Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2013, Article ID 302627, 8 pages
http://dx.doi.org/10.1155/2013/302627

Research Article

A Hybrid Algorithm of Traffic Accident Data Mining on Cause Analysis

Jianfeng Xi,[1,2] Zhenhai Gao,[1] Shifeng Niu,[3] Tongqiang Ding,[2] and Guobao Ning[4]

[1] *State Key Laboratory of Automobile Dynamic Simulation, Jilin University, Changchun 130022, China*
[2] *College of Traffic, Jilin University, Changchun 130022, China*
[3] *School of Automobile, Chang'an University, Xi'an 710064, China*
[4] *School of Automotive Engineering, Tongji University, Shanghai 201804, China*

Correspondence should be addressed to Guobao Ning; guobao.ning@tongji.edu.cn

Received 6 October 2012; Revised 30 December 2012; Accepted 31 December 2012

Academic Editor: Baozhen Yao

Road traffic accident databases provide the basis for road traffic accident analysis, the data inside which usually has a radial, multidimensional, and multilayered structure. Traditional data mining algorithms such as association rules, when applied alone, often yield uncertain and unreliable results. An improved association rule algorithm based on Particle Swarm Optimization (PSO) put forward by this paper can be used to analyze the correlation between accident attributes and causes. The new algorithm focuses on characteristics of the hyperstereo structure of road traffic accident data, and the association rules of accident causes can be calculated more accurately and in higher rates. A new concept of Association Entropy is also defined to help compare the importance between different accident attributes. T-test model and Delphi method were deployed to test and verify the accuracy of the improved algorithm, the result of which was a ten times faster speed for random traffic accident data sampling analyses on average. In the paper, the algorithms were tested on a sample database of more than twenty thousand items, each with 56 accident attributes. And the final result proves that the improved algorithm was accurate and stable.

1. Introduction

In recent years, with the growth of the volume and travel speed of road traffic, the number of traffic accidents, especially severe crashes, has been increasing rapidly on a yearly basis. The issue of traffic safety has raised great concerns across the globe, and it has become one of the key issues challenging the sustainable development of modern traffic and transportation. Therefore, it is crucial for engineers to be able to extract useful information from existing data to analyze the causes of traffic accidents, so that traffic administrations can be more accurately informed and better policies can be introduced [1–3].

Traffic conditions are a complex system due to many stochastic factors [4–6], and traffic accident data has long been known to be very difficult to process. Many attempts have been made in recent years through applying different methodologies and algorithms. Association rules has captured wide attentions and careful studies because of its adoptability and the nature of being easily understood, the focus of study of which is how to increase the accuracy and efficiency of the calculation. Among the researches to date, Geurts et al. [7] used association rules to identify accident circumstances that frequently occur together at high frequency accident locations; Tesema et al. [8] developed an adaptive regression trees to build a decision support system to handle road traffic accident analysis; Marukatat [9] has made noticeable attempts at identifying the degree of importance of Information Entropy for road traffic accident analyses. Dong et al. [10], Lee et al. [11], Hassan and Tazaki [12], Zhang et al. [13], and other researchers have achieved multileveled data mining of traffic accidents by means of a comprehensive application of data mining techniques. The researches above all achieved the mining of accident data on a certain level; however, the overall calculating processes are largely too complicated and cannot be applied

to all types of data, especially the multiattribute ones. On the other hand, the PSO algorithm has been applied in many fields. Shi and Eberhart [14] studied the parameters' optimization, based on particle swarm optimizer. Wang et al. [15] propose an association rules algorithm based on particle swarm optimization algorithm to mining the transaction data in the stock market. Moreover, others [16–20] improved and applied PSO algorithm to their purpose. So far, there have been a lot of researches targeting at different types of data, and due to the "capricious" nature of real-world data, coupled with the innate shortcomings of the algorithm, the association rules still falls short of people's expectations in being less complicated, less time and space-consuming, and more efficient.

In this paper, a new method of traffic accident data mining, based on PSO, association rules, and Information Entropy theories and through a comparative analysis of a variety of traffic accident data mining techniques, is put forward to identify the importance of different attributes and their respective values. The method is an attempt to come up with a multidimensional, all-inclusive method of data analysis to simplify existing algorithms as well as apply computational intelligence algorithms such as PSO to road traffic data analyses.

