Association Rule Mining for Road Traffic Accident Analysis: A Case Study from UK

Chapter · February 2020

DOI: 10.1007/978-3-030-39431-8_50

4 authors, including:

Mingchen Feng
Northwest A & F University

Jinchang Ren
University of Strathclyde

Yue Xi
University of Technology Sydney

Association Rule Mining for Road Traffic Accident Analysis: A Case Study from UK

Mingchen Feng[1(✉)], Jiangbin Zheng[1], Jinchang Ren[2], and Yue Xi[1]

[1] School of Computer Science, Northwestern Polytechnical University, Xi'an, China
{mingchen,yuexi}@mail.nwpu.edu.cn, zhengjb@nwpu.edu.cn

[2] Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, UK
Jinchang.ren@strath.ac.uk

Abstract. Road Traffic Accidents (RTAs) are currently the leading causes of traffic congestion, human death, health problems, environmental pollution, and economic losses. Investigation of the characteristics and patterns of RTAs is one of the high-priority issues in traffic safety analysis. This paper presents our work on mining RTAs using association rule based methods. A case study is conducted using UK traffic accident data from 2005 to 2017. We performed Apriori algorithm on the data set and then explored the rules with high lift and high support respectively. The results show that RTAs have strong correlation with environmental characteristics, speed limit, and location. With the network visualization, we can explain in details the association rules and obtain more understandable insights into the results. The promising outcomes will undoubtedly reduce traffic accident effectively and assist traffic safety department for decision making.

Keywords: Association rules · Data mining · Data visualization · Traffic accident analysis

1 Introduction

Traffic accident analysis continues to be a hot topic in the traffic engineering field. According to global status report on road safety 2018, the number of annual road traffic deaths has reached 1.35 million [1] and around fifty million get injuries or disabilities [2]. Besides, the economic impact is also notable: in a one-year period, the cost of medical care and productivity losses associated with occupant injuries and deaths from motor vehicle traffic crashes exceeded $63 billion [3]. Thus, more research and new models are needed to analyze RTAs in order to discover associated factors and reduce large number of accidents and fatalities. Data mining techniques still need improvement to discover more interesting rules and assist police decision makers in the formulation of new policies and traffic rules from some hidden patterns.

As one of the fundamental algorithms of data mining, association rule mining is an innovative, interdisciplinary, and growing research area, which are used to find all rules in the database satisfying some minimum support and minimum confidence constraints.

J. Ren et al. (Eds.): BICS 2019, LNAI 11691, pp. 520–529, 2020.
https://doi.org/10.1007/978-3-030-39431-8_50

Apriori [4] is a powerful algorithm for mining frequent itemsets for boolean association rules, where two separated steps are performed to generate association rules, i.e. applying minimum support to obtain frequent itemsets and utilizing these itemsets and the minimum confidence constraint to generate rules. Association rules have a unique advantage in discovering frequent patterns with a large dataset.

In this paper, we employ association rule method on UK traffic accident data. Along with some useful rules, we also explored high support and high lift of the rules. With network and other visualization methods, further interpretation of the rules can be obtained.

The major contributions of this paper can be summarized as follows:

(1) By applying association rules, interesting patterns are obtained to investigate the hidden relationship among multi-attributes and factors to RTAs;
(2) With network visualization technique, the association rules obtained above can be easily interpreted, which can provide valuable insights in decision making and in reducing accident risks.

2 Related Work

To date, a number of models have been developed for RTAs analysis. Weng et al. [5] proposed a novel association rules to analyze the characteristics and contributory factors of work zone crash casualties. Montella [6] identified crash characteristics and contributory factors at urban roundabouts and employed the association rule approach to explore their relationships at different crash types. Subasish [7] applied association rules mining method to explore potential patterns of RTAs data under rainy conditions. Gao et al. [8] performed association rules on traffic accidents from Shanghai expressway data and put forward a method to extract strong rules automatically. Priya et al. [9] proposed a technique to categorize the twist of fate information into different classes and carried out apriori algorithm to uncover the semantics of coincidence incidence. Xu [10] used descriptive statistics to illustrate the characteristics of serious casualty crashes in terms of road user behavior, vehicle conditions, geometric characteristics, and environmental conditions and applied association rule mining technique to identify sets of crash contributory factors that often occur together in serious casualty crashes. Das et al. [11] utilized apriori algorithm of supervised association mining technique to discover patterns from the vehicle-pedestrian crash database. Using association rule mining, Gariazzo et al. [12] assessed the relationship between the use of mobile phones at population level and road crash fatalities in large urban areas. Xi et al. [13] utilized association rules to categorize accidental factors and analyze the degree of an accident or the level of influence. Our previous work [14] has visualized and predicted crime trends to discover key factors and patterns related to crimes.

