sustainability

MDPI

Article

Identifying Factors that Influence the Patterns of Road Crashes Using Association Rules: A case Study from Wisconsin, United States

Shuai Yu *, Yuanhua Jia and Dongye Sun

School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China; yhjia@bjtu.edu.cn (Y.J.); 14114218@bjtu.edu.cn (D.S.)

* Correspondence: 14114217@bjtu.edu.cn; Tel.: +86-152-1057-6646

Received: 14 February 2019; Accepted: 25 March 2019; Published: 1 April 2019

Abstract: Road traffic injury is currently the leading cause of death among children and young adults aged 5–29 years all over the world. Measures must be taken to avoid accidents and promote the sustainability of road safety. The current study aimed to identify risk factors that are significantly associated with the severity in crash accidents; therefore, traffic crashes could be reduced, and the sustainable safety level of roadways could be improved. The Apriori algorithm is carried out to mine the significant association rules between the severity of the crash accidents and the factors influencing the occurrence of crash accidents. Compared to previous studies, the current study included the variables more comprehensively, including environment, management, and the state of drivers and vehicles. The data for the current study comes from the Wisconsin Transportation crash database that contains information on all reported crashes in Wisconsin in the year 2016. The results indicate that male drivers aged 16–29 are more inclined to be involved in crashes on roadways with no physical separation. Additionally, fatal crashes are more likely to occur in towns while property damage crashes are more likely to occur in the city. The findings can help government to make efficient policies on road safety improvement.

Keywords: traffic safety; significant factor; association rules; Apriori

1. Introduction

The number of road traffic deaths in the world remains unacceptably high and increases continuously, reaching 1.35 million in 2016 [1]. However, the fact is, every one of those deaths and injuries is avertible. Improving traffic safety levels is one of the great opportunities to save lives around the world, which does not receive anywhere near the attention it deserves [2].

Traffic crashes can be decreased significantly and identifying the causes of a traffic crash is the most critical procedure in adopting precautionary measures to reduce the severity and quantity of traffic crashes. However, some previous studies estimated a model of crash frequency and severity using only the volume of traffic as an explanatory variable, while clearly many other factors affect the frequency and severity of crashes, such as environmental conditions, roadway geometrics, driver characteristics, and so on. Due to the complex nature of traffic crashes, the policy decision makers must consider numerous contributory factors when making decisions on the improvement of safety [3]. It is vital for decision makers to find the most significant factors that affect the occurrence and consequence of traffic crashes. After years of research, it is generally accepted that through recognizing risk factors as shown in Figure 1, which affect the severity of a crash and corresponding coping strategies, the impact of crashes can be significantly reduced [4–6].

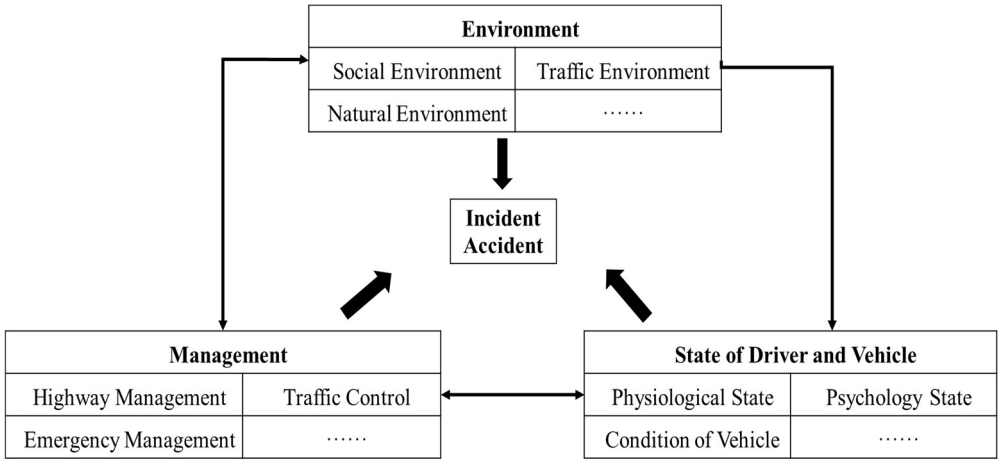

Figure 1. The causative mechanisms of traffic incidents/accidents.

Some previous studies have been devoted to identifying the contributing factors that affect the occurrence and severity of traffic crashes through traffic data. Various approaches were proposed by these studies such as binary logit/probit models [7,8], multinomial logit models [9,10], nested logit models [11,12], log-linear models [13], artificial neural networks [14,15], spatial and temporal correlations [16], Markov switching models [17], and genetic algorithms [18], etc. Meanwhile, various contributing factors to frequency and severity of traffic crashes have been identified in the above literature, such as weather, gender and age of drivers, posted speed, roadway geometrics, condition of drivers, and so on.

In recent years, the analysis of the various types of data using data mining techniques has been attracting more and more attention among researchers. Data mining technology has been employed in traffic crash analysis and achieved satisfactory results in areas such as assessing the inherent connection between crashes and road geometry [19–21], critical points identification [22], factors that contribute to the severity of traffic crashes [23], and the relationship between driver characteristics and traffic crashes [24]. Many studies have analyzed crash data with data mining techniques. Agrawal et al. utilized the data mining technique of association analysis for crash data analysis [25]. Golob and Recker used clustering analysis for relating prevailing traffic conditions on freeways with type of collision most likely to occur [26]. Prati et al. applied a decision tree technique and Bayesian network to predict the severity of bicycle crashes [27]. However, some of these studies are based on the hypotheses that these factors are independent of one another, which might misunderstand the contribution of every single factor.