2. Characteristics of Road Traffic Accident Data

Road traffic accident is under the influence of many factors, which makes it a complicated, and as far as information is concerned, an unfinished, uncertain gray system. There are different databases of traffic accident in different countries [8, 21, 22]. At present, roughly 60 items of information are contained in the China "Database of Road Traffic Accident" which is used by Chinese traffic administrative agencies, spawning off approximately 130 items of single-unit information, which can be used to reconstruct the whole process of the accident in a relatively full and objective manner. It provides more than adequate the information and references for road traffic accident analyses.

Road traffic accidents have their innate, random nature but are also subject to other factors. If the connections of those factors could be identified, through manual control and interference, the rates of traffic accidents could be lower.

Traffic accident data is the foundation of traffic accident analysis, the form and structure of which determine the form and structure of the analysis model. From in-depth analysis of the traffic accident database operated by the Ministry of Public Security, the data could be regards as a radial, multidimensional, and multilayered structure, as shown in Figure 1.

The structure of the data determines the structure of the causation-analysis model. This paper designs a double-layered analysis model and provides an improved algorithm according to the hyperstereo structure of the data. The purpose is to analyze the importance of each value on the attribute value layer with the association rules method and to compare the importance of each attribute on the attribute value layer with the Information Entropy method.

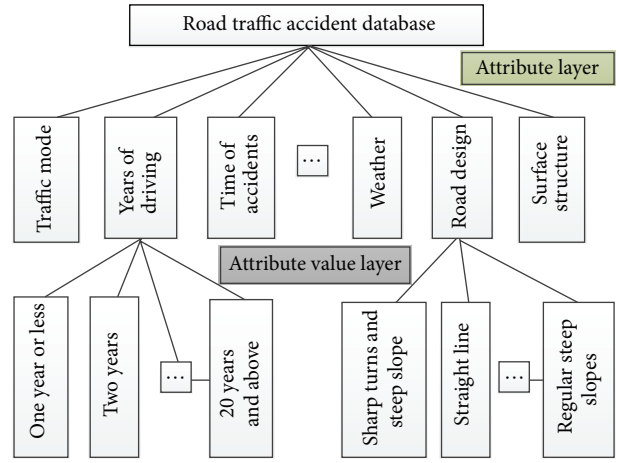

Figure 1: Structure of the traffic accidents data.

3. Characteristics of Data Analysis Algorithms

3.1. Association Rules. Association rules is a data mining method for investigating the associative property of different events, which can be used in traffic accident data mining to mine the importance of attributes, that is, the associative relationship of events with certain types of accident. Its basic idea is to treat each characteristic as an item. Accident site, number of death, and so on can all be called an item. The higher the association, the more likely one event is directly linked to the cause of a certain type of accident.

To decide how related two items are, we need to identify how many times some characteristics appear at the same time in a large number of similar events. If items show up at the same time frequently, indicating that there is a statistic pattern behind it, we can start to believe that the items are relevant.

In the association rules algorithm, X and Y are two random events, which can be thought as relevant and also can be thought as seemingly irrelevant. Assume there is a causal relationship between X and Y, which means when X happens, Y also happens (and vice versa). "$X \rightarrow Y$" is used to indicate this relationship, while "support" and "confidence" are used to measure the degree of it.

Agrawal et al. first put forward the problem of mining the association rules of datasets in consumer transaction databases and designed a simple algorithm [7, 23], the basic idea of which is based on the recursive method of frequent set theory. The classical frequent set algorithm is as follows.

(1) The basic idea. First, find out all the frequent sets and create the association rules from them.

(2) The procedure. The Apriori algorithm [23] is originated by Agrawal in 1994, the basic idea of which is to scan the database repeatedly. A brief description of the essential program is as follows:

L_1 = {large 1 – itemsets};

for ($k = 2; L_k - 1 \neq <; k{+}{+}$) do begin

C_k = apriori_gen ($L_k - 1$); //New candidate set

for all transactions $t \in D$ do begin

C_t = subset ($C_{k,t}$); //candidate sets that contained in event t do

for all candidates $c \in C_t$

c. count ++;

end

$L_k = \{c \in C_k \mid c.\text{count} \geq \text{min\_sup}\}$

end

Answer = $\bigcup_k L_k$.

But the main drawback of the Apriori algorithm is that it produces a large number of candidate sets during the computation process and the database may need to be scanned once again. In order to avoid the repetitiveness, an easier method must be developed.

