Cause Analysis of Traffic Accidents on Urban Roads Based on an Improved Association Rule Mining Algorithm

QIURU CAI
School of Computer Engineering, Jiangsu University of Technology, Changzhou 213001, China
e-mail: cai_qiuru@sina.com

This work was supported in part by the National Natural Science Foundation of China under Grant 61401226.

ABSTRACT The traffic accidents on urban roads are result of joint actions between multiple factors, namely, human, vehicle, road and environment. To identify the key causes to such accidents, it is necessary to mine the association rules between relevant risk factors out of the statistics on these accidents. Considering the multiple layers and dimensions of accident data, this paper improves the Apriori algorithm to mine the association rules between risk factors, and probes deep into the causes of traffic accidents on urban roads. According to the layer and dimension of specific attributes, the parameters like support, confidence and lift were adjusted to find the qualified association rules between risk factors. The results were further screened to obtain a series of meaningful association rules. The research results enable the traffic department to formulate pertinent accident control measures, and promote the traffic safety on urban roads.

INDEX TERMS Urban roads, causes of traffic accidents, data mining, association rules, Apriori algorithm.

I. INTRODUCTION
Traffic accidents on urban roads are an inevitable problem in urban transport, causing massive casualties and property losses. Currently, China has achieved good results on traffic safety, and curbed road traffic safety from further deterioration. However, the traffic safety situation in urban areas of China is not yet optimistic. Statistics from the World Health Organization (WHO) show that China ranks the 98th among 178 countries in road traffic death rate (10.45 deaths per 10,000 vehicles). According to China’s statistics on traffic accidents in 2015, an average of 23 accidents occur on every 100kms of urban roads in China, which is far greater than that on expressways, first-class highways and second-class highways.

Considering the severity of traffic accidents on urban roads, Chinese and foreign scholars and experts have explored deep into the influencing factors and analyzed the causes of urban road accidents. Concerning the influencing factors, 80-90% traffic accidents are related to human factors like the driver [1], [2]. For instance, Larsen and Kines [3] drew the following conclusions after investigating specific types of accidents: the main causes of head-on collisions include excessive speed, drunk driving and driving under the influence of illegal drugs; the drivers involved in these accidents are all males; in the left-turn accidents, the most common causes are attention errors and advanced age. McGwin and Brown [4] discussed the types and causes of traffic accidents that are common among drivers in different age groups, revealing that old drivers are prone to turning and changing lanes due to slow reaction and lack of observation, while young drivers may face accidents because they are risk-taking and unskilful. Ballesterors et al. [5] found that the youngest drivers have the highest rate of traffic accidents under safe road conditions.

With the application of mathematical statistics in traffic accident analysis, many scholars have attempted to identify the causes of traffic accidents from multiple perspectives. Based on the records of 1,606 accidents, Abdel-Aty and Radwan [6] disclosed the influence of multiple factors (i.e. traffic volume, vehicle speed, number of lanes and lane width) on the frequency of traffic accidents through negative binomial modeling, and evaluated how the influence of each factor varies with the genders and ages of the drivers. Chen et al. [7] applied a hierarchical Bayesian logistic model to examine the significant factors on the severity of driver injury, and concluded that: injury severity is enhanced by road curve, type of accident, as well as the gender, age and alcohol/drug...
involvement of the driver, and suppressed by wet road surface, male drivers and seatbelt use.

Concerning the cause analysis, the causes of traffic accidents are often examined by aggregate methods like linear regression [8], clustering [9], [10] and time series [11], and disaggregate methods like Bayesian network [12] and discrete selection model [13]. Treating all accidents as a whole, aggregate methods are easy to implement, but not comprehensive enough to consider the features of each accident. Meanwhile, disaggregate methods are complicated in computation, but excel in modelling effectiveness and prediction accuracy.

In addition, many intelligent models, which are based on machine learning, have been introduced to risk modelling of traffic accidents on urban roads, namely, neural network (NN) [14], Bayesian neural network (BNN) [15], and support vector machine (SVM) [16]. Compared with traditional models, these models are good at linear or nonlinear approximation. Nevertheless, the generalization ability of such models is rather limited, calling for verification against massive measured data [17], [18].