Among these data mining techniques, association rules mining is a valid technique to analyze traffic crashes since data mining methods do not rely on any hypothesis and can discover meaningful connections hidden in large datasets. There are three kinds of basic algorithms for association rules mining, which are the Apriori algorithm, an algorithm based on partition, and the Frequent Pattern tree algorithm. The Apriori algorithm is succinct and clear, which adopts an iterative method of layer-by-layer search. Compared to the other two algorithms, the Apriori algorithm is more capable of processing large-scale datasets. In the current study, the Apriori algorithm was used to discover the significant rules between the factors and crashes in Wisconsin.

2. Data Description and Processing

2.1. Raw Data and Study Area

The raw crash data for the current study was collected from the Wisconsin Transportation crash database that contains information about all reported crashes in Wisconsin in 2016. A reportable crash was a crash leading to injury or death of any person, total damage to property owned by any one

person to an apparent extent of $1000 or more, or any damage to government-owned non-vehicle property to an apparent extent of $200 or more.

The crash data included 129,051 crashes that occurred in Wisconsin and were described by 49 variables including calendar date on which the crash occurred, crash severity, type of crash, age of the driver, etc. However, not all the reported crashes listed in the database are described by all the 49 variables, and not all the variables were necessarily significant for the crashes. Therefore, in the current study the dataset needs to be pretreated with the following process as shown in Figure 2.

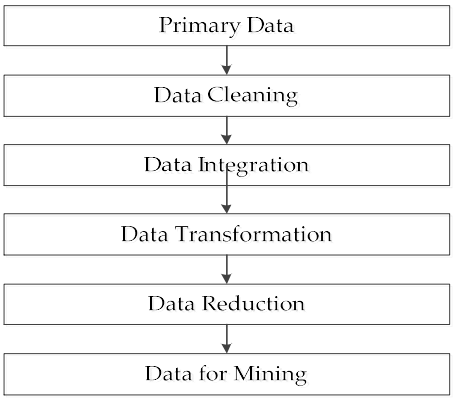

Figure 2. The procedure of data pretreatment.

2.2. Crash Data Processing

First, a clustering algorithm of k-means was used to clean the noise data, which were erroneous or abnormal [28]. Meanwhile, each reported crash needed to be checked for missing values. A reported crash would have to be removed if it had noise data or lacked key information, such as reasons of crash, the condition of the road, weather condition, injury condition, driver information, etc.

Because the data for the current study came from crash and spot investigations with combing meticulously, variables in the dataset were independent and the problem of data conflict does not exist. There was no need to clean up the redundant data and integrate the data. In order to mine association rules more efficiently, variables such as calendar date on which the crash occurred, the name of the street, name of the highway, house, fire, railroad, or other numbers that contributed little to the traffic crash were removed.

Some variables that had the same range of value such as NTFYHOUR (the one-hour range in which the enforcement agency was notified of the crash) and POSTSPD (posted speed) were converted into a different range of value as shown in Table 1. Boolean variables or discrete numeric variables were required to mine association rules using the Apriori algorithm, so that the continuous numerical variable AGE needed to be dispersed as shown in Table 2. Since the residents can get a driver's license at the age of 16 in the United States, the age value of the first group was set by (0,15).

Table 1. Variable conversion.

Variable	NTFYHOUR	POSTSPD	ROADCOND	VEHDMG	SAFETY
Initial data	X (hour) e.g., 5	XX (mile/h) e.g., 55	SNOW	NONE	NONE
Converted data	HX (hour) e.g., H5	SXX (mile/h) e.g., S55	SNOWY	VNONE	SNONE

Table 2. Variable discretization.

Initial Age	(0,15]	[16,25]	[26,35]	[36,45]	[46,55]	[56,65]	[66,75]	[76,85]	[86,99]
Discretization	A1	A2	A3	A4	A5	A6	A7	A8	A9

2.3. Structured Dataset Construction

Twenty-one variables and 63,325 reported crashes were filtered from 129,051 reported crashes by data processing. The description and range of value of the twenty-one variables are cataloged in Table 3.

Table 3. Description and information field of corresponding variables.

NO.	Variables	Description	Information Fields	Percentage (%)
1	MUNITYPE	The municipality type	• C = City	57.9
			• T = Town	29.6
			• V = Village	12.5
2	INTDIS	Intersection Distance in hundredths of a mile from intersection location listed (1 = approx. 50 feet)	• 0	40.8
			• [0,288]	59.2
3	MNRCOLL	Manner (first harmful event) in which participants collided in the crash	• ANGL = Angle	23.6
			• HEAD = Head on collision	1.5
			• NO C = No collision with another vehicle	31.1
			• REAR = Rear end	30.0
			• RTR = Rear to rear	0.3
			• SSO = Sideswipe/opposite direction	2.8
			• SSS = Sideswipe/same direction	10.7
			• GORE = Gore	0.2
			• LTSH = Outside shoulder-left	4.8
			• MED = Median	2.0
4	RLTNRDWY	Location of first harmful event in relation to a roadway	• OFF = Off roadway—location unknown	0.8
			• ON = On roadway	78.8
			• PLOT = Private lot or private prop	0.0
			• RAMP = On ramp	0.7
			• RTSH = Outside shoulder-right	9.1
			• SHLD = Shoulder	3.6
5	HWYCLASS	The type of road the crash took place on	• R CITY = City street rural	3.5
			• R CTH = County trunk rural	8.4
			• R IH = Interstate highway rural	3.2
			• R STH = State highway rural	14.8
			• R TOWN = Town road rural	6.8
			• U CITY = City street urban	40.3
			• U CTH = County trunk urban	0.1
			• U IH = Interstate highway urban	5.1
			• U STH = State highway urban	17.7
6	ACCDSVR	The worst level of the crash severity to life and property	• FAT = Fatal accident	0.5
			• INJ = Injury occurred	31.2
			• PD = Property damage only	68.3
7	POSTSPD	Posted speed for a vehicle unit at the location where a crash occurred	• [S5; S10; S15; S20] mile/hour	1.0
			• S25 mile/hour	26.5
			• [S30; S35; S40; S45; S50] mile/hour	43.3
			• S55 mile/hour	19.2
			• [S60; S65; S70; S77] mile/hour	10.0
8	TRFCWAY	Text describing areas designed for motor vehicle operation	• ND = Not physically divided	60.7
			• D/WO = Divided highway without traffic barrier	21.5
			• D/B = Divided highway with traffic barrier	13.6
			• OW = One-way traffic	4.2
9	AGE	The age of the driver who causes the crash	• A1	0.3
			• A2	35.3
			• [A3, A9]	64.7
10	SEX	The sex of the driver	• Male	57.4
			• Female	42.6
11	VEHDMG	The extent of the worst vehicle damage	• V MNR = Very minor	7.4
			• MNR = Minor	19.1
			• MOD = Moderate	38.6
			• SVR = Severe	22.7
			• V SVR = Very severe	8.3
			• VNONE = None	3.8
12	ROADCOND	Surface condition of the road	• DRY	67.9
			• MUD	0.2
			• SNOWY	14.0
			• ICE	3.1
			• WET	14.8