3.2. PSO Algorithm. PSO was originated by Kennedy and Kberhart [15] and was intended for simulating social behavior, which has many advantages such as higher convergence rates and being more applicable than other algorithms. However, it generally has a lower accuracy than genetic algorithms and under certain initial conditions it can only reach optima in a subset of the problem.

PSO is a population-based search algorithm and is initialized with a population of random solutions, called particles. Unlike in the other evolutionary computation techniques, each particle in PSO is also associated with a velocity. Particles fly through the search space with velocities which are dynamically adjusted according to their historical behaviors. Therefore, the particles have the tendency to fly towards the better search area over the course of search process. The PSO was first designed to simulate birds seeking food which is defined as a "cornfield vector."

3.3. Definition of Information Entropy. Traffic accidents are events with strong randomness, while entropy is the mathematical method for analyzing an event's uncertainty. From the perspective of statistical mathematics, entropy is a measure of system randomness. Information Entropy is a value for characterizing the statistical characteristics of random variables. It is a measure of the average uncertainty of random variables, an objective description of statistical characteristics of the population.

In the traffic accident data gathering process, due to the influence and limitations from many factors, the number of traffic accident data items is usually insufficient, which cannot be used to analyze the statistical characteristics. The use of Information Entropy as a statistical measure of the uncertainty information does not require the probability distribution of the data to be known and does not require the distribution to be single humped; that is, no prior information is needed. So Information Entropy is very suitable for testing the degree of discreteness of the population. As far as Information Entropy is concerned, the decrease of uncertainty means the reduction of entropy value.

According to the definition of Information Entropy by Shannon [13],

$$H = -\sum_{k=1}^{m} p_k \log_2 p_k, \tag{1}$$

where m is the dimension of state space and p_k is the probability of the kth state.

If n^k is the number of times when k happens, and the number of samples is N, then $p_k = n_k/N$.

Information Entropy is a measure of the degree of how sequenced a system is. Data measuring with Information Entropy does not require the probability distribution to be known, and the probability distribution of the data does not have to be in a single-hump shape; therefore, it is suitable for examining the discrete level of the distribution. The value indicates how certain factor is affecting the system. The lower the value, the more important factor in the system. Applying the Information Entropy theory to traffic accident, when the attribute spreads evenly across the distribution, indicates that the accident depends weakly on that attribute, which suggests that the importance of the accident is lower.

4. The Method of Traffic Accident Data Preprocessing

Due to human factors, in the actual accident data gathering process, some data items may be unavailable [10], affecting the integrity of the data and causing the results of the causation analysis to be considered not convincible. Data preprocessing is an essential step in any of the data mining processes. Researches on data preprocessing techniques mainly focus on the preprocessing of data which follows obvious patterns, the widely used method of which is to find the patterns, characteristics, and properties so that data can be preprocessed in a certain way. In contrast, it is rarely seen that data in nondigital form is processed in the same manner. The process includes the following parts: data washing, data filling, data integration, and data transformation.

(1) This paper adopts the data-washing methodology according to the characteristics of traffic accidents to build a traffic accident database using a form which resembles how "antivirus" software works, which has an "antivirus" definition library of its own, that is, an error database that is artificially created, and to predefine a set of rules for the "software" to use in order to hunt down all the "viruses." However, before the data-washing can be initiated, a database of all the errors needs to be created. The error database comprises a gathering of area-specific knowledge and common sense, through which data that contains errors are marked through comparison. Of course, the data washing only could correct and delete the error data which is detected in logic, not all error types.

(2) Due to the fact that data format and structure in data mining are not entirely identical with those in the database, data needs to be transformed before it

can be mined, so that existing data can be changed into proper format or form to be mined through data mining techniques. The current data inside the traffic accident database happens to be coded; only part of the statistical attributes is in coded format which is required by the association rules; therefore, the data in the database should be kept as close to their original form as possible. Attributes whose attribute values are not coded but sequential ought to be scattered to a certain extend and coded in orders (seen in Table 1).

Table 1: Coding schedule of attributes on road condition.

