Analysis of Causes of Traffic Accidents Based on Improved Apriori Association Rules

Chunhe Shi[a,*], Yu Ding[b], Gaofeng Yue[c], and Yinghong Xie[d]

School of Information Engineering, Shenyang University, Shenyang 110044, China

[a]schsydx@163.com; [b]384800457 @qq.com; [c]643375731 @qq.com; [d] Xieyinghong @163.com

Keywords: Association Rules; Apriori Algorithm; Data Mining; Road Traffic

Abstract: Road traffic Safety is a public safety issue, with the people of deaths from traffic accidents each year accounting for the highest proportion of total deaths due to all safety incidents. With the development of big data intelligent analysis technology, it is helpful to put forward targeted measures to prevent the occurrence of traffic accidents. This paper uses the traffic accident data source of a city in southern China to extract the related factors of traffic accident, such as time, weather, location and the type of accident, and then uses Apriori algorithm to mine the related factors, find out the various combination factors that lead to the accident, thus summed up the law of multiple traffic accidents. Some of the conclusions drawn from these rules could be made available to the authorities in order to take preventive and regulatory measures to reduce the incidence of accidents.

1. Introduction

With the sustained and rapid development of economy and society, the people's living standards gradually improve, and the rigid demand for automobiles to maintain a strong situation. According to statistics, as of March 2017, China's motor vehicle ownership exceeded 300 million, including 200 million vehicles, the average annual increase reached the highest level in history, motor vehicle drivers more than 364 million people, nearly five years an average annual increase of 24.5 million people. With the increasing number of urban motor vehicle, traffic congestion, traffic safety accidents and other problems occur frequently in all countries of the world every day, while China is one of the countries with the highest death rate from traffic accidents[1]. The annual traffic accident accounts for about 80% of all safety incidents [2], which not only caused huge economic losses, but also seriously threatened the safety of people's lives . According to data released by the National Bureau of Statistics, in 2015, a total of 187781 traffic accidents occurred in the country, the death toll was 58022, the number of injuries was 199880, resulting in direct property damage of 1,036,920,000 yuan[3]; Although the number of accidents and casualty losses decreased in the last 5 years, the relevant figures are still not often amazing . There are subjective factors in the causes of traffic accidents, as well as objective factors. Subjective factors include the driver's personal factors; objective factors include vehicle factors, road factors and environmental factors. Traffic accidents are accidental, but also because under certain conditions, due to the human, car, road, Environment composed of the dynamic traffic system in a certain link caused by the imbalance.

The frequent occurrence of traffic accidents and the huge losses caused by them have attracted people's attention, and a large number of studies have been carried out to analyze the causes of traffic accidents. The research shows that traffic accidents are not the result of the action of a single factor, but the combination of multiple factors.

This paper uses the traffic accident data source of a city. Firstly, it has been preprocesses, extract the relevant factors of traffic accident, such as the main factors such as time, weather, location and the type of accident, and then use Apriori algorithm to mine the related factors, find out the various combinations of factors leading to the accident, and then summarize the law of multiple traffic accidents. Some conclusions drawn from these rules can be made available to the relevant departments to facilitate their qualitative analysis of the frequent combinations of multiple traffic

DOI: 10.25236/meici.2019.056

accidents, so as to take preventive and regulatory measures to reduce the occurrence of accidents.

2. Apriori Algorithm

Apriori algorithm is a classic association rule mining algorithm proposed by R. Agralwal et al. The basic idea of the algorithm is to first find out all the frequent itemset, and then use the frequent itemset found in the first step to generate strong association rules, which must meet the minimum support and minimum confidence.

The Apriori algorithm has two main steps, namely, the two steps of the connection step and the pruning step to find the largest frequent item sets.

Apriori uses breadth-first search and a Hash tree structure to count candidate item sets efficiently. It generates candidate item sets of length k from item sets of length k-1. Then it prunes the candidates which have an infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent k-length item sets. After that, it scans the transaction database to determine frequent item sets among the candidates.