Table 3. *Cont.*

NO.	Variables	Description	Information Fields	Percentage (%)
13	WTHRCOND	The weather condition at the time of a crash	• CLR = Clear	49.1
			• CLDY = Cloudy	31.5
			• RAIN = Rain	7.7
			• SNOW = Snow	10.0
			• FOG = Fog/smog/smoke	0.5
			• SLET = Sleet/hail	0.7
			• WIND = Blowing sand/dirt/snow	0.4
			• XWIND = Severe crosswinds	0.0
14	DRVRDO	What the driver of unit was doing at the time of the crash	• BACKING = Backing up	3.4
			• CHG LN = Changing lanes	3,7
			• GO STR = Going straight	55.4
			• IL PRK = Illegally parked	0.0
			• LG PRK = Legally parked	0.0
			• LT TRN = Making left turn	13.4
			• MERGING = Merging into traffic	1.4
			• NEGCRV = Negotiating curve	7.0
			• NPASZN = Violate no pass zone	0.1
			• OVT LT = Overtaking on the left	0.7
			• OVT RT = Overtaking on right	0.4
			• PARKNG = Parking maneuver	0.3
			• RT TRN = Right turn	5.9
			• RTOR = Right turn on red	0.0
			• SL/ST = Slowing or stopped	7.3
			• STOPED = Stopped in traffic	0.3
			• UTURN = U turn	0.7
15	DRVRPC	The possible driver contributing circumstances (Driver Factors) in a collision	• DC = Driver condition	2.2
			• DIS = Physically disabled	0.0
			• DTC = Disregard traffic control	3.4
			• FTC = Following too close	11.1
			• FTY = Failure to yield	20.8
			• FVC = Failure to keep vehicle under control	13.6
			• IC = In conflict	0.0
			• ID = Inattentive driving	24.2
			• IO = Improper overtake	1.4
			• IT = Improper turn	2.5
			• LOC = Left of center	1.1
			• SPD = Exceed speed limit	2.6
			• TFC = Too fast for conditions	14.5
			• UB = Unsafe backing	2.6

3. Methodology

3.1. Basic Conceptions

In the current study, the item set is a set of items and it includes at least one reported crash. An item is one element of an item set, which represents a reported crash. A k-item set is defined as an item set consisting of k items. A frequent pattern means that the same combination of eigenvalues occurs a certain number of times in the dataset [29]. The association pattern represents the association and correlation between several items. Association rules are association patterns that satisfy user-specified support [30].

Given a finite set of items $I = \{i_1, i_2 \ldots\ldots, i_m\}$. Let D be a dataset including plenty of transactions that are subsets of I [31]. An extracted association rule is an implication of the form $X \Rightarrow Y$, where X is the antecedent, and Y is the consequent. X and Y are item sets, which belong to D, and $A \cap B = \emptyset$. Support and confidence are the two most commonly used criteria for measuring the importance of association rules. The support indicates the frequency of the association rule in the transaction set containing X and Y, which is defined as $Sup\ (X \Rightarrow Y) = P\ (X \cap Y)$:

$$Sup(X \Rightarrow Y) = \frac{|X \cup Y|}{|D|} \quad (1)$$

|D| is the total number of transactions, while $|X \cup Y|$ is the number of transactions that include both item sets X and Y.

The confidence indicates the credibility of the association rule $X \Rightarrow Y$, which is defined as $Conf(X \Rightarrow Y)$:

$$Conf(A \Rightarrow B) = \frac{|X \cup Y|}{|X|} = \frac{Sup(X \cup Y)}{Sup(X)} \quad (2)$$

|X| is the number of transactions that only contain item set X, while $|X \cup Y|$ is the number of transactions that include both item sets X and Y. The association rules whose value of support and confidence are equal to or bigger than the threshold defined by users are valid rules, which deserve to be analyzed.

To avoid generating a great number of uninteresting association rules, many algorithms for mining association rules use criteria based on minimum support and minimum confidence. Due to lacking consideration of correlation between the support of X and the support of (X, Y), useless association rules may still be generated when the support value of the consequent is too high. In order to solve this problem, previous researchers have proposed several valid measures. Lift is the most widely used measure of them, which is defined as

$$\text{lift}(X \Rightarrow Y) = \frac{Conf(X \Rightarrow Y)}{Sup(Y)} \quad (3)$$

Conf(X⇒Y) is the confidence of association rule $(X \Rightarrow Y)$, while *Sup(Y)* is the support value of item set Y. There is no correlation between item set X and Y with lift = 1, while the occurrence of item set X is exclusive to item set Y with lift < 1. Only if lift > 1, the association rules are recognized as valuable rules.

3.2. Association Rule Mining

Extracting important and hidden information from a large dataset by mining association rules is one of the most common tasks in data mining [32]. The association rule mining can be described as a two-step process [33]:

- Generating frequent item sets—find all frequent item sets whose support value is equal to or greater than the minimum support value;
- Generating association rules—generate association rules from frequent item sets under the condition of minimum confidence.