The name of attributes	Coding of association rule	
	Attribute value	Coding
Road surface condition	Intact	1
	In construction	2
	Concave-convex	3
	Collapsed	4
	Roadblock	5
	Others	6
Road physical separation	Not separated	1
	Median separated (1)	2
	Separated between vehicles and nonvehicles (2)	3
	(1) and (2)	4
Road alignment	Straight line	1
	Common turn	21
	Sharp turn	22
	Common slope	31
	Steep slope	32
	Continuous downward slope	33
	Regular slope	41
	Sharp turn and steep slope	42
	Regular slope turn	51
	Regular turn slope	52
Roadside safeguarding types	Guard rail	1
	Barrier wall	2
	Barrier barrel	3
	Others	4
	None	5

5. The Improved Algorithm in Traffic Accident Data Mining

5.1. Fundamental Principle of the Algorithm. Focusing on the characteristics of road traffic accident data, causation analyses need to examine the double-leveled structure. Unfortunately, although they can analyze the causes of accident from different angels and each method has its advantages, none of the data mining methods currently being widely used can accomplish an overall, multiangled, multilayered data mining task on its own.

With the accumulation of traffic accidents database, the data quantity is more and more huge, so how to obtain the effective knowledge, hiding rules, and fundamental causes is changing into one key issue for road traffic administration.

To meet the demand of better accuracy and efficient analysis of traffic accident causes, this paper combines the binary PSO algorithm to improve the association rules. The reason is that the speed of the PSO algorithm does not decrease with the increase of the number of datasets. To solve the problem that accident data needs to be analyzed in different layers, this paper introduces the Information Entropy theory into road traffic accident analysis, with the help of the Association Rules theory, and puts forward the concept of Association Entropy and its algorithm.

With the introduction of PSO and Association Entropy, traffic accident causes can be analyzed from all angles and on all layers, satisfying the requirement that the association rules have to be within a certain support level. In the meanwhile, causes on different levels can provide references for different traffic administrations at different levels, so that more effective preventative measures can be taken.

5.2. Importance of Traffic Accident Attribute Value. Conducting data mining with association rules, first large item sets must be found from the original information datasets. Later, the association rules are made up of all the large item sets. It is required that the frequency of any item must be greater than the min_sup, otherwise the item is considered not common enough because it falls short of the frequency requirement, which may render it meaningless. Meanwhile, the rules calculated from the large item sets must be greater than the minimum confidence. Otherwise the results from the rule cannot be trusted. However, due to the fact that the information in the database does not include all types of traffic accident information, the ratios of the samples themselves do not match reality. Therefore, traditional data mining methods often lead to large sample volume, that is, high importance mistakes. To solve this problem, this paper made some modifications to the traditional methods, as shown in the following.

(1) According to the Association Rules theory, it is when confidence level reaches the minimum confidence min_conf and support level reaches the minimum threshold min_sup that the importance of attribute value layer starts to make sense. The function for calculating the importance of the association rules is

$$c_{ijk}^{z} = \begin{cases} \dfrac{s_{ijk}^{g}}{s_{ij}^{t}} = \dfrac{B_{ijk}/A}{D_{ij}/A} = \dfrac{B_{ijk}}{D_{ijk}}, & s_{ij}^{t} > \min\_\sup_1,\ s_{ijk}^{g} > \min\_\sup_2 \\ 0, & \text{others}, \end{cases} \tag{2}$$

where c_{ijk}^{z} is the attribute i, j's value of importance; s_{ijk}^{g} is the attribute i, j's support on k level of traffic accident severity; s_{ij}^{t} is the attribute i, j's support; B_{ijk} is the entry number of i and j on k level in the database; D_{ij} is the entry number in the database that contains I and j; A is the entry number in the database; $\min\_\sup_1$ is the minimum support of i and j; $\min\_\sup_2$ is the minimum support of i and j on k level;

The function for conditional support is

$$s_{ij}^{t} = \begin{cases} \dfrac{D_{ij}}{A}, & \dfrac{D_{ij}}{A} > \min\_\sup_1 \\ 0, & \text{others.} \end{cases} \tag{3}$$

In order to tell the severity, traffic accidents are graded according to the number of deaths. Therefore, the function for calculating the association rule support is

$$s_{ijk}^{g} = \begin{cases} \dfrac{B_{ijk}}{A}, & \dfrac{B_{ijk}}{A} > \min\_\sup_2 \\ 0, & \text{others.} \end{cases} \tag{4}$$

(2) In the traffic accident database, the following attributes are used to describe the severity of the accident: number of deaths, number of injuries, and property loss. Through surveys on experts, this model takes number of death as the determining attribute to indicate severity. The group of experts consists of 4 researchers of traffic safety at university, 4 staffs of traffic management, and 4 policemen of traffic accidents, who are at least 5 years of work experience. In order to make sure that the attribute value meets the requirement of support level, the model divides traffic accident into different levels based on the graded idea. It used the Delphi method to calculate the grade standards and the degree of importance of traffic accident p^k, shown in Table 2.