The above analysis shows that the distribution laws of influencing factors can be obtained through statistical analysis of traffic accidents on urban roads, shedding light on the causes of such accidents [19]. However, it is impossible to identify the associations between multiple risk factors of traffic accidents through data analysis on the statistical distribution level, not to mention mining out the key accident chains from accident data [20].

To solve these defects, the classic Apriori algorithm was improved from the angles of data structure, measuring indices, and subjective constraints, in the light of the statistical features of traffic accidents on urban roads. The algorithm flow was optimized by the R programming language. In the improved algorithm, the association rules were mined in four steps: data processing, modelling of multidimensional data, algorithm mining and rule interpretation. Then, the improved Apriori algorithm was adopted to identify the key causes of traffic accidents on urban roads, through multidimensional and multilayered mining of accident data. The improved algorithm proposed in this paper further improves the analysis method system of causes of urban road traffic accidents, which has a strong practical significance.

The remainder of this paper is organized as follows: Section 2 explains how to mine the association rules between risk factors of traffic accidents on urban roads with the improved Apriori algorithm; Section 3 verifies the improved Apriori algorithm with actual data; Section 4 puts forward the conclusions.

II. METHODOLOGY

Traffic accidents on urban roads are attributable to many risk factors, which are complexly associated with each other. To understand the accident causes, it is helpful to firstly identify the association rules between these risk factors.

Many analysis strategies have emerged for risk factors of traffic accidents on urban roads. Most of them only evaluate the accident causes by a single, static index, considering linear factors (e.g. slope, horizontal curve radius, and turning angle), driver factors (e.g. gender and age), and traffic conditions (e.g. traffic flow, vehicle speed, and road occupancy).

On actual roads, however, traffic accidents are the result of joint actions between multiple factors, namely, human, vehicle, road and environment. These factors often have complicated associations. At present, most of the studies on the causes of urban road traffic accidents have not been able to divide these interrelated rules effectively. It is urgent to propose an association rule mining algorithm based on the statistical characteristics of urban road traffic accidents to improve the efficiency and accuracy of accident cause analysis.

Therefore, this paper attempts to mine the causes of traffic accidents on urban roads based on association rules. Firstly, the statistics on such accidents were processed into structured data. Next, the Apriori algorithm was improved to extract the association rules between influencing factors of the accidents. The extracted rules lay the basis for traffic department to formulate risk control measures.

A. APRIORI ALGORITHM

The Apriori algorithm is a famous association rule mining method proposed by Agrawal et al. in 1993. The algorithm adopts a two-phase iterative search for frequent itemsets [21]: (1) Find all itemsets for which the support is greater than the threshold support in the entire database through level-wise search; (2) Create rules from each frequent itemset by binary partition, and look for the ones with high confidence [22].

The Apriori algorithm, a milestone of data mining, provides a level-wise 1D data mining method that finds frequent itemsets in a dataset for Boolean association rule [23]. The algorithm is simple in structure, and easy to understand, involving no complex formula derivation. The relevant concepts and formulas are defined as follows:

Let $T = \{t_1, t_2, \ldots, t_n\}$ be the set of transactions called database (the database of traffic accidents on urban roads), and $I = \{i_1, i_2, \ldots, i_n\}$ a set of $n$ attributes called items (the accident attributes). Each transaction $t_i = (i = 1, 2, \ldots, n)$ corresponds to a subset of $I$ ($t_i \subseteq I$) (every record of traffic accident contains several accident attributes).

A rule can be defined as an implication, $A \Rightarrow B$, where $A_i (i = 1, 2, \ldots, n)$ and $B_j (j = 1, 2, \ldots, m)$ be two subsets of $I$ (two accident attributes in a record of traffic accident). $A \Rightarrow B$ implies that, if $A$ occurs in an accident, $B$ will also occur ($A \subseteq I$, $B \subseteq I$ and $A \cap B = \emptyset$).