Figure 3 shows the process of association rule mining.

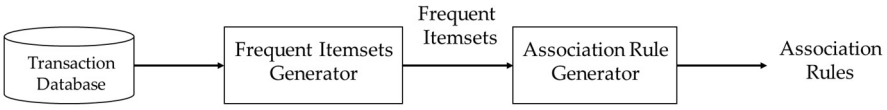

Figure 3. Association rule mining process.

The association rules mining algorithms include Apriori, SETM [34], ECLAT [35], Pincer Search [36], and MAFIA [37], which are based on a support-confidence framework proposed by Agrawal and Srikant. The Apriori algorithm is succinct and clear, which adopts an iterative method of layer-by-layer search. In the current study, the Apriori algorithm was used to discover the significant rules between the factors and crashes in Wisconsin.

3.3. Validity Test of Association Rules

An extreme risk of type-I error exists because of the large number of association rules, which needs a process of validity tests to evaluate the statistical significance of the rules obtained [38].

The validation process is generally distinguished in two ways. The first approach is the direct adjustment approach, which requires all association rules to pass statistical tests at the adjusted critical value. The second approach is the holdout approach, which divides the data into exploratory data for generating association rules without regard for the problem of multiple testing and holdout data for statistical tests.

In the current study, a direct adjustment approach was applied to test the validation of association rules, as it has an advantage of data usage for both association rule discovery and statistical evaluation [38]. Meanwhile, no more statistical tests will be required under this approach than under the holdout approach. A number of direct adjustment approaches were employed to perform multiple hypothesis tests, such as Bonferroni correction [39], sequentially rejective Bonferroni [40], adaptive Benjamini–Hochberg algorithm [41], and so on. The Bonferroni correction states that if an experimenter is testing n independent hypotheses on a set of data, then the statistical significance level that should be used for each hypothesis separately is 1/n times what it would be if only one hypothesis was tested. Because of the principle and characteristics of Bonferroni correction, it made the results more rigorous with a tight upper bound. Thus, the method of Bonferroni correction was applied in the current study. The definition of Bonferroni correction is as follows:

Let H_1, H_2,..., H_n be a family of hypotheses and p_1, p_2, ... , p_n be their corresponding p-values. The n is the total number of null hypotheses, while n_0 is the number of true hypotheses. The familywise error rate (FWER) is the probability of rejecting at least one true H_i; in other words, of making at least one type I error. The Bonferroni correction rejects the null hypothesis for each $p_i \leq \alpha/n$, while α is the global significance level. Proof of this control follows from Boole's inequality, as follows:

$$FWER = P\left\{\bigcup_{i=1}^{n_0}\left(p_i \leq \frac{\alpha}{n}\right)\right\} \leq \sum_{i=1}^{n_0}\left\{P\left(p_i \leq \frac{\alpha}{n}\right)\right\} = n_0\frac{\alpha}{n} \leq n\frac{\alpha}{n} = \alpha \quad (4)$$

4. Results and Discussions

Through the procedure of data pretreatment, 63,325 pieces of valid reported crashes data were filtrated. Among them, there were 43,239 pieces of property damage only (PD) crashes, 19,766 injuries occurred (INJ) crashes, and 320 fatal crashes (FAT) as in Figure 4. Based on the dataset, the current study then used the mathematical programming software Python 3.5 on a Lenovo laptop with Intel Core i5-5200U 2.20GHz CPU and 8 GB RAM to generate association rules. There were 766 pieces of association rules that were obtained with filter criteria of minimum support equal to 0.1, minimum confidence equal to 0.14, and minimum lift greater than 1.0, as shown in Figure 5.

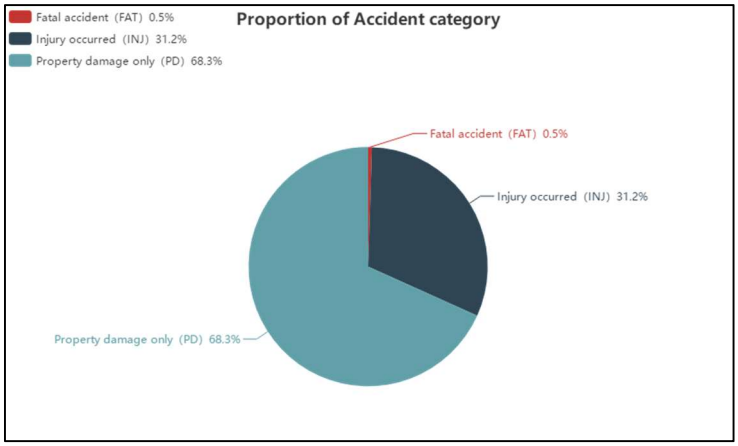

Figure 4. The proportion of accident category.

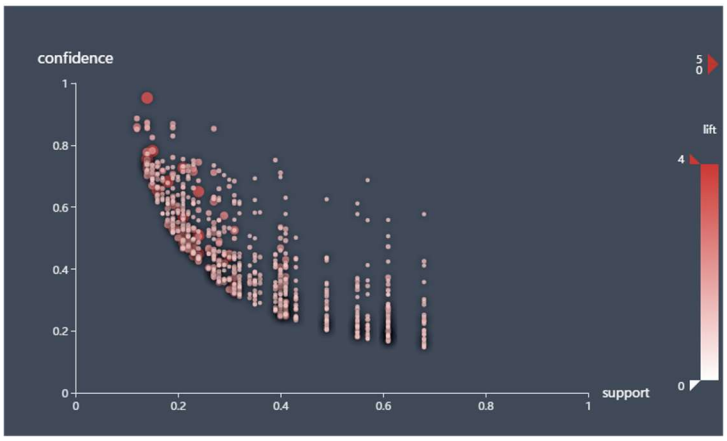

Figure 5. Seven hundred and sixty-six pieces of association rules.