According to the calculation method of association rules, with reference of the importance value in Table 2, the importance of attribute j with respect to severity is calculated as such:

$$\rho^{j} = \sum_{k=1}^{4} p_i^k \cdot c_{ijk}^{z} - 1, \tag{5}$$

where p^k is the relative degree of importance with regard to the kth severity level; ρ^j is the attribute j's degree of importance with regard to accident severity under the association rules.

As shown in the above models, in the calculations of attribute value importance according to the association rules, it has the advantage of being able to effectively shield out certain attribute values of high importance because they have higher frequencies. It is a method of quantifying severity importance of traffic accidents from single datum under the requirement of confident level.

Table 2: Graded severity of accident and importance.

Grade	Set	Importance p^i
1	$[0, 1)$	1
2	$[1, 3)$	2
3	$[3, 10)$	6
4	$[10, \infty)$	24

In the calculation results of association rule, the value of degree of importance with regard to accident severity, the association of attributes with consequence of traffic accident, can be used to indicate the relationship of attributes and the causes of traffic accident; that is, factors with higher value of importance are the major causes in traffic accidents.

5.3. The Application of PSO Algorithm. Assume there are N particles, and each individual is treated as a volume-less particle (a point) in the D-dimensional search space. The speed, location, individual best position, and swarm best position of the ith particle at "t" moment are represented as $v_i(t), x_i(t), p_i(t)$, and $g_i(t)$, respectively, from which we can have the following recursive equations of the speed and location of the ith particle:

$$\begin{aligned} v_{id}(t+1) &= w \cdot v_{id}(t) + c_1 r_1 \left(p_{id}(t) - x_{id}(t)\right) \\ &\quad + c_2 r_2 \left(g_d(t) - x_{id}(t)\right), \\ v_{id}(t+1) &= \begin{cases} 1, & r_3 < \mathrm{Sig}\left(v_{id}(t+1)\right) \\ 0, & r_3 \geq \mathrm{Sig}\left(v_{id}(t+1)\right) \end{cases} \end{aligned} \tag{6}$$

while $i = 1, \ldots, N$, N is the size of the group, usually assigned 20; $d = 1, \ldots, D$, represents the number of dimensions of each individual, which is decided by specific problems; r_1, r_2, r_3 are random numbers within the range (0,1); c_1, c_2 are two learning factors, usually given as $c_1 = c_2 = 2$; w is greater or equal to 0 and is called the inertia factor, usually given the value of 1; $\mathrm{Sig}(\cdot)$ is the sigmoid function; that is, $\mathrm{Sig}(x) = 1/(1 + \exp(-x))$. The X_{id} of each particle in the search space is given the value of 0 or 1; $v_{id} \in [-v_{\max}, v_{\max}]$ is the probability of X_{id} equals 1; hence, when $\mathrm{Sig}(v_{id}) = 0$, X_{id} equals 0.

Because $\mathrm{Sig}(x) = 1/(1 + \exp(-x))$ when x is from the range $[-10, 10]$, we usually has a result within the range $[0, 1]$; therefore, the maximum speed $v_{\max}$ is usually less than 10.

With PSO, determining attributes of traffic accident can be sequenced to form a series of attributes all at once, no longer needing to change the sequence during the mining process, which conforms to the procedure of discrete binary PSO algorithm.

In the binary space, the moving of particles is done through switching the position values, and the speed of the particle is the change of digits after each recursive calculation. Attribute swarm is for finding out frequent item sets; each individual particle in a swarm is represented by m (m is the number of accident determining attributes) digits of $1'$s and $0'$s. And in the determining attribute swarm, each digit of

the particle is used to indicate whether the corresponding attributes appear or not; 1 means the attribute appears and 0 not.

The adaptability function is for measuring the quality of particles' rule sets. In the competition of all the rules, only the ones with high confidence and credibility may survive.