In summary, association rule mining is effective in dealing with datasets which contain large number of attributes. With a proper support and confidence, it can also interpret hidden relationship among them without predetermining the assumptions and functional forms. Moreover, association rule could also reflect the fact that risk factors may exhibit heterogeneous or hidden effects at various circumstances. Thus, we performed association rule mining to explore accident data from UK to uncover more potential patterns.

3 Methodology

3.1 Association Rules

Association rules mining is a well-known data mining algorithm for detecting potential patterns within huge datasets. Among other machine learning methods, association rules mining is flexible due to its no specified function and no dependent variables nature. Guided by the rules, countermeasures can be taken to make quicker decisions and to reduce the risk for accidental incidents. For example, the rules {weather = rainy, light = dark, time = 22–24} ⇒ {accident = rollover} indicates that on rainy days between hours 22 to 24 when the road lights are off, rollover is more likely to happen. Thus, in order to reduce accident we suggest turn on the road lights.

So far, a series of association rules has been put forward, such as apriori and Fp-growth. Apriori was proposed by Agrawal et al. [15] to mine association rules from transaction data. In this paper, we apply Apriori to analyze accident data from UK, the details of this method shown below.

Let $I = \{i_1, i_2, \ldots, i_n\}$ be the set of literals, let $T = \{t_1, t_2, \ldots, t_m\}$ be a set of accident incidents. Each incident in T is a subset of items in I. A rule is defined as $X \Rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \varnothing$. The sets of itemsets X and Y are called antecedent (left-hand-side, LHS) and consequent (right-hand-side, RHS) of the rule.

3.2 Interesting Rule Mining

There are three parameters controlling the number of rules to be generated i.e. Support, Confidence and Lift. Support refers to the proportion of an accident incident, Confidence can be interpreted as an estimate of the probability P(Y|X), and The lift of the rule shows the frequency of co-occurrence of the antecedent and the consequent. They are defined below.

$$Supp(X) = \|\{t \in D | X \subseteq t\}\| / \|t \in D\| \tag{1}$$

$$\mathrm{Conf}(X \Rightarrow Y) = Supp(X \cup Y)/Supp(X) \tag{2}$$

$$\mathrm{lift}(X \Rightarrow Y) = Supp(X \cup Y)/(Supp(X)Supp(Y)) \tag{3}$$

Moreover,

$$Supp(X \cup Y) \geq \sigma \tag{4}$$

$$\mathrm{Conf}(\mathrm{X} \cup \mathrm{Y}) \geq \delta \tag{5}$$

Where, σ and δ are the minimum of support and confidence.

4 Results and Discussion

4.1 Data

The accident data comes from Department for Transport, UK [16], which amassed traffic data from 2005 to 2017, recording over 2 million incidents in the process. Each record includes two types of information: environmental factors and crash information. Environmental factors refer to factors such as road condition, road type, weather, light condition, junction, and speed limit of the road. Crash information contains the detailed information regarding number of vehicles, time, number of casualties and location. It should be pointed out that these data are real-world data and are high in quality, thus the analysis results based on the collected data are plausible.

4.2 Rules Generation and Visualization

To generate interesting rules from UK traffic accident data, we performed Apriori algorithm using package "arrules" provided by R software. The dataset is first transformed to data-frame, and then converted to transaction for further processing. Besides, the minimum support and minimum confidence are set to be 0.4 and 0.7 respectively. After that a totally 195 rules are obtained. As shown in Fig. 1, by utilizing grouped matrix plot, the column is the LHS items which are grouped into 20 groups while the rows are consequents of the rules. The color stands for the number of the lift, and the size of the circles represents the support values. The plot in the top-left corner also shows that there are 3 rules contain {speed_limit = 30, road_surface = dfry}. Besides, we also noticed that the high lift rules and high support rules are separated. So in order to explore more information about support and lift values, after scatter plot of the 195 rules in Fig. 2, a support >0.6 is determined as high support rules and a lift >1.2 is considered as high lift rules.