Mining the association rules by frequent sets, it is necessary to mine the matching association rules from the obtained frequent items. The method adopted is to traverse all the frequent itemset, and then take 1, 2, ..., k elements from each frequent item set as the latter, and the other elements in the project set as the predecessor, according to the support and confidence. The calculation formula is calculated and filtered, and the definition of association rule and confidence [4] is as follows:

The association rule is an implication of the form X→Y, where X∩Y=Z, and the support of the association rule X→Y is the percentage of the transaction inclusion (X∪Y) in D, that is, the probability P(X∪Y), which is Eq.1:

$$Support(X \to Y) = \frac{|X \cup Y|}{|D|} = P(X \cup Y) \tag{1}$$

The confidence of the association rule X→Y is the percentage of the transaction in the D that contains the X transaction and also contains the Y, that is, the conditional probability P(Y/X), which is Eq.2:

$$\text{Confidence}(X \to Y) = |X \cup Y||X| \times 100\% = P(Y / X) \tag{2}$$

From the calculation formula of the confidence, an association rule based on the example can be generated.

3. Improvement of Apriori Algorithm

3.1 Row Vector Algorithm for Boolean Matrix (RVABM).

The row vector algorithm of the Boolean matrix sets the number of items included in the item set to m. The number of items k (k≤m) contained in each transaction in the transaction database D is calculated, and the number of all items with the same k value is counted. N, if N is greater than or equal to the given minimum support number, then k is the number of items of the largest frequent item, and find all candidate item sets containing k items. Then determine whether these itemset are frequent k itemset. If a frequent k itemset is a frequent item set, then all of its non-empty subsets are frequent itemset, and all frequent itemset can be found. The flow chart of the row vector algorithm is shown in Fig 1.

3.2 Column Vector Algorithm for Boolean Matrix (CVABM).

The column vector algorithm of the Boolean matrix can calculate the number of occurrences of each item in the item set 1 (that is, the number of "1") through. If it is greater than or equal to the given minimum support degree, the item belongs to the frequent 1 itemset. If any number of "1" is greater than or equal to the given minimum support degree, the combination of the two items belongs to the frequent 2 itemset. By analogy, when a frequent k item set is required, two eligible items are connected from the known frequent k-1 itemset, and the items in the connected item set are AND operation. If it is greater than or equal to the minimum support, the item belongs to the frequent k itemset. The column vector algorithm flow is shown in Fig 2.

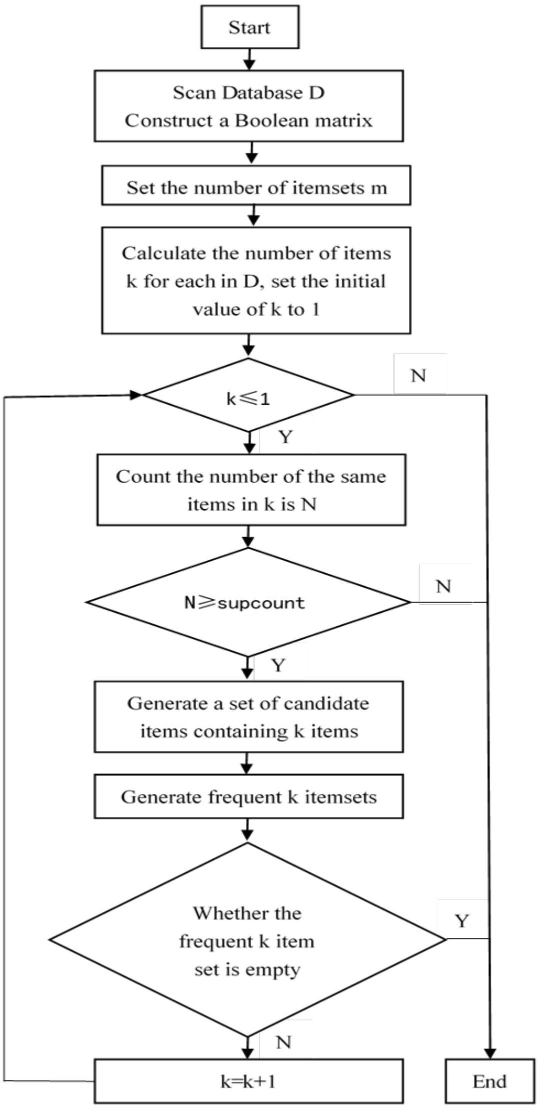
Start
Scan Database D
Construct a Boolean matrix
Set the number of itemsets m
Calculate the number of items k for each in D, set the initial value of k to 1
k≤1
N
Y
Count the number of the same items in k is N
N≥supcount
N
Y
Generate a set of candidate items containing k items
Generate frequent k itemsets
Whether the frequent k item set is empty
Y
N
k=k+1
End

