriously injured, then no one dies. (Sup:0.8843 Conf:0.9882 Lift:1.001)

These rules tell us that if a crash occurs and no one dies, then the chance of serious injuries is also minimal. Also, if a crash occurs and no one is seriously injured then the chance of fatalities is minimal. These rules help us classify a crash as serious or non-serious, where serious is leading to death. However, the lift measure indicates that fatalities and serious injuries are independent, and there is no correlation between them.

Rule 10: If there are no fatalities and the second vehicle involved is a car, van or pickup then there are no serious injuries. (Supp:0.5411 Conf:0.9150 Lift:1.5850)

Rule 11: If there are no serious injuries and the second vehicle involved is a car, van or pickup then there are no fatalities. (Supp:0.5411 Conf:0.9374 Lift:1.5850)

It seems that Rule 8 and 9 are misleading because the lift was equal to one. After Rule 10 and 11 are analyzed it is understood that when a crash occurs between two motor vehicles, being cars vans or pickups and the fatalities are zero then serious injuries are also equal to zero. The principle also works vice versa. The lift is very elevated indicating the positive correlation between each of these rules.

5 Conclusion

This study uses a new and comprehensive dataset encompassing several years. A methodical examination of all factors went into the study, followed by the omission of all factors found to be outside the scope of the study's objective; factors determined to be consequential to the study include those pertaining to weather, road, and vehicle conditions in precise time frames.

From our established findings on the categorical factors of collisions, we have developed a guideline to encourage drivers to take pre-emptive safety measures on the road:

1. Be more vigilant in the afternoon, between 12 and 6 pm, when most crashes occur.
2. Young drivers should drive responsibly since most people crash their cars in their early twenties.
3. Be extra cautious on Fridays because most accidents occur on Fridays.
4. Do not go over the speed limit.
5. Do not tailgate.
6. When making a right turn, make sure to check your blind spot.
7. Be on the lookout for traffic control signs; when there are none, be extra cautious.
8. When the road conditions are not ideal (clear, dry and daylight), accidents can be very dangerous.
9. If possible, drive during daylight, clear, dry days to reduce possibilities of collisions.

We could embed the results in some navigation software, such as Google maps, so that they would be highlighted in colored indicators reflecting the probability of crash (red for high probability, green for low probability and yellow for intermediate). The probability would be based on the checklist. For example, when the road conditions are dangerous and the timeframe is between 12 and 6 pm, the area would be highlighted in red. This would create awareness for the drivers and possibly reduce the number of crashes and deaths. The speed of the driver could also be tracked and the indicator would be updated accordingly.

Future work would generate new and improved safety practices to be appended to the current checklist based on the data analyzed and possibly finding another, more up to date, dataset to study. Keep in mind that our dataset was last updated in 2013. A new dataset would work on improving the quality of the checklist and possibly adding more points of interest.

References

1. Association for safe international road travel [http://asirt.org/].
2. https://opendata.socrata.com/Education/Five-Years-of-Motor-Vehicle-Crashes/tbyb-ykru
3. X. Xu, Ž. Šarić, A. Kouhpanejade, Traffic&Transportation, **Vol. 26**, (2014), No. 3, 191-199 191.
4. M. Fogue , P.Garrido, F J. Martinez, Juan-Carlos Cano, Carlos T(2012), Advances in Intelligent Systems and Computing, pages 37-46.
5. S.Krishnaveni , Dr.M.Hemalatha(2011), International Journal of Computer Applications, **Volume 23– No.7**
6. R. Nayak, N. iyatrapoomi, J. Weligamage, Proceedings of the 4th World Congress on Engineering Asset Management (WCEAM 2009), (28-30 September 2009).
7. T. Beshah and S. Hill (2010), Proceedings of AAAI Artificial Intelligence for Development (AI-D'10).
8. X F Zhang; L. Fan, Electrical and Computer Engineering (CCECE), 2013 26th Annual IEEE Canadian Conference on, **vol., no., pp.1,4,** 5-8 (May 2013).
9. Xi Jianfeng; C. Xiaodong; W. Shuangwei; Zhurong Tao, Computer Application and System Modeling (ICCASM), 2010 International Conference on , **vol.13, no., pp.V13-230,V13-233, (**22-24 Oct. 2010).
10. Y. Haixia; N. Zhihong, Computer Science and Education (ICCSE), 5th International Conference on, **vol., no., pp.1355, 1358, (**24-27 2010)
11. D. Emerson, R. Nayak, J. Weligamage, N. Piyatrapoomi, Proceedings of the 5th World Congress on Engineering Asset Management (WCEAM 2010).
12. R. G. Ramani,. S. Shanthi, Computational Intelligence & Computing Research (ICCIC), IEEE International Conference on, **vol., no., pp.1, 4**, (18-20 Dec. 2012).
13. T. Beshah, D. Ejigu, A. Abraham, V. Snasel, Pavel Kromer,"Mining Pattern
14. A. Pakgohar, R. Sigari Tabrizi, M. Khalili, Alireza Esmaeili, World Conference on Information Technology, **Vol 3**, (2011), Pages 764–769.
15. T. Akomolafe, Akinbola Olutayo, American Journal of Database Theory and Application (2012), **1**(3), pages 26-38.
16. J. Han, M. Kamber, J. Pei, *Data Mining: Concepts and Techniques*, 3rd Edition, (June 2011).