ISBN: 978-981-18-3959-7
2022 the 12th International Workshop on Computer Science and Engineering (WCSE 2022)
doi: 10.18178/wcse.2022.06.018

Characterization of Urban Road Traffic Accidents based on Data Mining

Weifan Zhong [1] Jie Yuan [2], Yue Ren [3] and Lijing Du [4+],

Wuhan University of Technology, China

Abstract. Road traffic safety is a social issue of concern, and China is one of the countries with the largest number of total road traffic accidents and accident fatalities. The occurrence of traffic accidents mainly includes four factors: people, vehicles, facilities, and the environment. Under the background of big data, to explore the relationship between various factors of urban road traffic accidents and analyze the characteristics of traffic accidents given the current situation of frequent traffic accidents and serious losses, this paper collects week, time period, weather, road conditions, alarm subcategories, collision types, and accident vehicles as research factors. The overall association rule, week and time period association rule, road and climate association rule, and accident vehicle association rule are applied to the road traffic accident factors. The results show that there is a correlation between road traffic accident factors and accident outcomes, revealing important factors that cause traffic accidents.

Keywords: urban roads, traffic accidents, accident features, association rules, Apriori algorithm

1. Introduction

According to the report released by the World Health Organization (WHO), [1] about 1.35 million people die in urban road traffic accidents every year. On average, one person dies on the road every 24 seconds. Road traffic injuries are the leading cause of death among children and young people aged 5 to 29, and the global road traffic safety situation is very serious.[2-3] In September 2019, China issued the "Outline of the Construction of a Strong Transportation State", which proposed to "promote data resources to empower transportation development" and "build a comprehensive transportation big data center system", which clearly defined the development of comprehensive transportation big data in the new period. Big data has become an important carrier to realize the interconnection of the comprehensive transportation system and an important tool to improve the governance capability and service level of China's transportation industry. To search the influencing factors in the problem of traffic accident, many scholars have done research on it.

S.Luo et al.[4] used the severity fields of a single accident in the dataset as data labels, the data fields related to the spatial environment and the data fields describing the basic attributes of the accident as data features, and built an analysis and prediction model based on logistic regression models and random forest models of influencing factors for the whole accident data based on various environmental factors; M.Manzoor et al. [5] used random forest and deep learning algorithms to predict the severity of traffic accidents; P.Kromer et al.[6] used artificial evolution and fuzzy system machine learning methods to mine traffic accident features based on factors such as accident time, accident vehicle, and accident location; C.Ting [7] used the random forest, XGBoost, CART, neural network, plain Bayesian and support vector machine for prediction; Y.Ding [8] used logistic whitening weight clustering algorithm and apriori algorithm for urban traffic accident driver characteristics; Y.Zhang [9] used data mining methods such as the random forest, multiple linear regression and ARIMA time series to propose a traffic accident combination analysis and prediction model to comprehensively analyze the causes of traffic accidents from road and environment perspectives; F.Zhou [10] used multicategorical logistic regression model and Bayesian network model to explore about traffic factors such as vehicle and environment; in many studies of traffic accident

+ Corresponding author. Tel.: + 86 18086635191
E-mail address: dulijing@whut.edu.cn.

characteristics at present, many relevant studies focus on one aspect of the research factors, and also use one aspect of the factors for accident severity prediction, lacking the study between different factors of accidents.

2. Data Resource

2.1. Data Collection

The authors taken urban road traffic accidents as the research object, and collected road traffic accident data from the Traffic Management Bureau to ensure the accuracy and plurality of the data, which contained more than 4,000 pieces of data including Week, Slot, Weather, Road_Conditions, Alarm_Categories, Active_Hit, Postive_Hit, Collision_Type, Casualties, Road_section, Rodatype,Speed and type of driver. This paper analyzes some previously ignored data types

2.2. Data Collation

The general traffic accident influencing factors are divided into four aspects: person, vehicle, weather and environment. Since the factors of person are highly subjective and susceptible to environmental factors, after screening, the data content includes only some environmental factors and some vehicle factors, and according to the characteristics of the data, the influencing factors are divided into the factors of the time of day, climate of road and vehicle in accident. The constructed urban road traffic accident factors and conditions consisted of nine fields, as shown in Figure 1.