The current study estimated the smallest p-value for the association rules based on the upper bound of 0.1/766 that equals $1.3*10^{-4}$, while 766 pieces of association were obtained with a minimum support value that equals 0.1 [42]. Only two rules had p-values higher than $1.3*10^{-4}$—the p-value of rule WET, MALE ⇒ ND is 0.012 and the p-value of rule LT TRN ⇒ PD is 0.029. The reason for the extremely low number of false discoveries is that the support, confidence, and lift threshold already do an excellent job of pruning out most rules that are not statistically significant.

High support rules indicate a high frequency of association rules (i.e., events that occur frequently in a crash), while high confidence indicates the probability of occurrence of a consequent event when the antecedent item occurred (i.e., the antecedent event is more likely to occur when the antecedent event happens in a crash). Rules with high lift value, which are greater than 1.0, are valid rules and indicate strong associations between the factors (i.e., there is a strong positive correlation between the two events in a crash). The current study screened out the top 20 support association rules of the highest value as in Table 4, the top 20 confidence association rules of the highest value as in Table 5, and the top 20 lift association rules of the highest value as shown in Table 6.

Table 4. Top 20 support association rules of the highest value.

Rules	Antecedent	Consequent	Support	Confidence	Lift
1	PD	S25, 0	0.68	0.15	1.01
2	PD	S25, CLR	0.68	0.15	1.05
3	PD	U CITY, CLR, ND	0.68	0.15	1.01
4	PD	S25, M	0.68	0.16	1.08
5	PD	MOD, A2	0.68	0.16	1.13
6	PD	U CITY, ND, M	0.68	0.16	1.04
7	PD	TFC	0.68	0.16	1.08
8	PD	REAR, M	0.68	0.17	1.02
9	PD	0, MOD	0.68	0.17	1.05
10	PD	S25, U CITY, ND	0.68	0.18	1.07
11	PD	MOD, U CITY	0.68	0.18	1.09
12	PD	F, U CITY	0.68	0.19	1.01
13	PD	MOD, F	0.68	0.19	1.10
14	PD	MOD, CLR	0.68	0.20	1.06
15	PD	S25, U CITY	0.68	0.20	1.07
16	PD	A2, M	0.68	0.20	1.02
17	PD	U CITY, M	0.68	0.22	1.03
18	PD	MOD, GO STR	0.68	0.23	1.07
19	PD	MNR	0.68	0.23	1.21
20	PD	MOD, M	0.68	0.24	1.11

Following are the analysis of results from Table 4:

- Due to the PD (property damage only) crashes having a proportion of 68.3% in the whole dataset, the top 20 support association rules of highest value are all related to PD. It indicates that most of the crashes are not related to injury and fatalities, which is consistent with the findings of the Global status report on road safety 2018 [1].
- The significant factors for the high value of support association rules are the type of road, the extent of the worst vehicle damage, posted speed, male drivers, and a roadway with no physical separation, weather, location, and age of drivers.
- It is obvious that the extent of vehicle damage is more likely to be moderate (MOD) in a property damage only crash (rule 5, 9, 11, 13, 14, 18, and 20).
- The crashes mostly occurred in urban areas (rule 11, 12, and 17) with no physical separation (rule 3 and 6), while Abdel-Aty and Radwan found that highway geometry is the second important factor in occurrence of traffic crashes [24], and a lower posted speed (rule 15). Especially, the rule *PD → S25, U CITY, ND (support = 0.68, confidence = 0.18, lift = 1.07)* clearly expresses the relationship between them. Through the revelation of the above rules, decision makers can reduce the occurrence of crashes by setting up physical separations on crash-prone sections.
- Male drivers are more prone to be associated with property damage only traffic crashes than female drivers, which can be observed from the rules (4, 6, 8, 16, 17, and 20) and rules (12 and 13). On the one hand, male drives are more likely to drive drunk and/or speed than female drivers [43]. On the other hand, it is probable that male drivers are less likely to comply with traffic rules and are generally overconfident while driving [44].

Table 5. Top 20 confidence association rules of the highest value.

Rules	Antecedent	Consequent	Support	Confidence	Lift
1	FTC	REAR	0.14	0.95	3.17
2	S55, NO C	ND	0.12	0.89	1.46
3	S25, GO STR	ND	0.14	0.87	1.44
4	S25, U CITY, PD	ND	0.14	0.87	1.43
5	S25, U CITY	ND	0.19	0.87	1.43
6	S25, CLR	ND	0.14	0.86	1.42
7	ANGL, GO STR	0	0.12	0.86	2.10
8	S25, PD	ND	0.19	0.86	1.41
9	S25, M	ND	0.14	0.85	1.41
10	S25	ND	0.27	0.85	1.41
11	S25, F	ND	0.12	0.85	1.40
12	MNR	PD	0.19	0.83	1.21
13	S25, 0	ND	0.15	0.82	1.36
14	FTY, ANGL	0	0.15	0.78	1.92
15	0, FTY	ANGL	0.15	0.78	3.31
16	ND, FVC	NO C	0.14	0.78	2.5
17	MOD, A2	PD	0.14	0.77	1.13
18	S55	ND	0.19	0.76	1.25
19	FTY, ND	ANGL	0.14	0.75	3.20
20	MOD, M	PD	0.21	0.75	1.11

Following are the analysis of results from Table 5:

- The highest confidence value rule *FTC (following too close) → REAR (rear end) (support = 0.14, confidence = 0.95, lift = 3.17)* indicates that following too close will lead to rear ending between cars, which is widely known.
- Same as the result from Table 4, low posted speed and roadways with no physical separation (rule 3, 4, 5, 6, etc.) are significant elements that affect the occurrence of crashes. The large deviation

of speed, which is generated by drivers that ignore the posted speed and speed a lot, is perhaps the reason why crashes happen in the location with low posted speed. Elvik found that lower posted speed is prone to lead to a crash as a result of a high deviation of speed [45]. Roadways with no physical separation often cause the problem that drivers sometime occupy the opposite lanes, which probably leads to a collision.