There are three key indices in the Apriori algorithm, namely, support, confident and lift. The support of association rule $A \Rightarrow B$ in database $T$ refers to the proportion of accidents involving both $A$ and $B$ in all accidents:

$$
\text{Support} (A \Rightarrow B) = P(A \cup B) = \frac{|\{t | A \cup B \in t, t \in T\}|}{|T|}
$$

(1)
The confidence of association rule $A \Rightarrow B$ in database $T$ refers to the probability for an accident containing $A$ to contain $B$, i.e. the proportion of accidents involving both $A$ and $B$ to those containing $A$:

$$\text{Confident} (A \Rightarrow B) = \frac{P(B | A)}{P(A, B)} = \frac{P(A, B)}{P(A)} = \frac{\text{Support} (A \Rightarrow B)}{\text{Support} (A)} \quad (2)$$

The lift of $A \Rightarrow B$ in database $T$ [24] refers to the ratio of the probability for an accident to contain both $A$ and $B$ to that for an accident to contain $B$ only:

$$\text{Lift} (A \Rightarrow B) = \frac{P(B | A)}{P(B)} = \frac{\text{Confident} (A \Rightarrow B)}{\text{Support} (B)} = \frac{\text{Support} (A \Rightarrow B)}{\text{Support} (B) \cdot \text{Support} (A)} \quad (3)$$

B. IMPROVED APRIORI ALGORITHM

According to the above analysis on association rule mining, all the attributes must be converted into the same dimension, before applying the Apriori algorithm to mine the association rules between risk factors of traffic accidents on urban roads. This normalization process might reduce the scanning efficiency of the database in the search for frequent itemsets, and output meaningless search results. In other words, the Apriori algorithm is not directly applicable to the mining of the multilayer, multidimensional dataset of traffic accidents on urban roads.

Therefore, the Apriori algorithm was improved to suit the association rule mining of the multilayer, multidimensional data of traffic accidents on urban roads. Specifically, each accident attribute was regarded as a predicate. Since the accident data are multilayered, the layers of attributes related to the mining demand were selected for association analysis. In this way, the multilayer mining problem is converted into single-layer mining problem, and the associations between different layers of attributes could be revealed. This method can avoid a large number of meaningless errors in the output results of traditional association rule algorithm due to too many factors, and improve the efficiency of data mining algorithm to some extent. The flow of the improved Apriori algorithm is explained as follows:

Step 1. Generate the frequent itemset.

Input: an $n$-dimensional database; the minimum threshold support, min-sup.

Output: an $n$-dimensional frequent itemset.

(1) Initialize the parameters: $k=1$, $L = \emptyset$ and $C_1 = \emptyset$;
(2) Generate the 1$^{st}$-$n$-dimensional candidate itemset $C_1$;
(3) Generate the 1$^{st}$ frequent itemset $L_1$;
(4) Generate the k-th candidate itemset $C_k$ based on the (k-1)$^{th}$ frequent itemset $L_{k-1}$;
(5) Generate the k-th frequent itemset $L_k$ based on the k-th candidate itemset $C_k$; Calculate the sup of each item in $C_k$ by formula (1), and add all the items whose sup $\geq$ min-sup into $L$;
(6) Obtain the $n$-dimensional frequent itemset $L_n$, and go to Step 2.

Step 2. Generate the set of multidimensional association rules $R$

Input: Frequent itemset L; the minimum threshold confidence, min-conf.

Output: A set of association rules $R$.

(7) Initialize $R = \emptyset$;
(8) Calculate the conf and lift between the nonempty subsets of each frequent itemset by formulas (2) and (3), respectively;
(9) Add all the association rules whose conf $\geq$ min-conf and lift $\geq$ 1 into $R$.
(10) Terminate the algorithm.

III. EXAMPLE ANALYSIS AND DISCUSSION

A. DATA COLLECTION AND PROCESSING

The mining of association rules is greatly affected by the structure of data samples on traffic accidents on urban roads. The mining efficiency can be enhanced if the data are structured. However, the traffic accident records have many missing items. Before applying to association rule mining, the original records should be processed into structured data. Based on the original records, the structured data are often obtained in three steps (Figure 1): data cleaning, data transformation and data reduction.