With support >0.6 as high support rules, we obtain 20 rules, as shown in Table 1. The high support means high proportion of the items i.e. the high frequent occurrence of accidents. We noticed that Severity = Slight, Weather = Fine no high winds, Number_of_Casualties = 1, RoadType = Single carriageway, Light = Daylight are high frequent rules which are highly related to accident. So in order to get rid of traffic accident, traffic department should pay more attention to these characteristics.

With lift >1.2, we get another 20 rules as shown in Table 2, these are all stronger associations, which tell us that speed_limit = 30, area = urban, road_surface = dry and Weather = Fine no high winds are critical factors that cause accident. These rules remind us that we can reduce speed_limit to 20 or 50 (we can conclude from the data that when speed limit is 20 and 50 the accident rate is 1.8% and 3.4% respectively). For the environmental factors, we can not change them, but traffic safety department can set up warning signs to inform drivers for cautious driving.

Table 1. Association rules with high support.

Rules	LHS	RHS	Support	Confidence	Lift
1	{}	{Severity = Slight}	0.847	0.847	1.00
2	{}	{Weather = Fine no high winds}	0.801	0.801	1.00
3	{}	{Number_of_Casualties = 1}	0.770	0.770	1.00
4	{}	{RoadType = Single carriageway}	0.746	0.746	1.00
5	{}	{Light = Daylight}	0.731	0.731	1.00
6	{Weather = Fine no high winds}	{Severity = Slight}	0.675	0.843	0.99
7	{Severity = Slight}	{Weather = Fine no high winds}	0.675	0.797	0.99
8	{Road_Surface = Dry}	{Weather = Fine no high winds}	0.664	0.958	1.19
9	{Weather = Fine no high winds}	{Road_Surface = Dry}	0.664	0.829	1.01
10	{Number_of_Casualties = 1}	{Severity = Slight}	0.659	0.855	1.01
11	{Severity = Slight}	{Number_of_Casualties = 1}	0.659	0.778	1.01
12	{Light = Daylight}	{Severity = Slight}	0.627	0.858	1.01
13	{Severity = Slight}	{Light = Daylight}	0.627	0.739	1.01
14	{RoadType = Single carriageway}	Severity = Slight}	0.626	0.838	0.99
15	Severity = Slight}	{RoadType = Single carriageway}	0.626	0.738	0.98
16	{Number_of_Casualties = 1}	{Weather = Fine no high winds}	0.619	0.804	1.00
17	{Weather = Fine no high winds}	{Number_of_Casualties = 1}	0.619	0.773	1.00
18	{Light = Daylight}	{Weather = Fine no high winds}	0.609	0.834	1.04
19	{Weather = Fine no high winds}	{Light = Daylight}	0.609	0.761	1.04
20	{RoadType = Single carriageway}	{Weather = Fine no high winds}	0.601	0.805	1.00