- In comparison with other drivers, the drivers aged 16–25, which are presented by A2 in Tables 4 and 5, are most likely to be involved in crashes (rule 5, 16 in Table 4, rule 17 in Table 5), because drivers aged 16–25 are a large proportion of the whole drivers, and they are more likely to violate driving rules. Decision makers can strengthen traffic safety education for drivers aged 16–25 to reduce the occurrence of traffic crashes.
- '0' indicates that the crash occurred at an intersection. Four rules (rule 7, 13, 14, and 15) show that crashes are more likely to occur at an intersection. The intersection is a convergence area of city traffic flow and flow of people, which have complex traffic conditions and are more likely to lead to a crash. Wang et al. found that a crash is more prone to occur at an intersection [46]. An appropriate organization of intersection flow might help decision makers control the occurrence of crashes effectively.
- Following too close (FTC), failure to yield (FTY), and failure to keep the vehicle under control (FVC) are perhaps the significant driver-contributing circumstances in a crash (rule 1, 14, 15, 16, and 19). Abdel-Aty and Radwan found that driver conditions were the most important factors in the occurrence of traffic crashes [24].

Table 6. Top 20 lift association rules of the highest value.

Rules	Antecedent	Consequent	Support	Confidence	Lift
1	0, FTY	ANGL	0.15	0.78	3.31
2	FTY	0, ANGL	0.21	0.57	3.23
3	FTY, ND	ANGL	0.14	0.75	3.20
4	FTC	REAR	0.14	0.95	3.17
5	S55	ND, NO C	0.19	0.54	2.52
6	ND, FVC	NO C	0.14	0.78	2.50
7	M, FVC	NO C	0.14	0.73	2.36
8	FVC, GO STR	NO C	0.14	0.72	2.33
9	M, NO C	FVC	0.20	0.53	2.31
10	FVC	NO C	0.23	0.71	2.30
11	PD, FVC	NO C	0.15	0.71	2.28
12	ND, S55	NO C	0.15	0.71	2.27
13	U CITY, PD, ND	S25	0.20	0.59	2.24
14	FVC	ND, NO C	0.23	0.48	2.24
15	FVC	PD, NO C	0.23	0.47	2.21
16	S25	U CITY, ND	0.27	0.62	2.16
17	U CITY, ND	S25, PD	0.29	0.42	2.14
18	ANGL, GO STR	0	0.12	0.86	2.10
19	0, GO STR	ANGL	0.21	0.50	2.08
20	0, U CITY	ANGL	0.22	0.49	2.05

Following are the analysis of results from Table 6:

- High lift values suggest a strong interdependence between the antecedent and the consequent. Three rules with high lift values indicate that drivers failing to yield, crash occurring at the intersection, and the collision type of angle have a strong connection [24].
- The rule with highest lift value is *0, FTY → ANGL (support = 0.15, confidence = 0.78, lift = 3.17)*. The support value shows that 15% of crashes result from failing to yield at an intersection [46]. The confidence value proves that 78% of the crashes occurred due to angle collision. The ratio of angle collision crashed was 3.17 times the ratio of other types of collision.

- The crash is more likely to happen when drivers go straight (rule 8, 18, and 19), because drivers might tend to be more relaxed with their vigilance during going straight than when crossing a curve.
- There are nine rules with *NO C = no collision* as a consequent, which indicates that most of the crashes with no collision happened between vehicles because most of the vehicles had a collision with a physical barrier.
- Male drivers are more prone to fail to keep the vehicle under control. Das et al. also found a higher number of males are associated with crashes [47].

With the percentage of fatal crashes (0.5%) being too small, it is impossible to produce high values of support and confidence. To discuss the influence factors of fatal crashes, the dataset applied only included fatal crashes. Twelve pieces of association rules that were obtained with filter criteria of minimum support that equaled 0.5, minimum confidence that equaled 0.5, and minimum lift that was greater than 1.0 is shown in Table 7.

Table 7. Association rules related to fatal crashes.

Rules	Antecedent	Consequent	Support	Confidence	Lift
1	T	M	0.68	0.76	1.02
2	M	T	0.75	0.69	1.02
3	DRY	CLR	0.84	0.69	1.14
4	CLR	DRY	0.60	0.96	1.14
5	M	ND	0.75	0.77	1.02
6	ND	M	0.76	0.76	1.02
7	M	DRY	0.75	0.87	1.03
8	DRY	M	0.84	0.77	1.03
9	T	ND	0.68	0.86	1.13
10	ND	T	0.76	0.77	1.13
11	V SVR	ND	0.68	0.76	1.02
12	ND	V SVR	0.75	0.69	1.02

The following are the analysis of results from Table 7:

- The significant factors for fatal crashes are location, male drivers, the extent of the worst vehicle damage, roadway with no physical separation, weather and road surface condition.
- Different from property damage only crashes, fatal crashes are more likely to occur in town instead of the city. Compared with the city road, there are fewer vehicles, police, and less supervision in town. Drivers tend to be more relaxed with their vigilance and speeding.
- Male drivers are prone to be involved in fatal crashes, which has the same reason with other types of crashes.
- Drivers are more likely to get involved in fatal crashes when the weather condition is clear, and the road surface condition is dry. It is perhaps because drivers would pay more attention to driving when the weather and road surface condition are dangerous. Karlaftis and Yannis suggest a negative relationship between adverse weather and road safety, mainly because drivers are not used to driving under adverse weather conditions and consequently adjust their behavior by driving more carefully [48].
- Roadways with no physical separation have always been a problem threatening traffic safety.