Fig.1 Row vector algorithm flow chart

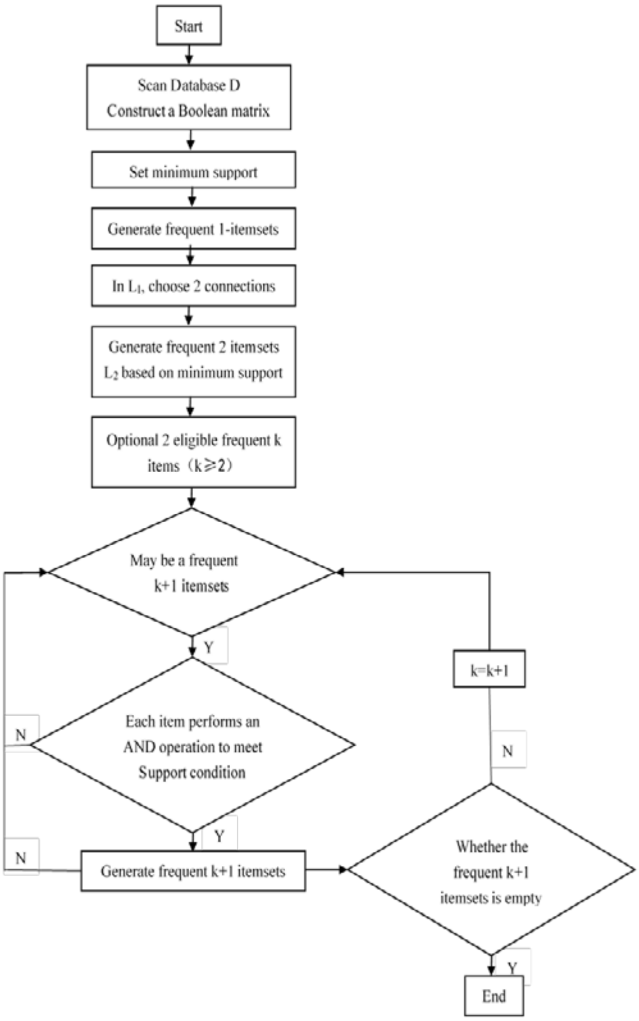
Start
Scan Database D
Construct a Boolean matrix
Set minimum support
Generate frequent 1-itemsets
In L1, choose 2 connections
Generate frequent 2 itemsets
L2 based on minimum support
Optional 2 eligible frequent k items（k≥2）
May be a frequent k+1 itemsets
Y
k=k+1
N
Each item performs an AND operation to meet Support condition
N
Y
N
Generate frequent k+1 itemsets
Whether the frequent k+1 itemsets is empty
Y
End

Fig.2 Column vector algorithm flow chart

3.3 Algorithm Performance Analysis.

In order to test the performance of improved Apriori algorithm, the test data set Mushroom (transaction number 8124, 23 transactions per transaction) was selected as experimental data, and the algorithm performance test was performed by MATLAB platform. The results of the operation are shown in Fig 3. In Fig. 3, the comparison of the time taken to mine the frequent itemset is set in the case where the same number of transactions and the minimum support degree are different. Experiments show that, because the improved algorithm only needs to scan the database once, and use the row vector or column vector mining process to improve the mining efficiency of frequent k items set by increasing the frequent k-1 item set as much as possible, and improve the running speed. The improved Apriori algorithm can also perform frequent item set mining, which is completely better than the original Apriori algorithm.

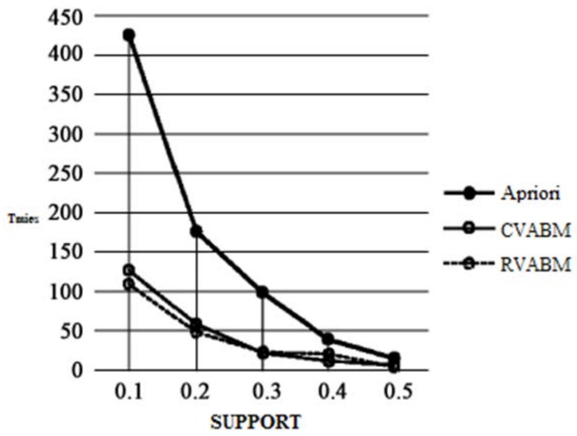

Fig.3 Performance comparison between Apriori algorithm and improved algorithm under different support