Construct a series of rule structures with all the determining and task attributes in the form of $\{B_1, B_2, \ldots, B_m, A_1, A_2, \ldots, A_n\}$, where A_i represents determining attributes, which is the severity of traffic accidents and B_i is task attributes, which are driving years, weather, and so forth.

The rule support is

$$\text{Sup}(R) = \text{Cover}(A + B) \tag{7}$$

while R is the rule; Cover$(A + B)$ is the ratio of two events in a database.

The rule confidence is

$$\text{Conf}(R) = \frac{\text{cover}(\text{A} + \text{B})}{\text{cover}} = \frac{\text{Sup}(R)}{\text{Sup}(R_A)} \tag{8}$$

while R_A represents the datasets in R that match the attributes of rule R.

Then the adaptability function of the swarm is defined as such: Fitness $= a\,\text{Sup}(R) + b\text{Conf}(R)$, while a and b are constants, and $0 \le a \le 1, 0 \le b \le 1$; the value of a and b can be adjusted according to specific problems.

5.4. Calculation of Accident Causation Attributes Association Entropy. In the calculation of Association Entropy, using each attribute value's association rule importance as the source of information analyzes the entropy to compare the importance between different attributes. Borrowing from the calculation method of Information Entropy, the Association Entropy of traffic accident is calculated as follows:

$$\widetilde{H} = \frac{H}{H_{\max}} \tag{9}$$

while H is the Association Entropy of attributes; $H_{\max}$ is the maximum association entropy value from m traffic accident attributes; According to the definition of Shannon entropy [4, 8, 10],

$$H = -\sum_{k=1}^{m} p_k \log_2 p_k, \tag{10}$$

where m is the attribute number under attribute i's values; p_{ij} is the degree of importance probability under attribute j; p_{ij} is calculated as follows:

$$p_{ij} = \frac{\rho_{ij}}{\sum_{j=1}^{m} \rho_{ij}} \tag{11}$$

while ρ_{ij} is the association rule importance of attribute j's relative determinative attribute; $(j = 1, 2, \ldots, m)$.

From (10) and (11), association importance probability reaches maximum entropy value when distributed according to importance. And the maximum entropy value is calculated as follows:

$$\begin{aligned} H_{\max} &= -\sum_{i=1}^{m} p_i \log_2 p_i = -\sum_{k=1}^{m} \frac{1}{m} \log_2 \left(\frac{1}{m}\right) \\ &= -m\frac{1}{m} \log_2 \left(\frac{1}{m}\right) = \log_2 m. \end{aligned} \tag{12}$$

Carry (10) and (12) into (9), the entropy value of traffic accident attribute importance is calculated as such:

$$\widetilde{H} = \frac{H}{H_{\max}} = -\frac{\sum_{j=1}^{m} p_{ij} \log_2 p_{ij}}{\log_2 m}. \tag{13}$$

Sequence the results according to entropy value. The larger the entropy, the more evenly distributed the association among all the attributes and the more uncertain the result; that is, the lower the importance of traffic accident attribute, and in contrast, the higher the importance of accident attributes.

6. Application and Validation of the Algorithm

6.1. Sample Data Declaration. The sample data sets in the model and algorithm validation of the paper are more than twenty thousand in total and each of them contains 56 accident attributes. And all of them are randomly collected from traffic accident database, which come from traffic accident data in Northeast China, North China, East China, South Central China, Southwest China, and Northwest China based on equivalent sampling principle, so that it can reflect the whole picture of road traffic accident nationwide better.

6.2. Validation Method. It is needed to know the computation result and standard answer of a method to test and verify the accuracy of the new model method and to compare and analyze the data. While in the validation of the traffic accident factor analysis model, because the standard answer is unavailable, accurate analysis cannot be implemented using the traditional method. But we can use another completely different and widely accepted method to carry on the analysis for the same target and then use the consistency of two kinds of methods to verify the relative accuracy of the existing model calculation results. Therefore, the Delphi method is used to compare the computation result and standard answer exactly, which is a common method in testing uncertain event. We can design expert experience questionnaire by using Delphi method to get the attribute value of pavement condition and the importance weight of road traffic accidents attribute through the way of expert experience questionnaire.

The degree of importance of attribute and attribute value are calculated through different methods in the model of the paper, so it will be conducted into two parts when analyzing the results. First, test and verify the part of attribute value which is calculated through improved association rules.

(1) For attribute value, use formulae (1) and (2) to conduct support test to the data. In consideration of the sample size in the model, we have $\text{min\_sup}_1 = 0.0001$ and $\text{min\_sup}_2 = 0.005$.