After visiting traffic departments, the authors learned that the statistics of traffic accidents on urban roads in China are
Currently collected and archived by traffic police division, and most data are kept confidential.

Considering statistical availability, the authors paid a visit to the traffic police division in a city in northern China, and acquired 589 records on traffic accidents in that city, which happened between January and September, 2018.

The 589 records were processed by the workflow in Figure 1, creating 576 valid records of structured data. Each valid record contains 15 attributes. The serial number, name, type, and meaning of each attribute are listed in Table 1 below.

The data of traffic accidents on urban roads belong to multiple layers and dimensions. It is unwise to directly apply all the accident attributes to the association rule mining between risk factors. For one thing, the numerous attributes will drag down the mining efficiency. For another, it is difficult to find useful association rules with high confidences between the multidimensional data on different layers.

Based on multilayer multidimensional association rules, this paper puts forward an accident risk analysis approach to mine the data of traffic accidents on urban roads. To facilitate the mining of association rules, the accident attributes in the dataset of traffic accidents on urban roads were divided into three layers (Figure 2).

As shown in Figure 2, the data samples on traffic accidents of urban roads are divided into three layers. On the first layer, there are four attributes, including accident, time, external environment, and road condition.

On the second layer, the four primary attributes were broken down into 15 secondary attributes: accident was split into four parts (accident form, accident cause, accident type, and accident location); time was split into three parts (season, week, and hour); external environment was split into four parts (terrain, weather, illumination and visibility); the road condition was split into five parts (pavement condition, intersection/section type, road shape and isolation facility).

On the third layer, each secondary attribute was further split into a number of tertiary attributes. The full list of all attributes is provided in Table 1.

**B. ASSOCIATION RULE MINING**

1) ASSOCIATION RULES BETWEEN SECONDARY ATTRIBUTES UNDER ACCIDENT

From the arules package for R programming language, the Apriori algorithm was selected and improved to suit multilayer, multidimensional data, according to the procedure in Subsection 2.2.

Firstly, the targets of association rule mining were selected, considering the features of attributes on different layers. Here, the mining focuses on the association rules between tertiary attributes of accident and those of other secondary attributes. Secondly, the parameter thresholds of the association rule mining algorithm were adjusted for each target. Finally, the association rules thus acquired were analyzed, in view of the actual meaning of each attribute.

The causes of traffic accidents on urban roads can be analyzed in the following steps:

**Step 1. Selection of analysis targets:** The attributes to be analyzed are selected based on their levels, forming a dataset of relevant attributes.
TABLE 1. The attributes of traffic accidents on urban roads.

| Serial number | Primary attributes | Secondary attributes | Tertiary attributes                                                                 | Remark   |
|---------------|--------------------|----------------------|-------------------------------------------------------------------------------------|----------|
| 1             | Accident form      | Accident             | Rollover, hitting pedestrians, running over pedestrians, crash into stationary vehicles, collision with moving vehicles, bumping into non-fixed objects, bumping into fixed objects, crash, and others. | 9 items  |
| 2             | Accident cause     | Accident             | Driving against regulations, parking against regulations, fatigue driving, pedestrians entering expressways, driving without holding a driver's license, driving with other behaviors that undermine driving safety, unqualified vehicle, overload/over speed, driving a vehicle not specified on the driver's license, other improper operations, tailgating, escape after creating a traffic accident (which does not constitute a crime), and drunk driving. | 13 items |
| 3             | Accident type      | Accident             | Fatal accident, injury accident, and property loss accident                           | 3 items  |
| 4             | Accident location  | Accident             | Motor vehicle travel lane, emergency lane, and others                                | 3 items  |
| 5             | Season             | Time                 | Spring, summer, autumn and winter                                                  | 4 items  |
| 6             | Week               | Time                 | Workday, and holiday                                                                | 2 items  |
| 7             | Hour               | Time                 | Morning, noon, afternoon, dusk, evening, and late night                             | 6 items  |
| 8             | Terrain            | External environment | Plains, mountains, and hills                                                       | 3 items  |
| 9             | Weather            | External environment | Sunny, foggy, snowy, overcast, rainy                                               | 5 items  |
| 10            | Illumination       | External environment | Day, dusk, dawn, evening without street lighting, and evening with street lighting | 5 items  |
| 11            | Visibility         | External environment | <50m, 50-100m, 100-200m, and >200m                                                | 4 items  |
| 12            | Pavement condition | Road condition       | Dry, wet, and ponding                                                               | 3 items  |
| 13            | Road shape         | Road condition       | Straight, gentle curve/slope, and sharp curve/steep slope                           | 3 items  |
| 14            | Intersection/section type | Road condition | Ordinary section, ramp entrance/exit, elevated section of bridge/tunnel, junction of branch roads, and other special sections | 5 items  |
| 15            | Isolation facility | Road condition       | Corrugated beam barrier, isolation pier (column), concrete barrier, metal barrier, green belt, and others | 6 items  |