	A	B	C	D	E	F	G	H	I	J	K
1	Week	Slot	Weather	Road_Condi	Alarm_Cate	Active_Hit	Positive_H	Collision_	Casualties		
2	wed	dawn	sunny	dry	vehicle	vehicle	vehicle	crash	N		
3	wed	dawn	sunny	dry	vehicle	vehicle	motorcycle	crash	Y		
4	wed	morning	sunny	dry	vehicle_bc	electric_v	motorcycle	crash	Y		
5	wed	morning	sunny	dry	vehicle_bc	vehicle	electric_v	crash	N		
6	wed	morning	sunny	dry	vehicle	vehicle	vehicle	crash	N		
7	wed	morning	sunny	dry	vehicle	vehicle	other	other	N		
8	wed	forenoon	sunny	dry	vehicle	vehicle	vehicle	crash	N		
9	wed	forenoon	sunny	dry	vehicle_bc	vehicle	electric_v	crash	Y		
10	wed	forenoon	sunny	dry	vehicle_bc	vehicle	electric_v	crash	Y		
11	wed	forenoon	sunny	dry	vehicle_bc	vehicle	electric_v	crash	Y		
12	wed	forenoon	sunny	dry	vehicle	vehicle	vehicle	crash	N		
13	wed	forenoon	sunny	dry	other	vehicle	other	other	N		
14	wed	forenoon	sunny	dry	vehicl_pec	vehicle	pedestriar	crash	Y		
15	wed	forenoon	sunny	dry	both_non	electric_v	electric_v	crash	Y		
16	wed	forenoon	sunny	dry	vehicle_bc	vehicle	electric_v	crash	Y		
17	wed	forenoon	sunny	dry	vehicle_bc	bus	electric_v	crash	Y		
18	wed	forenoon	sunny	dry	vehicle	vehicle	vehicle	crash	Y		
19	wed	forenoon	sunny	dry	vehicle_bc	truck	electric_v	crash	Y		
20	wed	forenoon	sunny	dry	vehicle	vehicle	motorcycle	crash	Y		
21	wed	forenoon	sunny	dry	vehicle_bc	vehicle	electric_v	crash	Y		
22	wed	forenoon	sunny	dry	vehicle	off_road	motorcycle	crash	Y		
23	wed	noon	sunny	dry	unilateral	van	other	other	Y		
24	wed	noon	sunny	dry	vehicle	motorcycle	small_truc	crash	Y		
25	wed	noon	sunny	dry	vehicle	vehicle	vehicle	crash	N		
26	wed	noon	sunny	dry	vehicle_bc	vehicle	electric_v	crash	Y		
27	wed	noon	sunny	dry	vehicle	vehicle	vehicle	crash	Y		
28	wed	noon	sunny	dry	vehicle_bc	vehicle	electric_v	crash	N		
29	wed	afternoon	sunny	dry	other	motorcycle	animal	crash	N		
30	wed	afternoon	sunny	dry	vehicle	bus	vehicle	crash	N		

Fig. 1: Traffic accident factors and conditions table

3. Development of Methodology

3.1. Introduction of the Apriori Algorithm

Through data mining, we reveal hidden, unknowable and mining-worthy information from the large amount of data in a database. Data mining association rule mining is a frequency rule based machine learning algorithm that finds relationships in large databases. It aims to obtain the strong association rules present in the database using many metrics. The Apriori algorithm uses an iterative method of layer-by-layer search to find the relationships of sets of items in a database. [11] It has been widely used in data mining. There are many algorithms in data mining, such as k-means algorithm, Support Vector Machine, C4.5. Apriori algorithm uses Apriori property to produce candidate item sets, which greatly compresses the size of frequent sets and achieves good performance. The generated data features play a more significant role in traffic data analysis. So in this paper, Apriori algorithm is our first choice.