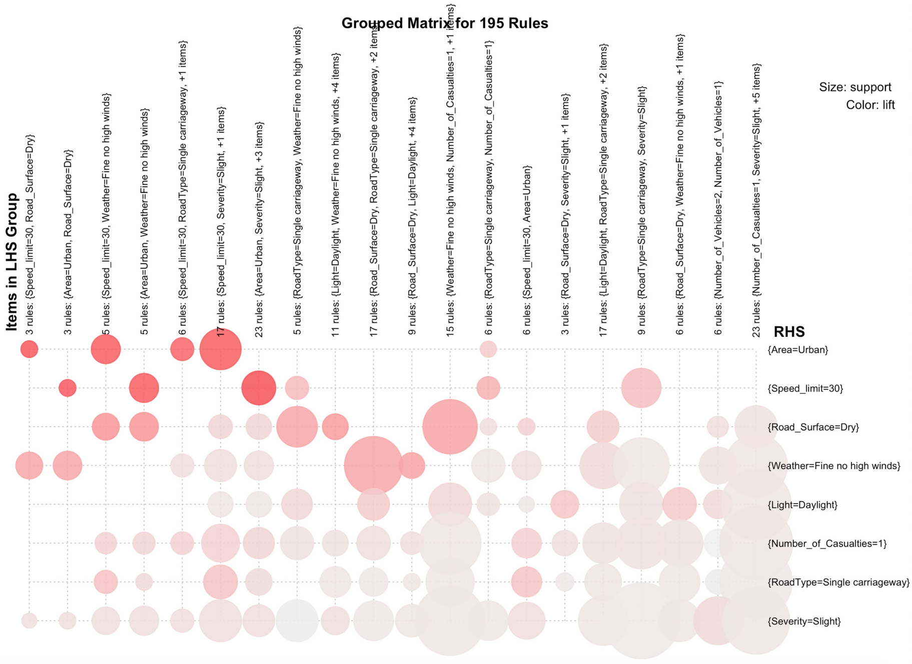

Fig. 1. Grouped matrix plot of the 195 rules. (Color figure online)

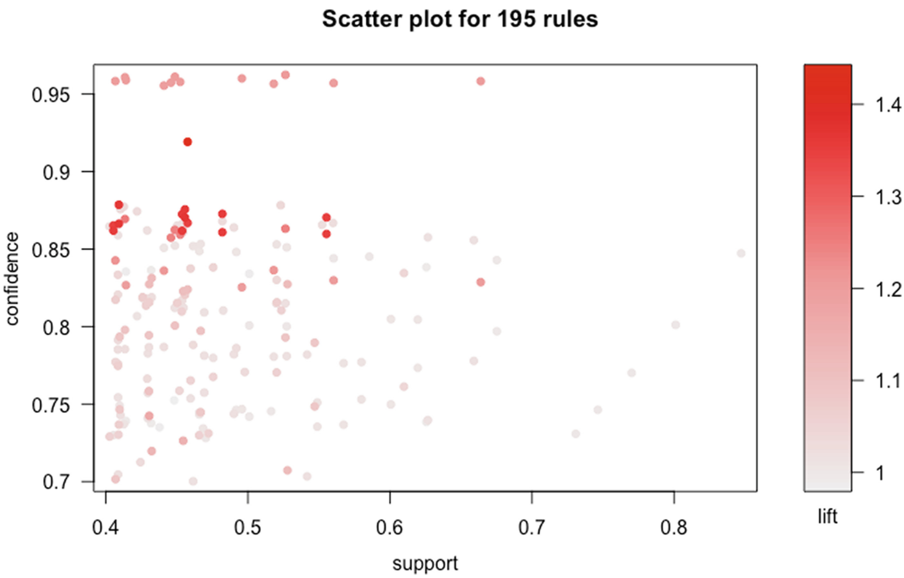

Fig. 2. Scatter plot of the 195 rules. (Color figure online)

Table 2. Association rules with high lift.