Step 2. Parameter setting: The minimum support (\(\text{min-sup}\)), minimum confidence (\(\text{min-conf}\)) and lift (\(\text{lift}\)) are determined according to the accident attributes for the association rule mining.

Step 3. Association rule mining: Following the flow of the improved Apriori algorithm, the frequent itemset satisfying the minimum support is solved, and the qualified association rules are then identified based on the minimum threshold confidence and threshold support.

Step 4. Interpretation and application of association rules: The qualified association rules are further analyzed and screened, and the consequents are extracted as the association rules between accident form, accident type and accident cause. Next, the practical meanings of these association rules are interpreted to guide the application.

According to the abovementioned flow of association rule mining for risk factors, the authors probed into the relationship between the three secondary attributes under accident, namely, accident form, accident type and accident cause.

Firstly, the urban road traffic accident dataset, which only contains no other secondary attribute but accident form, accident type and accident cause, was inputted to the improved algorithm. Next, the initial parameter threshold is finally determined by analyzing the results of several trials. The minimum threshold support (\(\text{min-sup}\)) and minimum threshold confidence (\(\text{min-conf}\)) were initialized as 0.1 and 0.25, respectively, and 21 initial association rules were found to satisfy these initial thresholds.

The visualization package arulesViz of R programming language was adopted to prepare the scatter plot for the 21 rules. The scatter plot is presented as Figure 3, where x-axis is support, y-axis is confidence, and the shade of green color is the degree of lift.

It can be seen from Figure 3 that the supports of some association rules fell between 0.1 and 0.2, and those of other association rules ranged between 0.25 and 0.35, indicating the coexistence of high- and low-frequency qualified association rules. In terms of confidence, most association rules fell between 0.4 and 0.7, and only a few remained below 0.4. This means the initial threshold confidence is too small. Thus, the confidence was increased to 0.4 to facilitate rule screening. Moreover, half of all rules were below 1 in terms of lift, revealing that many rules fail to satisfy the constraint that the lift should be greater than 1.
The above analysis shows the necessity of parameter adjustment, before further screening effective association rules. Due to the severity of the consequences of traffic accidents, the association rules between important yet low-frequency risk factors might be filtered out, if the threshold support is too high. Hence, the minimum threshold confidence was reset to 0.4 and lift to 1, without changing the minimum threshold support. In this way, a total of 10 association rules were found to satisfy constraints: \( \text{sup} > 0.1, \text{conf} > 0.4 \) and \( \text{lift} > 1.0 \).

Considering the meanings of secondary attributes under accidents, the consequents cannot be the causes of accident, but the form and type of accident. In addition, the relationship between accident form and accident type is only meaningful, when accident form is taken as the antecedent and accident type as the consequent. Thus, the inspect function of the arules package was called to extract the association rules for the consequents of accident type and accident form. The association rules with practical value were thus obtained (Table 2).