3.2. Degree of Support (X=>Y)

Degree of support is the frequency of an item set, which is a method of the significance of association rules and is used to indicate the importance of this set. It is assumed that the rate of the number of deals containing itemset X and Y to the total number of deals countAll reflects the frequency of simultaneous occurrence of itemset X and Y.

$$support(X => Y) = \frac{count(X \cap Y)}{countAll}$$

3.3. Degree of Confidence (X=>Y)

Degree of Confidence is used to determine the frequency of occurrence of Y in the transactions containing X, i.e., the conditional probability of Y conditional on X. It is a method of the accuracy of the association rule and indicates the reliability of the rule. Suppose the rate of the number of deals containing the item set X, Y to the number of deals in the item set X.

$$Confidence(X => Y) = \frac{support(X \cap Y)}{support(X)}$$

3.4. Frequent Item Set and Min Degree of Support

Min degree of Support is a pre-set lower limit for the itemset to satisfy the support, denoted by Min_sup, which reflects the minimum importance of the itemset of interest. X is a frequent itemset when the Support of the itemset X is not less than the minimum support threshold.

$$support(X => Y) = \frac{count(X)}{countAll} \geq Min\_sup$$

3.5. Strong Association Rules and Minimum Degree of Confidence

The minimum degree of Confidence is a pre-set lower limit of the Confidence level that the item set satisfies, denoted by Min_conf, which reflects the minimum reliability of the item set of interest. An association rule R is said to be strongly associated when it satisfies both support and confidence not less than a minimum threshold.

$$\begin{cases} support(X => Y) \geq Min\_sup \\ confidence(X => y) \geq \min\_conf \end{cases}$$

3.6. Lifting Degree (X=>Y)

Lifting degree represents the ratio of the probability of occurrence of thing Y given a known condition X, to the probability of occurrence of thing Y alone. Defined as the ratio of the confidence level to the frequency of the posterior term, it indicates the influence of the antecedent condition on the posterior matter.

$$lift(X => Y) = \frac{support(X => Y)}{support(X) \times support(Y)}$$

3.7. Linked Data of the Accidents in Data Mining

The occurrence of traffic accidents has many initiating factors, and the joint action of these factors promotes the final occurrence of accidents. Through the characteristics of the algorithm, we can analyze the hidden relationship among the factors. We will analyze from three aspects below, which are rarely mentioned and excavated before.

4. Experimental Research

4.1 Overall Linked Data

In the first step we perform data mining on the overall dataset and start the overall study of the association rule. Based on the dataset, the single dataset whose Support is greater than 0.5 are shown in the clustered bar Figure 2. As can be seen in the graph, the weather conditions are mainly cloudy, the probability of producing casualties in accidents is higher than the probability of not producing casualties, the traffic path where accidents occur are often congested and busy, the types of impact are mainly collisions, the vehicles

driven by both the crushers and the crushed are mainly small vehicles, the road conditions are mainly dry, and the road types are mainly straight roads. The week and time of day data do not have a large Support with the other data, in fact, their Support does not exceed 0.2, so we can assume that the other factors of accident occurrence and the week and time of day are not related.

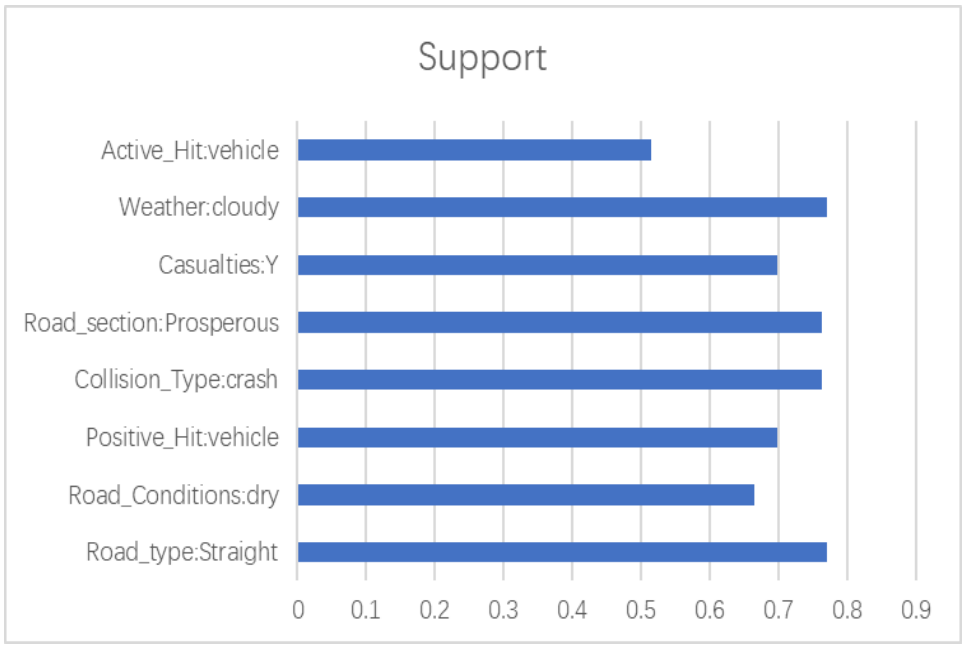

Fig. 2: Incident elements with support greater than 0.5.