Rules	LHS	RHS	Support	Confidence	Lift
1	{RoadType = Single carriageway, Area = Urban}	{Speed_limit = 30}	0.458	0.919	1.44
2	{Number_of_Casualties = 1, Area = Urban}	{Speed_limit = 30}	0.456	0.870	1.36
3	{Speed_limit = 30, Road_Surface = Dry}	{Area = Urban}	0.409	0.878	1.36
4	{Road_Surface = Dry, Area = Urban}	{Speed_limit = 30}	0.409	0.866	1.36
5	{Number_of_Casualties = 1, Speed_limit = 30}	{Area = Urban}	0.455	0.876	1.35
6	{Speed_limit = 30, Severity = Slight}	{Area = Urban}	0.482	0.872	1.35
7	{Light = Daylight, Area = Urban}	{Speed_limit = 30}	0.405	0.862	1.35
8	{Speed_limit = 30, Weather = Fine no high winds}	{Area = Urban}	0.454	0.862	1.35
9	{Weather = Fine no high winds, Area = Urban}	{Speed_limit = 30}	0.454	0.862	1.35
10	{Severity = Slight, Area = Urban}	{Speed_limit = 30}	0.482	0.861	1.35
11	{Area = Urban}	{Speed_limit = 30}	0.555	0.859	1.34
12	{Speed_limit = 30}	{Area = Urban}	0.555	0.870	1.34
13	{Speed_limit = 30, RoadType = Single carriageway}	{Area = Urban}	0.458	0.867	1.34
14	{Speed_limit = 30, Light = Daylight}	{Area = Urban}	0.405	0.865	1.33
15	{Number_of_Casualties = 1, Weather = Fine no high winds, Light = Daylight}	{Road_Surface = Dry}	0.413	0.869	1.25
16	{Weather = Fine no high winds, Light = Daylight}	{Road_Surface = Dry}	0.526	0.863	1.25
17	{Weather = Fine no high winds, Severity = Slight, Light = Daylight}	{Road_Surface = Dry}	0.449	0.862	1.24
18	{Weather = Fine no high winds, Area = Urban}	{Road_Surface = Dry}	0.452	0.859	1.24
19	{Speed_limit = 30, Weather = Fine no high winds}	{Road_Surface = Dry}	0.446	0.857	1.24
20	{Number_of_Vehicles = 2, Weather = Fine no high winds}	{Road_Surface = Dry}	0.407	0.843	1.21

Finally we plot the 100 rules using a network visualization, the plot indicates that road type, light, speed limit and road surface play central roles to accident incidents. Besides, number of vehicles and urban area are also decisive factors for traffic accident. These factors also have some connections, so in order to reduce the risk of traffic accident, we can take proper action to change or eliminate any factors (Fig. 3).

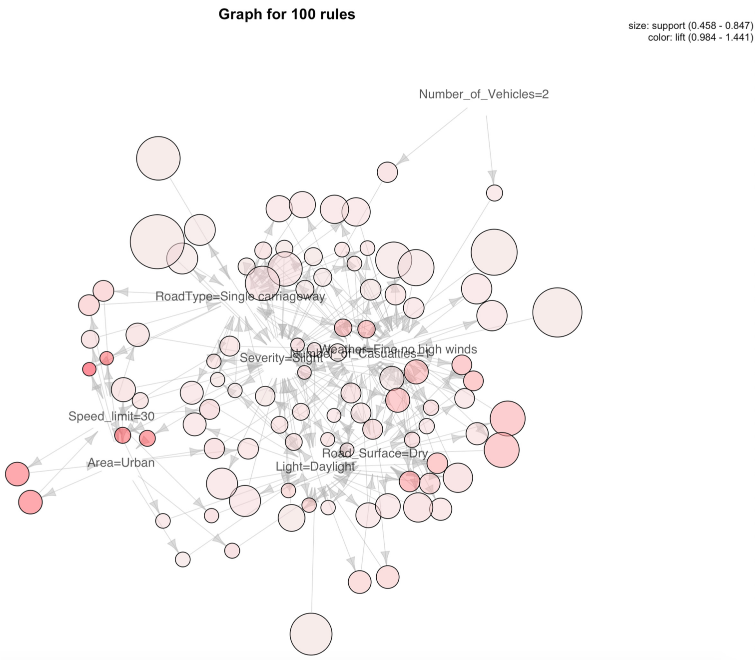

Fig. 3. Network plot of 100 rules.

5 Conclusion and Future Work

In this paper we conduct a study to perform a data mining technique to explore traffic accident data in UK. By mining association rules, we obtain a series of interesting rules and discover critical factors that related to accident. Through using different supports and confidences, we discover the potential reason of the accident. Together with data visualization, a further understanding of the rules and more information are provided.

In the future we will utilize the results to explore more in detail of the accident on how the rules changed in different cities, and whether we can predict or not. Besides, we also want to perform this technique on crime data [17] to mining more interesting rules. Moreover, as Artificial Intelligence [18, 19] and feature selection [20–22] has achieved big success in many field, we also want to seek combination of artificial intelligence algorithms with association rules and add more useful features on our datasets.

Acknowledgement. This work has been supported by HGJ, HJSW and Research & Development plan of Shaanxi Province (Program No. 2017ZDXM-GY-094, 2015KTZDGY04-01).