As shown in Table 2, a total of 5 meaningful association rules were obtained between the secondary attributes under accident. The following laws can be extracted from these rules: (1) The fatal traffic accidents on urban roads are mostly caused by driving with other behaviors that undermine driving safety; (2) Collisions with moving vehicles often lead to injury accidents; (3) The collisions with moving vehicles, which arise from driving with other behaviors that undermine driving safety, are very likely to cause fatal accidents; (4) Crashes into stationary vehicles often result in fatal accidents; (5) Tailgating can easily bring about collision with moving vehicles. Based on these rules, the traffic department can prepare countermeasures to reduce the accident risk.

2) ASSOCIATION RULES BETWEEN ATTRIBUTES UNDER ACCIDENT AND THOSE UNDER THE OTHER PRIMARY ATTRIBUTES

In the preceding subsection, the association rules between secondary attributes under accident were mined out, reflecting the correlations between accident type, accident form and accident cause. This subsection aims to identify how the causes of traffic accidents on urban roads correlate with the attributes under other primary attributes. For this purpose, four primary attributes (accident, time, road condition and external environment), 15 secondary attributes and 74 tertiary attributes were taken into account.

The improved apriori algorithm was applied to mine the association rules between attributes under accident and those under the other primary attributes. Firstly, the thresholds (minimum support, minimum confidence and lift) were adjusted to find the effective association rules. A lot of effective rules were obtained, due to the sheer number of tertiary attributes. To facilitate subsequent analysis, the subset function and sort function of the arules package were called to further screen the set of effective rules. Finally, the useful association rules were obtained, and used to interpret the correlations between attributes under accident and those under the other primary attributes.

(1) Association rules between attributes under time and those under accident

If all attributes were considered, the extracted association rules will change with the frequency of multiple attributes. To disclose how the attributes under time affect those under accident, three secondary attributes under time (season, week, and hour) and the secondary attributes under accident (e.g. accident type and accident form) were selected as the targets of association rule mining. A total of 37 effective association rules were obtained through the abovementioned procedure. For the lack of space, the ten association rules with the highest confidence (Table 3) were selected for further analysis.
The following laws can be extracted from these rules:
(1) On workdays in summer, the traffic accidents occurring at deep night are often fatal, most of which are collisions with moving vehicles; (2) The accidents on workdays are often fatal, most of which are collisions with moving vehicles; (3) The accidents on holidays are often injury accidents. In addition, more accidents tend to occur on workdays than on holidays, due to the high travel frequency on these days.

(2) Association rules between attributes under external environment and those under accident

To disclose how the attributes under external environment affect those under accident, four secondary attributes under external environment (weather, terrain, illumination and visibility) and the secondary attributes under accident (e.g. accident type and accident form) were selected as the targets of association rule mining. A total of 99 effective association rules were obtained through the abovementioned procedure. The ten association rules with the highest confidence (Table 4) were selected for further analysis.

Since the minimum threshold support was set to 0.1, the low-frequency attributes were not included in the frequent itemset, and thus not present in the effective association rules. However, the following laws can be derived from the high-confidence rules: (1) On a sunny day, fatal accidents are likely to occur on plains in the evening without street lighting; most of these accidents are collisions with moving vehicles; (2) During daytime, the collisions with moving vehicles are often injury accidents; (3) On a sunny day, most injury...
TABLE 5. Strong association rules between attributes under road condition and those under accident.