However, the association rules between accident factors cannot be fully obtained by analyzing only a single factor. In fact, accidents always occur under the mutual influence and superposition of multiple factors, so the two-dimensional and three-dimensional correlations of the factor association rules of traffic accidents are our next step to do. We set the minimum Support to 0.3 and the minimum Confidence to 0.4 to get the Figure 3 of the distribution of the correlation dataset. It can be found that the data are highly concentrated and has strong correlation.

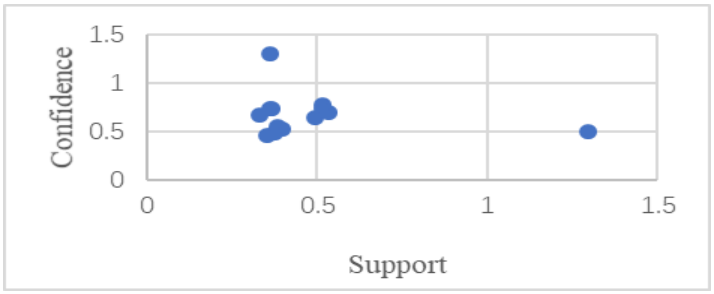

Fig. 3: Hotspot distribution of support and confidence for the overall correlation portfolio

Table 1: Traffic accident association rules

Number	Leading field	Post field	Pre support	Post support	Total support	Confidence	Lifting degree
1	**Road_type:Straight**	**Road_Conditions:dry**	**0.665**	**0.771**	**0.516**	**0.776**	**1.007**
2	**Positive_Hit:vehicle**	**Collision_Type:crash**	**0.764**	**0.698**	**0.535**	**0.7**	**1.004**
3	**Collision_Type:crash**	**Casualties:Y**	**0.698**	**0.494**	**0.384**	**0.551**	**1.116**

4	**Collision_Type:crash**	**Road_section:Prosperous**	**0.698**	**1.297**	**0.515**	**0.739**	**1.004**
5	**Road_Conditions:dry**	**Casualties:Y**	**0.771**	**0.494**	**1.297**	**0.496**	**1.005**
6	**Road_type:Straight**	**Casualties:Y**	**0.665**	**0.494**	**0.331**	**0.67**	**1.008**
7	**Casualties:Y**	**Road_section:Prosperous**	**0.494**	**0.736**	**0.364**	**0.737**	**1.001**
8	**Road_Conditions:dry**	**Weather:cloudy**	**0.771**	**0.495**	**0.495**	**0.642**	**1.297**
9	**Road_Conditions:dry**	**Road_type:Straight Road_section:Prosperous**	**0.771**	**0.456**	**0.352**	**0.457**	**1.001**
10	**Positive_Hit:vehicle**	**Collision_Type:crash Road_section:Prosperous**	**0.764**	**0.515**	**0.396**	**0.518**	**1.006**
11	**Road_Conditions:dry**	**Active_Hit:vehicle Weather:cloudy**	**0.771**	**0.375**	**0.375**	**0.486**	**1.297**
12	**Weather:cloudy**	**Road_Conditions:dry Road_section:Prosperous**	**0.495**	**0.561**	**0.363**	**0.732**	**1.305**
13	**Road_Conditions:dry**	**Weather:cloudy Road_section:Prosperous**	**0.771**	**0.363**	**0.363**	**1.297**	**1.297**

Through data mining of association rules of Apriori algorithm, we extracted the overall thirteen association rules of the accident, whose association rules are shown in Table 1. As shown in the table, from two-dimensional data to three-dimensional data, we can consider that these thirteen accident factors combination all have a certain strong interconnection, which has important tips for accident prevention. We know that road sections where crashes occur are generally prosperous and congested, which makes it necessary to pay attention to reducing the probability of other accident factors, thus reducing the accident rate.