References

1. World Health Organization: Global status report on road safety 2018. World Health Organization (2018)
2. Road Safety Facts. https://www.asirt.org/safe-travel/road-safety-facts/. Accessed 25 Oct 2018
3. US Department of Health and Human Services, CDC. https://www.cdc.gov/injury/wisqars. Accessed 14 Jan 2018
4. Bhandari, A., Gupta, A., Das, D.: Improvised apriori algorithm using frequent pattern tree for real time applications in data mining. Procedia Comput. Sci. **46**, 644–651 (2015)
5. Weng, J., Zhu, J.Z., et al.: Investigation of work zone crash casualty patterns using association rules. Accid. Anal. Prev. **92**, 43–52 (2016)
6. Montella, A.: Identifying crash contributory factors at urban roundabouts and using association rules to explore their relationships to different crash types. Accid. Anal. Prev. **43**(4), 1451–1463 (2011)
7. Subasish, D., Sun, X.: Investigating the pattern of traffic crashes under rainy weather by association rules in data mining. In: Transportation Research Board 93rd Annual Meeting, No. 14-1540. Transportation Research Board, Washington DC (2014)
8. Gao, Z., Pan, R., et al.: Research on automated modeling algorithm using association rules for traffic accidents. In: 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), Shanghai, pp. 127–132 (2018)
9. Priya, S., Agalya, R.: Association rule mining approach to analyze road accident data. In: 2018 International Conference on Current Trends Towards Converging Technologies (ICCTCT), Coimbatore, pp. 1–5 (2018)
10. Xu, C., Bao, J., et al.: Association rule analysis of factors contributing to extraordinarily severe traffic crashes in China. J. Saf. Res. **67**, 65–75 (2018)
11. Das, S., Dutta, A., et al.: Supervised association rules mining on pedestrian crashes in urban areas: identifying patterns for appropriate countermeasures. Int. J. Urban Sci. **23**(1), 30–48 (2019)
12. Gariazzo, C., Stafoggia, M., et al.: Association between mobile phone traffic volume and road crash fatalities: a population-based case-crossover study. Accid. Anal. Prev. **115**, 25–33 (2018)
13. Deng, X., Zeng, D., Shen, H.: Causation analysis model: based on AHP and hybrid Apriori-Genetic algorithm. J. Intell. Fuzzy Syst. **35**(1), 767–778 (2018)
14. Feng, M., Zheng, J., et al.: Big data analytics and mining for effective visualization and trends forecasting of crime data. IEEE Access **7**(1), 106111–106123 (2019)
15. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. ACM SIGMOD Rec. **22**(2), 207–216 (1993)
16. UK Road Safety Dataset. https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data
17. Feng, M., Zheng, J., Han, Y., Ren, J., Liu, Q.: Big data analytics and mining for crime data analysis, visualization and prediction. In: Ren, J., Hussain, A., Zheng, J., Liu, C.-L., Luo, B., Zhao, H., Zhao, X. (eds.) BICS 2018. LNCS (LNAI), vol. 10989, pp. 605–614. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00563-4_59

18. Yan, Y., Ren, J., et al.: Cognitive fusion of thermal and visible imagery for effective detection and tracking of pedestrians in videos. Cogn. Comput. **10**, 94–104 (2017)
19. Yan, Y., Ren, J., et al.: Unsupervised image saliency detection with Gestalt-laws guided optimization and visual attention based refinement. Pattern Recogn. **79**, 65–78 (2018)
20. Cao, F., Yang, Z., Ren, J., et al.: Local block multilayer sparse extreme learning machine for effective feature extraction and classification of hyperspectral images. IEEE Trans. Geosci. Remote Sens. **57**, 5580–5594 (2019)
21. Sun, H., Ren, J., et al.: Superpixel based feature specific sparse representation for spectral-spatial classification of hyperspectral images. Remote Sens. (MDPI) **11**(5), 536 (2019)
22. Zhang, A., Sun, G., Ren, J., et al.: A dynamic neighborhood learning-based gravitational search algorithm. IEEE Trans. Cybern. **48**, 436–447 (2017)