| Rule number | Antecedents => Consequences                                                                 | Support | Confidence | Lift |
|-------------|-------------------------------------------------------------------------------------------|---------|------------|------|
| 1           | (Pavement condition=Dry, Road shape=straight, Intersection/section type=Ramp entrance/exit) => {Accident type=Fatal accident} | 0.12    | 0.69       | 1.29 |
| 2           | (Pavement condition=Dry, Intersection/section type=Ramp entrance/exit) => {Accident type=Fatal accident} | 0.13    | 0.68       | 1.27 |
| 3           | {Highway grade=National highway, Intersection/section type=Ramp entrance/exit} => {Accident type=Fatal accident} | 0.10    | 0.68       | 1.27 |
| 4           | {Road shape=straight, Intersection/section type=Ramp entrance/exit} => {Accident type=Fatal accident} | 0.12    | 0.67       | 1.25 |
| 5           | (Pavement condition=Dry, Highway grade=National highway, Isolation facility=Corrugated beam barrier) => {Accident form=Collision with moving vehicles} | 0.15    | 0.67       | 1.17 |
| 6           | {Intersection/section type=Ramp entrance/exit} => {Accident type=Fatal accident} | 0.14    | 0.66       | 1.25 |
| 7           | {Pavement condition=Dry, Road shape=straight, Highway grade=National highway, Isolation facility=Corrugated beam barrier} => {Accident form=Collision with moving vehicles} | 0.13    | 0.65       | 1.15 |
| 8           | {Accident type=Fatal accident, Pavement condition=Dry, Highway grade=Provincial highway} => {Accident form=Collision with moving vehicles} | 0.10    | 0.65       | 1.14 |
| 9           | {Pavement condition=Dry, solution facility=Corrugated beam barrier} => {Accident form=Collision with moving vehicles} | 0.17    | 0.65       | 1.14 |
| 10          | {Highway grade=National highway, Isolation facility=Corrugated beam barrier} => {Accident form=Collision with moving vehicles} | 0.16    | 0.65       | 1.14 |

Accidents occurring on plains with high visibility are resulted from collisions with moving vehicles.

(3) Association rules between attributes under road condition and those under accident

To disclose how the attributes under road condition affect those under accident, four secondary attributes under road condition (pavement condition, road shape, intersection/section type and isolation facility) and the secondary attributes under accident (e.g. accident type and accident form) were selected as the targets of association rule mining. A total of 122 effective association rules were obtained through the abovementioned procedure. The ten association rules with the highest confidence (Table 5) were selected for further analysis.

The following laws can be derived from these strong association rules: (1) Fatal accidents are likely to occur at ramp entrance/exit, if the pavement is dry and the road is straight; (2) Collisions into moving vehicles often take place on a dry road section with corrugated beam barrier; (3) Fatal accidents are likely to occur on a dry road section, due to collisions into moving vehicles; (4) Injury accidents are likely to occur on a dry and straight road sections with corrugated beam barrier, due to collisions into moving vehicles.

IV. CONCLUSION

The cause analysis of traffic accidents helps the traffic department to identify the key influencing factors of accidents, and disclose the correlations between these factors, laying the basis for formulation of risk control measures. Based on association rule mining, this paper proposes a cause analysis strategy for traffic accidents on urban roads. Considering the multiple layers and dimensions of the data on such accidents, the authors improved the flow of the apriori algorithm, and applied the improved algorithm to mine the association rules between risk factors of traffic accidents on urban roads. According to the layer and dimension of specific attributes, the parameters like support, confidence and lift were adjusted to find the qualified association rules between risk factors. The results were further screened to obtain a series of meaningful association rules. In addition, it should be pointed out that although this research conducts an example analysis on the sample data of urban road traffic accidents, the improved algorithm proposed in this study can be extended to practical application problems with multiple layers and dimensions sample data characteristics.