4.2 Accident and Vehicle Association Data

In this article, by data mining the association data of accident factors and vehicles, the Support of the elements with minimum Support of 0.2 obtained is shown in Figure 4, which shows the one-dimensional focus elements of accidents and vehicles. The high-dimensional data association is shown in Table 2, which shows that the association of casualties occurring when an electric vehicle is hit is high.

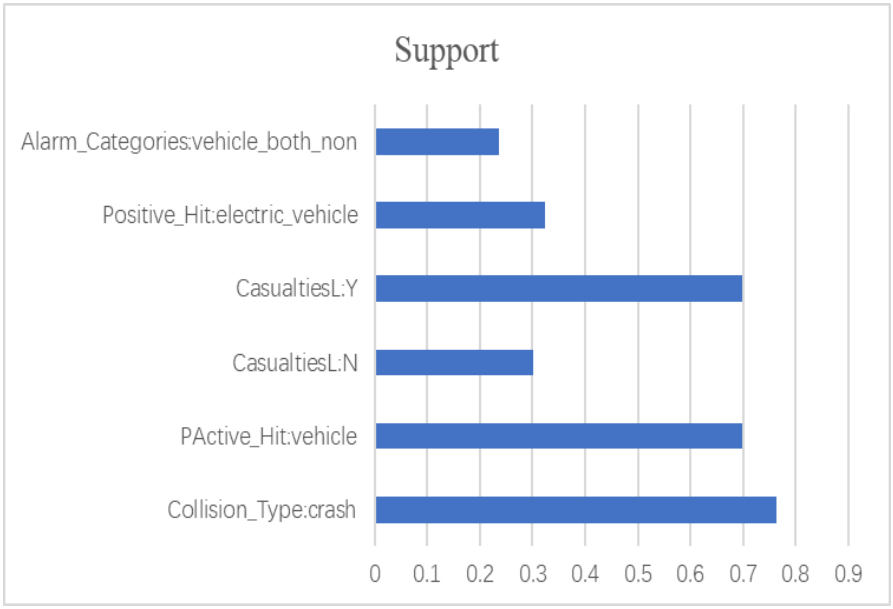

Fig. 4: Accident and vehicle dataset support

Table 2: Accident and vehicle association rules

Number	Leading field	Post field	Pre support	Post support	Total support	Confidence	Lifting degree
1	**Collision_Type:cras**	**Positive_Hit:electric_v**	**0.698**	**0.764**	**0.535**	**0.767**	**1.004**
2	**Casualties:Y**	**Collision_Type:crash**	**0.494**	**0.698**	**0.384**	**0.778**	**1.116**
3	**Alarm_Categories:v**	**Positive_Hit:electric_v**	**0.237**	**0.323**	**0.208**	**0.876**	**2.711**
4	**Active_Hit:vehicle**	**Collision_Type:crash Casualties:N**	**0.764**	**0.313**	**0.287**	**0.376**	**1.201**
5	**Collision_Type:cras h**	**Casualties:Y Positive_Hit:electric_v ehicle**	**0.698**	**0.28**	**0.22**	**0.315**	**1.124**

4.3 Week and Period Linked Data

We have conducted a separate data mining of the week and period data to try to obtain some association relationships from them, and we obtained the Support and Confidence Figure 5. But from the graphs, we unfortunately found that the association rules for the week and period are not obvious, and the Support only exists between 0.024 and 0.032, and the Confidence is only distributed between 0.14 and 0.24, so we can assume that there is no association relationship for the week and period. We can conclude that there is no association between week and period.