4. Research on the Traffic Accident Based on Apriori Algorithm

In the process of association rules analysis, users may be more interested in certain items, and hope to enhance their role when mining, while weakening the role of some other items. For example, for a large shopping mall, “high-grade perfume” and “platinum diamond ring” are valuable commodities, and their prices and profits are much higher than ordinary daily necessities “milk” and “bread”. Merchants are more concerned about the sales of “Platinum Diamond Ring” and “Advanced Perfume”, but because of the high price of “Advanced Perfume” and “Platinum Diamond Ring”, the sales total is very low, that is, the support in the database is very small, it is a non- frequent itemset, so even if the "Platinum Diamond Ring" => "Advanced Perfume" is a highly credible rule, it will be mined by traditional association rules because {"Platinum Diamond Ring", "Advanced Perfume"} is a non-frequent item sets. The algorithm ignores it. In contrast, "milk" and "bread" are sold in large quantities and are frequent itemset that will be included in the excavated rules. If they are treated equally, it will be difficult to extract information about the products that the merchant is highly concerned about.

Just like expensive goods and cheap goods, the proportion of driver segments in the data varies greatly, such as the proportion of male drivers is 91.72%, while female drivers only account for 8.28%. If the driver characteristics are directly correlated to a large extent, it does not reflect the potentially valuable rule information. Therefore, the driver's age, driving age, and gender distribution are correlated and analyzed according to the distribution of each segment in the statistical analysis to eliminate the influence of uneven distribution of feature data,and get more valuable rules.

4.1 Age Feature Association Analysis.

In the statistical analysis, the driver's age is divided into 7 segments. Now the driver of 18~24 years old, the driver of 32~45 years old and the driver of 60~70 years old are selected as representatives to analyze the relationship between different age characteristics and other dimensional data. Shown in Table.1, Table.2, Table.3.

Table 1 Association analysis results of 18 to 24 years old drivers

Rules		Support	Confidence
[7,8),Rear-end	=>Rate=high	1.1285%	100.0000%
[7,8),Grab the line	=>Rate=high	1.2111%	100.0000%
5M,Rear-end	=>Rate=high	2.8902%	100.0000%
9M,Rear-end	=>Rate=high	2.8902%	100.0000%
18~24years old,Grab the line,red	=>Rate=high	2.8076%	100.0000%
[16,17),Grab the line,male	=>Rate=high	2.8076%	100.0000%

Table 2 Association analysis results of 32 to 45 years old drivers

Rules		Support	Confidence
11M,black, No died no injuries	=>Rate=high	1.3173%	99.2857%
0~5Y,4M,white	=>Rate=high	1.8290%	99.4845%
[16,17),Rear-end	=>Rate=high	1.9186%	96.6184%
[18,19),balck	=>Rate=high	1.0265%	100.0000%
0~5Y,Rear-end,black	=>Rate=high	1.3239%	98.5714%
2M,Rear-end	=>Rate=high	1.2183%	98.4496%

Table 3 Association analysis results of 60 to 70 years old drivers

Rules		Support	Confidence
female,Rear-end	=>Rate=high	2.4020%	100.0000%
60~70years old,blue	=>Rate=high	2.5284%	100.0000%
[21,22),male	=>Rate=high	2.4020%	100.0000%
[21,22),60~70years old	=>Rate=high	2.7813%	100.0000%
[17,18),6~15Y	=>Rate=high	2.2756%	100.0000%
12M,60~70years old,Rear-end	=>Rate=high	2.2756%	100.0000%

Analysis of the correlation characteristics of different age groups, we can find that the accident rate of the driver in all ages is high in the morning and evening peak. The rear-end collisions and the grab the line of type accident rate is higher, but there is still a difference in the age of each age group. Drivers aged 18 to 24 have a high rate of rear-end collisions in May and September. Drivers aged 32 to 45 have a high rate of rear-end collisions in February. Drivers aged 60 to 70 occur in December. The accident rate of rear-end collision is high; and the accident rate of new drivers in the 32-45 age group is high in March and April; the driver of 60-70-year-old driver has many types of traffic accidents, and the accident rate is high at 16~22 clock. This is directly related to the vision of the elderly.

4.2 Analysis of Driving Age Characteristics.

We divided the driver's driving experience into 7 segments. Now we select the novice driver who is driving for 0~5 years and the driver who is driving for more than 15 years as the representative to analyze the relationship between different driving age characteristics and other dimensional data. Shown in Table.4, Table.5.