REFERENCES

[1] A. Papanikolaou and E. Eliopoulou, “Impact of ship age on tanker accidents,” in Proc. Greek Sect. Soc. Nav. Architects Mar. Eng., Athens, Greece, 2008, pp. 103–110.
[2] O. Agbonkhese, G. L. Yisa, E. G. Agbonkhese, D. O. Akanbi, E. O. Aka, and E. B. Mondigha, “Road traffic accidents in Nigeria: Causes and preventive measures,” Civil Environ. Res., vol. 3, no. 13, pp. 90–99, 2013.
[3] L. Larsen and P. Kines, “Multidisciplinary in-depth investigations of head-on and left-turn road collisions,” Accident Anal. Prevention, vol. 34, no. 3, pp. 367–380, May 2002.
[4] J. G. McGwin and D. B. Brown, “Characteristics of traffic crashes among young, middle-aged, and older drivers,” Accident Anal. Prevention, vol. 31, no. 3, pp. 181–198, May 1999.
[5] M. F. Ballesteros and P. C. Dischinger, “Characteristics of traffic crashes in Maryland (1996–1998): Differences among the youngest drivers,” Accident Anal. Prevention, vol. 34, no. 3, pp. 279–284, May 2002.
[6] M. A. Abdel-Aty and A. E. Radwan, “Modeling traffic accident occurrence and involvement,” Accident Anal. Prevention, vol. 32, no. 5, pp. 633–642, Sep. 2000.
[7] C. Chen, G. Zhang, X. C. Liu, Y. C. Huang, J. Ma, Y. Chen, and H. Guan, “Driver injury severity outcome analysis in rural interstate highway crashes: A two-level Bayesian logistic regression interpretation,” Accident Anal. Prevention, vol. 97, pp. 69–78, Dec. 2016.
[8] J. Hinde and C. G. B. Demétrio, “Overdispersion: Models and estimation,” Comput. Statist. Data Anal., vol. 27, no. 2, pp. 151–170, Apr. 1998.
[9] S. Tuncer and A. Alkan, “Spinal cord based kidney segmentation using connected component labeling and K-Means clustering algorithm,” Traitement Signal, vol. 36, no. 6, pp. 521–527, Dec. 2019.
[10] S. G. Langhnoja, M. P. Barot, and D. B. Mehta, “Web usage mining using association rule mining on clustered data for pattern discovery,” Int. J. Data Mining Techn. Appl., vol. 2, no. 1, pp. 141–150, 2013.
[11] B. Tianlong and H. Wentao, “Traffic accident prediction based on time series linear mode,” in Proc. IEEE Conf. Anthology, China, Jan. 2013, pp. 1–3.
[12] T. S. Shively, K. Kockelman, and P. Damien, “A Bayesian semi-parametric model to estimate relationships between crash counts and roadway characteristics,” Transp. Res. B, Methodol., vol. 44, no. 5, pp. 699–715, Jun. 2010.
[13] R. P. Cogdill and P. Dardenne, “Least-squares support vector machines for chemometrics: An introduction and evaluation,” J. Near Infr. Spectrosc., vol. 12, no. 2, pp. 93–100, Apr. 2004.
[14] L.-Y. Chang, “Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network,” Saf. Sci., vol. 43, no. 8, pp. 541–557, Oct. 2005.
[15] C. Riviere, P. Lauret, J. F. M. Ramsamy, and Y. Page, “A Bayesian neural network approach to estimating the energy equivalent speed,” Accident Anal. Prevention, vol. 38, no. 2, pp. 248–259, Mar. 2006.
[16] S. Chinnam, V. Sistla, and V. Kolli, “SVM-PUK kernel based MRI-brain tumor identification using texture and Gabor wavelets,” Traitement Signal, vol. 36, no. 2, pp. 185–191, Jul. 2019.
[17] K. Haleem, M. Abdel-Aty, and K. Mackie, “Using a reliability process to reduce uncertainty in predicting crashes at unsignalized intersections,” Accident Anal. Prevention, vol. 42, no. 2, pp. 654–666, Mar. 2010.
[18] D. Sun, Y. Jia, L. Qin, Y. Yang, and J. Zhang, “A variance maximization based weight optimization method for railway transportation safety performance measurement,” Sustainability, vol. 10, no. 8, p. 2903, 2018.
[19] M. Abdel-Aty, N. Uddin, A. Pande, M. F. Abdalla, and L. Hsia, “Predicting freeway crashes from loop detector data by matched case-control logistic regression,” Transp. Res. Rec., J. Transp. Res. Board, vol. 1897, no. 1, pp. 88–95, Jan. 2004.

QIURU CAI was born in Qinhuangdao, China, in September 1972. She received the B.S. degree in computer engineering from Northeastern University, Shenyang, China, in 1996, and the M.S. degree in computer application from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2008. She is currently a Lecturer with the School of Computer Engineering of Jiangsu University of Technology, China. Her current research interests include data mining, intelligent information systems, and information security.

* * *

VOLUME 8, 2020