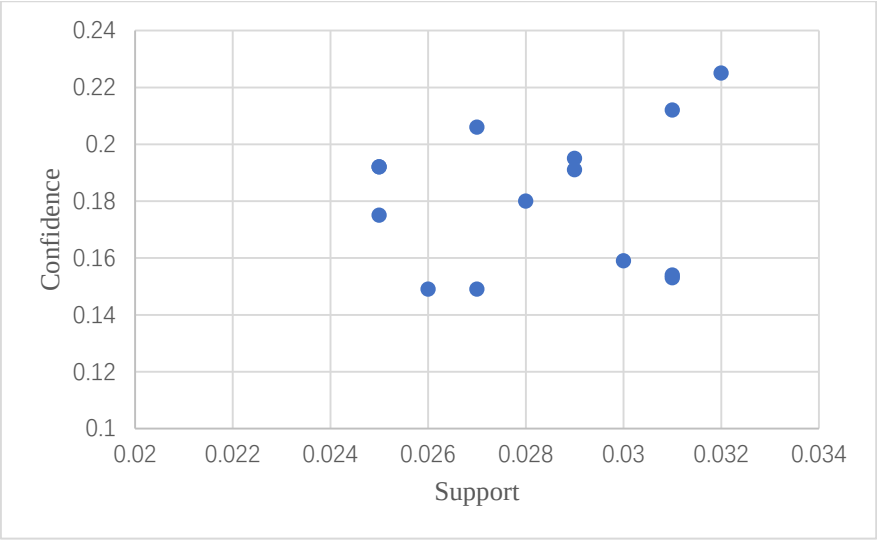

Fig. 5: Hotspot distribution of weekday dataset

4.4 Road and Climate Correlation Data

By data mining the road-climate dataset, we will explore the association rules that exist between roads and climate. The support degree of association data is obtained Figure 6, from which we can see that the span of its support degree is large, the minimum is 0.264 and the maximum is 0.771, but we can get some datasets with association relationship from it.

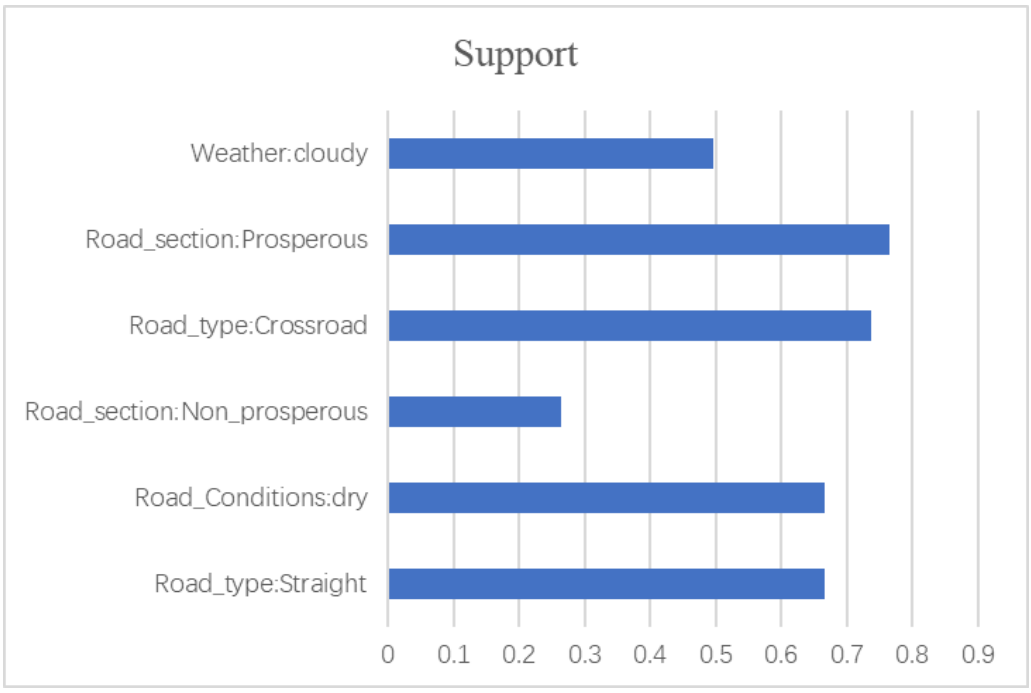

Fig. 6: Road and climate dataset support

We performed a high-dimensional association rule data mining process on the road climate dataset and obtained Table 3, from which a set of four-dimensional numbers Road_Conditions:dry, Road_type:Straight, Weather:cloudy, and Road_section:Prosperous have strong correlation and can be that these four conditions have strong correlation rules and are prone to congestion on straight roads during wet roads and cloudy weather.