5. Conclusions

Due to the complicated interaction among different factors—the situation of the driver, the condition of vehicle and road, environment and management—a traffic crash is a complex and systemic problem. In order to decrease the number of traffic crashes, fundamental reasons, which are the basis for promoting measures, need to be systematically analyzed. A large number of researchers

have made efforts to identify the vital factors that influence the severity and frequency of traffic crashes during recent years, in order to formulate effective safety countermeasures to enhance traffic sustainability [47].

In the current study, the Apriori algorithm was implemented to identify characteristics and factors impacting traffic crashes in Wisconsin, United States. By setting an appropriate threshold value of support and confidence, essential information of traffic crash characteristics can be gained to analyze the fundamental causes of a traffic crash. The association rules, which were generated in the current study, suggest a couple of significant factor groups: posted speed, driver condition, weather condition, road surface condition, distance from the intersection, a roadway with no physical separation, an administrative grade of crash location, male drivers, and the age of drivers. Taking these factors into account, the government can make countable measures to improve the sustainable level of traffic safety. The majority of the findings are consistent with previous studies. The variables considered are more comprehensive, including environment, management, and state of drivers and vehicles, which is the critical contribution of the current study.

Note that the present study did not optimize the parameters with any optimization method, for the current study obtained objective and significant results in the current size of the database. For future directions, efforts could be made on incorporating genetic algorithms and particle swarm optimization with the Apriori algorithm to optimize the values of the parameters, and to obtain significant results with high efficiency in analyzing large-scale databases.

Author Contributions: S.Y. and Y.J. developed the concept and designed the study. S.Y. performed the methodology. S.Y. and D.S. performed the data analysis. Y.J. and D.S. read and approved the final manuscript. All authors contributed to the result interpretation and the final version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: This research is supported by National Natural Science Foundation of China (71340020).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. World Health Organization (WHO). *Global Status Report on Road Safety 2018*; WHO: Geneva, Switzerland, 2018.
2. Wegman, F. *Road Accidents: Worldwide a Problem that Can Be Tackled Successfully*; Permanent International Association of Road Congresses: Paris, France, 1996.
3. Tešić, M.; Hermans, E.; Lipvac, K.; Pešić, D. Identifying the most significant indicators of the total road safety performance index. *Accid. Anal. Prev.* **2018**, *113*, 263–278. [CrossRef] [PubMed]
4. Haddon, W. Options for the prevention of motor vehicle crash injury. *Isr. J. Med Sci.* **1980**, *16*, 45–65. [PubMed]
5. Sun, Y.; Hrušovský, M.; Zhang, C.; Lang, M.X. A Time-Dependent Fuzzy Programming Approach for the Green Multimodal Routing Problem with Rail Service Capacity Uncertainty and Road Traffic Congestion. *Complexity* **2018**, *2018*, 1–22. [CrossRef]
6. Figueira, A.D.C.; Pitombo, C.S.; Oliveira, D.; Paulo, T.M.S.; Larocca, A.P.C. Identification of rules induced through decision tree algorithm for detection of traffic accidents with victims: A study case from Brazil. *Case Stud. Transp. Policy* **2017**, *5*, 200–207. [CrossRef]
7. Shibata, A.; Fukuda, K. Risk factors of fatality in motor vehicle traffic accidents. *Accid. Anal. Prev.* **1994**, *26*, 391–397. [CrossRef]
8. Moudon, A.; Lin, L.; Jiao, J.; Hurvitz, P.; Reeves, P. The risk of pedestrian injury and fatality in collisions with motor vehicles, a social ecological study of state routes and city streets in King County, Washington. *Accid. Anal. Prev.* **2011**, *43*, 11–24. [CrossRef] [PubMed]
9. Shankar, V.; Mannering, F. An exploratory multinomial logit analysis of single-vehicle motor cycle accident severity. *J. Saf. Res.* **1996**, *27*, 183–194. [CrossRef]
10. Yasmin, S.; Eluru, N.; Ukkusuri, S. Alternative ordered response frameworks for examining pedestrian injury severity in New York City. *J. Transp. Saf. Secur.* **2014**, *6*, 275–300. [CrossRef]