Table 4 Association analysis results of 0 ~ 5 years driving drivers

Rules		Support	Confidence
[7,8),white	=>Rate=high	2.2873%	99.7802%
[16,17),39~45 years old	=>Rate=high	1.1134%	96.5066%
39~45years old,black	=>Rate=high	2.0303%	98.2927%
46~52years old,white	=>Rate=high	3.3755%	97.8102%
0~5Y,18~24years old,female	=>Rate=high	1.3703%	97.4910%
female,white	=>Rate=high	4.4284%	97.9933%

Table 5 Association analysis results of more than 15 years driving drivers

Rules		Support	Confidence
8M,black	=>Rate=high	1.4668%	100.0000%
[11,12),≥15Y,black	=>Rate=high	1.2060%	100.0000%
12M,46~52years old	=>Rate=high	2.6402%	97.5904%
12M,≥15Y,46~52years old	=>Rate=high	2.6402%	97.5904%
female,Rear-end	=>Rate=high	1.9231%	95.9350%

By analyzing the correlation characteristics of drivers with low driving age and high driving age, it can be found that the driver of 0~5 years driving age has a high peak accident rate in the morning and evening, but the driver of the driving age of 15 years or more has a high accident rate at noon; the older female novice driver accident The rate is high; the driver of high driving age has a high accident rate in August and the driver accident rate of high driving age is high in 46~52 years old in December; the high accidental age of female drivers is high; the driver of 46~59 years old with high driving age has other types of accidents. The rate is high.

4.3 Correlation Analysis of Gender Characteristics.

We divided the driver gender into male drivers and female drivers, and analyzed the association between gender characteristics and other dimensional data. Shown in Table.6, Table.7.

Table 6 Association analysis results of male drivers

Rules		Support	Confidence
[8,9),white	=>Rate=high	4.7504%	100.0000%
0~5Y,25~31years old,Rear-end	=>Rate=high	4.5537%	100.0000%
3M,6~15Y	=>Rate=high	4.2444%	100.0000%
0~5Y,Reversing	=>Rate=high	4.0741%	100.0000%
0~5Y,[8,9),white	=>Rate=high	2.0982%	100.0000%
[16,17),Rear-end	=>Rate=high	2.0766%	100.0000%

Table 7 Association analysis results of female drivers

Rules		Support	Confidence
18~24years old,Rear-end	=>Rate=high	2.3737%	100.0000%
18~24years old,[8,9)	=>Rate=high	1.1041%	100.0000%
18~24years old	=>Rate=high	7.7008%	99.2883%
18~24years old,No died no injuries	=>Rate=high	7.6456%	99.2832%
18~24years old,Local car	=>Rate=high	7.5076%	99.2701%
18~24years old,Not allowed to go	=>Rate=high	2.7325%	99.0000%

By analyzing the related characteristics of drivers of different genders, it can be found that the male novice driver has a high accident rate and the female young driver has a high accident rate; the male driver's accident rate is high in March and December 6 to 15 years; 18 to 24 years old The female driver’s accident rate was high in December; the female driver driving a white silver vehicle had a high accident rate.

5. Summary

This paper first introduces the Apriori algorithm of association rules, and illustrates the data mining process of Apriori algorithm by example. Then, the association characteristics of the driver's age, driving age, gender and other dimension data are excavated separately, and the frequent items of the driver characteristics are obtained. The same accident rules and different accident rules of different feature intervals of the same feature are analyzed. By comparing and analyzing the rules of different feature intervals of the same feature, some potential, useful and valuable information is found.

Acknowledgements

Liaoning Provincial Natural Science Foundation Funded Project (20170540644, 20170540646), Shenyang Science and Technology Plan Project (18013015), the Doctoral Initiation Fund of Liaoning Province (20170520322)

References

[1] Kun Shen: Safety and Environmental Engineering, Vol.24 (2017) No.5, p.138.

[2] Hua Fei: China Economic Weekly, (2017) No.7, p.78.

[3] Chenlu Qiu, Jun Ji, Huiying Xu: Police Technology, (2017) No.3, p.29.

[4] R.Agralwal, R.Srikant: Proceedings of the 20th International Conference on Very Large Databases (Santiago, Chile, 1994). P487.

[5] Zhemin Yang: Software, (2017) No.11, p.71.

[6] Xinliang Li, Xiangtao Chen: Computer Engineering & Science, Vol.29 (2007) No.12, p.111.

[7] Chunhe Shi, Chengdong Wu, Xiaowei Han, et al: 6th International Conference on Electronic, Mechanical, Information and Management Society (Shenyang, China, April 1-3, 2016). Vol. 40, p.108.

[8] Chunhe Shi, Chengdong Wu, Xiaowei Han, et al: 6th International Conference on Electronic, Mechanical, Information and Management Society(Shenyang, China, April 1-3, 2016). Vol. 40, p.301.