Table 3: Accident and vehicle association rules

Number	Leading field	Post field	Pre support	Post support	Total support	Confidence	Lifting degree
1	**Road_type:Straigh**	**Road_Conditions:dry**	**0.665**	**0.771**	**0.516**	**0.776**	**1.007**
2	**Road_section:Non_prosperous**	**Road_Conditions:dry**	**0.264**	**0.771**	**0.21**	**0.796**	**1.032**
3	**Road_type:Straigh**	**Road_section:Non_prosperous**	**0.665**	**0.264**	**0.209**	**0.314**	**1.19**
4	**Road_type:Crossroad**	**Road_section:Prosperous**	**0.335**	**0.736**	**0.28**	**0.836**	**1.135**
5	**Weather:cloudy**	**Road_Conditions:dry**	**0.495**	**0.771**	**0.495**	**1.0**	**1.297**
6	**Road_Conditions:dry**	**Road_type:Straigh** **Road_section:Prosperous**	**0.771**	**0.456**	**0.352**	**0.457**	**1.001**
7	**Road_type:Crossroad**	**Road_Conditions:dry** **Road_section:Prosperous**	**0.335**	**0.561**	**0.209**	**0.624**	**1.112**
8	**Weather:cloudy**	**Road_Conditions:dry** **Road_section:Prosperous**	**0.495**	**0.561**	**0.363**	**0.732**	**1.305**
9	**Road_Conditions:dry**	**Road_type:Straight** **Weather:cloudy** **Road_section:Prosperous**	**0.771**	**0.225**	**0.225**	**0.292**	**1.297**

5. Conclusion

The general traffic accident factors are divided into three aspects: person, vehicles, and environment. In this paper, we have achieved more significant results by data mining Apriori algorithm on overall linked data, accident and vehicle linked data, week and period linked data, and road and climate linked data. For example, if the type of road is straight, the weather is cloudy, the road condition is dry, and the congestion condition is congested, the possibility of traffic accidents is greatly increased. Knowing the intrinsic relationship between the factors influencing accidents helps us to increase the vigilance of this type of road and better avoid accidents. It can be seen that data mining is very useful for discovering the intrinsic connections of traffic accidents, and the results achieved in this paper through data mining of actual data from traffic management departments are more credible in helping to improve the way of traffic accident warning, and provide a new way of judging the importance of traffic warning factors, which can effectively improve the traffic occurrence in urban areas of China.

6. Acknowledgements

We would be grateful to the anonymous reviews and the Editorial office for their constructive and thorough review. The authors wish to express sincere appreciation to Lijing Du of School of Safety Science and Emergency Management, Wuhan University of Technology, for her helpful comments on the drafts of the paper. This research was funded by the National Natural Science Foundation of China, grant number

72104190 and 72042015; Foundation of Social Science and Humanity, China Ministry of Education, grant number 20YJC630018; Hubei Provincial Natural Science Foundation, grant number 2020CFB162.

7. References

[1] World Health Organization. Global Status Report on Road Safety2015[R]. *Geneva：World Health Organization*，2015.

[2] L. Guo, J. Zhou, S. Dong. Analysis of urban road traffic accidents based on improved K-means algorithm[J]. In: L. Guo, et al. *Chinese Journal of Highways*. 2018, 31(4): 270-279.

[3] Y. Guo, P. Liu, Y. Wu. A traffic conflict model based on Bayesian multivariate Poisson-log-normal distribution[J]. In: Y. Guo, et al. *Chinese Journal of Highways*.2018, 31(1): 1-9.

[4] S. Luo. Spatial Feature Analysis of Road Traffic Accidents Based on Data Mining Technology[D]. Tsinghua University, 2019.

[5] M. Manzoor, M. Umer, C. Bisogni. RFCNN: Traffic Accident Severity Prediction Based on Decision Level Fusion of Machine and Deep Learning Model. *IEEE ACCESS 9*. 2021, pp.128359-128371.

[6] P. Kromer, T. Beshah, A. Abraham. Mining Traffic Accident Features by Evolutionary Fuzzy Rules. *IEEE Symposium on Computational Intelligence in Vehicles and Transportation Systems (CIVTS).* 2013, pp. 38-43.

[7] C. Ting, N. Tan, A. Shabadin. Malaysian Road Accident Severity: Variables and Predictive Models. *6th International Conference on Computational Science and Technology (ICCST).* 2020, pp.699-708.

[8] Y. Ding. Research on driver characteristics of urban traffic accidents based on data mining [D]. Shenyang University,2017.

[9] Y. Zhang. Accident causation analysis and black spot identification based on urban road big data[D]. Tongji University,2018.

[10] F. Zhou. Analysis of the causes of urban road traffic accidents based on Bayesian networks [D]. Hunan Normal University,2017.

[11] Y. Ding. Research on data mining techniques based on Apriori algorithm[J]. *Modern computing: second half of the month edition*, 2012.