11. Wu, Z.; Sharma, A.; Mannering, F.; Wang, S. Safety impacts of signal-warning flashers and speed control at high-speed signalized intersections. *Accid. Anal. Prev.* **2013**, *54*, 90–98. [CrossRef]
12. Savolainen, P.; Mannering, F. Probabilistic models of motor cyclists' injury severities in single- and multi-vehicle crashes. *Accid. Anal. Prev.* **2007**, *39*, 955–963. [CrossRef]
13. Chen, W.; Jovanis, P. Method for identifying factors contributing to driver-injury severity in traffic crashes. *Transp. Res. Rec.* **2000**, *1707*, 1–9. [CrossRef]
14. Abdelwahab, H.; Abdel-Aty, M. Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. *Transp. Res. Rec.* **2001**, *1746*, 6–13. [CrossRef]
15. Chimba, D.; Sando, T. Neuromorphic prediction of highway injury severity. *Adv. Transp. Stud.* **2009**, *19*, 17–26.
16. Castro, M.; Paleti, R.; Bhat, C. A spatial generalized ordered response model to examine highway crash injury severity. *Accid. Anal. Prev.* **2013**, *52*, 188–203. [CrossRef]
17. Xiong, Y.; Tobias, J.; Mannering, F. The analysis of vehicle crash injury-severity data: A Markov switching approach with road-segment heterogeneity. *Transp. Res. Part B: Methodol.* **2014**, *67*, 109–128. [CrossRef]
18. Martin, D.; Rosete, A.; Alcala-Fdez, J.; Herrera, F. A new multiobjective evo-lutionary algorithm for mining a reduced set of interesting positive and negative quantitative association rules. *IEEE Trans. Evol. Comput.* **2014**, *18*, 54–69. [CrossRef]
19. Miaou, S.; Lum, H. Modeling vehicle accidents and highway geometric design relationships. *Accid. Anal. Prev.* **1993**, *25*, 689–709. [CrossRef]
20. Shankar, V.; Mannering, F.; Barfield, W. Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. *Accid. Anal. Prev.* **1995**, *27*, 371–389. [CrossRef]
21. Milton, J.; Mannering, F. The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies. *Transportation* **1998**, *25*, 395–431. [CrossRef]
22. Tarko, A.P.; Kanodia, M. Effective and Fair Identification of Hazardous Locations. In *Transportation Research Record: Journal of Transportation Research Board, No. 1897*; Transportation Research Board of National Academics: Washington, DC, USA, 2004; pp. 64–70.
23. Bagdadi, O. Estimation of the severity of safety critical events. *Accid. Anal. Prev.* **2013**, *50*, 167–174. [CrossRef] [PubMed]
24. Abdel-Aty, M.; Radwan, A. Modeling traffic accident occurrence and involvement. *Accid. Anal. Prev.* **2000**, *32*, 633–642. [CrossRef]
25. Agrawal, R.; Imielinski, T.; Swami, A. Mining association rules between sets of items in large databases. *Proc. ACM Sigmod* **1994**, *22*, 207–216. [CrossRef]
26. Golob, T.; Recker, W. A Method for relating type of cash to traffic flow characteristics on urban freeways. *Transp. Res. Part A Policy Pract.* **2004**, *38*, 52–80.
27. Prati, G.; Pietrantoni, L.; Fraboni, F. Using data mining techniques to predict the severity of bicycle crashes. *Accid. Anal. Prev.* **2017**, *101*, 44–54. [CrossRef]
28. Kumar, S.S.; Kumar, J.S. A Study of K-Means and C-Means Clustering Algorithms for Intension Detection Product Development. *Int. J. Innov. Technol. Manag.* **2016**, *5*, 207–213.
29. Rodríguez, G.; Ansel, Y.; Martínez, T.; José, F. Mining frequent patterns and association rules using similarities. *Expert Syst. Appl.* **2013**, *40*, 6823–6836. [CrossRef]
30. Xue, C.J.; Song, W.J.; Qin, L.J.; Dong, Q.; Wen, X.Y. A mutual-information-based mining method for marine abnormal association rules. *Comput. Geosci.* **2015**, *76*, 121–129.
31. Lazzerini, B.; Pistolesi, F. Profiling risk sensibility through association rules. *Expert Syst. Appl.* **2013**, *40*, 1484–1490. [CrossRef]
32. Kabir, M.M.J.; Xu, S.X.; Kang, B.H.; Zhao, Z.Y. A new multiple seeds based genetic algorithm for discovering a set of interesting Boolean association rules. *Expert Syst. Appl.* **2017**, *74*, 55–69. [CrossRef]
33. Agrawal, R.; Srikant, R. Fast algorithms for mining association rules in large databases. *Process. 20th Int. Conf. Very Large Databases* **1994**, *1215*, 487–499.
34. Houtsma, M.; Swami, A. Set-oriented mining for association rules in relational databases. In Proceedings of the Eleventh International Conference on Data Engineering, Taipei, Taiwan, 6–10 March 1995; pp. 25–33.
35. Zaki, M.J. Scalable algorithms for association mining. *IEEE Trans. Knowl. Data Eng.* **2000**, *12*, 372–390. [CrossRef]

36. Lin, D.I.; Kedem, Z.M. Pincer-search: An efficient algorithm for discovering the maximum frequent set. *IEEE Trans. Knowl. Data Eng.* **2002**, *14*, 553–566. [CrossRef]
37. Burdick, D.; Calimlim, M.; Flannick, J.; Gehrke, J.; Yiu, T. MAFIA: A maximal frequent itemset algorithm. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 1490–1504. [CrossRef]
38. Webb, G.I. Discovering significant patterns. *Mach. Learn.* **2007**, *68*, 1–33. [CrossRef]
39. Scheffer, T. Finding association rules that trade support optimally against confidence. *Intell. Data Anal.* **1995**, *9*, 381–395. [CrossRef]
40. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
41. Benjamini, Y.; Hochberg, Y. On the Adaptive Control of the False Discovery Rate in Multiple Testing with Independent Statistic. *J. Educ. Behav. Stat.* **2000**, *25*, 60–83. [CrossRef]
42. Megiddo, N.; Srikant, R. Discovering predictive association rules. In Proceedings of the International Conference on Knowledge Discovery & Data Mining, San Diego, CA, USA, 15–18 August 1999.
43. Shinar, D.; Compton, R. Aggressive driving: An observational study of driver, vehicle, and situational variables. *Accid. Anal. Prev.* **2004**, *36*, 429–437. [CrossRef]
44. Zhang, G.N.; Yau, K.K.W.; Gong, X.P. Traffic violations in Guangdong Province of China: Speeding and drunk driving. *Accid. Anal. Prev.* **2014**, *63*, 30–40. [CrossRef]
45. Elvik, R. A comprehensive and unified framework for analysing the effects on injuries of measures influencing speed. *Accid. Anal. Prev.* **2019**, *125*, 63–69. [CrossRef]
46. Jinghui, Y.; Mohamed, A.A. Approach-level real-time crash risk analysis for signalized intersections. *Accid. Anal. Prev.* **2018**, *119*, 274–289.
47. Das, S.; Dutta, A.; Jalayer, M.; Bibeka, A.; Wu, L. Factors influencing the patterns of wrong-way driving crashes on freeway exit ramps and median crossovers: Exploration using 'Eclat' association rules to promote safety. *Int. J. Transp. Sci. Technol.* **2018**, *7*, 114–123. [CrossRef]
48. Karlaftis, M.; Yannis, G. Weather effects on daily traffic accidents and fatalities: A time series count data approach. In Proceedings of the 89th Annual Meeting of the Transportation Research Board, Washington, DC, USA, 10–14 January 2010.