Table 3: Graded severity of accident in relation to importance.

Attribute value	The average weight by expert	The degree of importance of associated regulation	The checking result $\alpha = 0.025$ $t = 2.3646$
Common bending	0.163	0.083	Zero difference
Sharp turn	0.138	0.099	Zero difference
Common bending slope	0.053	0.145	Different
Sharp turn abrupt slope	0.159	0.151	Zero difference
Common slope sharp turn	0.102	0.171	Zero difference
Common bending abrupt slope	0.052	0.351	Zero difference

Table 4: Part of road traffic accidents attributes importance for comparative table.

Attribute	The average weight by expert	The degree of importance of associated regulation	The checking result $\alpha = 0.025$ $t = 2.3646$
Driving years	0.048	0.206	Zero difference
Road alignment	0.177	0.196	Zero difference
Weather	0.109	0.202	Zero difference
The safeguard types on roadside	0.071	0.204	Zero difference
Vehicle safety condition	0.196	0.193	Zero difference

(2) Use formulae (3) and (4) to calculate the importance of attribute value and make normalization.

(3) Use the degree of importance of attribute value as the source of information and use formulae (10) and (12) to analyze the entropy value of the attribute.

Because of the complexity of the factors of traffic accident, in data mining process if we want to get the result smoothly, we should consider the calculation results of multigroup equal precision sample and then get their average value. And the more the sample groups, the closer the final average results to the truth. So, in order to test and verify the accuracy of the computing method of the model, make random sampling analysis ten times for the data and conduct check consistency with matching t-test for the model calculating results and Delphi method results.

6.3. Calculation Results and Verification Analysis. By applying the algorithm in Sections 5.2 and 5.3 and then using the verification method in Section 6.2, the final results are listed as follows.

From Table 3 we can see that general bend and sharp bend steep slopes have the most influence on accident severity, respectively, 0.163 and 0.159. The main reason is although the former alignment condition is good, but driver has a lower safety cognitive level, often because over speed driving led to accident, the later alignment condition and road side security level are relatively low; once a vehicle loses control in this section, the result must be serious. However, real traffic accident analysis result showed that, except for general bend slope factor, other road sections have no obvious difference to the influence on accident severity. This shows that the expert's subjective impression not necessarily represents the truth. In the meantime, the accident influencing factors' absolute value of importance from association rules algorithm in different in numbers, but before significance check, also cannot represent that it really has difference between accident influence factor.

Same as the result shown in Table 3, as far as people are concerned, it is clearly different how years of driving, road geometrics, weather, and so forth affect traffic accident severity; however, the result of data mining does not show the same degree of obviousness as how the corresponding factors differ, which suggests the limitedness of people's experiences; for example, years of driving is indeed one of the major factors triggering accidents, but that does not mean relatively inexperienced drivers tend to have more severe accidents than more experienced drivers.

According to the results in Tables 3 and 4, most of the data of importance in the two methods have no difference from a statistical point of view, and only the importance of common bending slope has difference. The reason may come from the difference of the data between the experts' judgment and the model use which are limited in sample size. But in general, the importance from the two methods has almost no difference, and the relative precision of the computing results from the method can meet requirements.

7. Conclusion

This paper aims at the hierarchically structured characteristics of road traffic accident databases, mixed using the method of associated rules, PSO, and Information Entropy to analyze the degree of importance of traffic accidents. Through a modification of traditional methodologies and algorithms, a PSO algorithm and associated entropy model are built for calculating the degree of importance of road accidents. Through applying the improved algorithm on

both the attribute and the attribute value layers, respectively, each accident-triggering factor's influence on the severity of accident is calculated. The algorithm this paper introduced has the advantage of better accuracy and higher mining rates over the traditional association rules and PSO algorithms, the result of which is quite different from what the experts concluded, which indicates when facing a large amount of random information, people's experiences and how people perceive things are limited. Of course, traffic accident data as a type of data possesses certain physical meanings—whether there really exist connections between certain types of data, and that certain types of data were manually gathered so they may not be error-free, as well as whether they can be fully applied to the models and algorithms of this paper in their entirety questions as those that are yet to be discussed in future researches. However, approved by real applications and tests of effectiveness, this type of data mining method which is based on traffic accident database provides yet another powerful tool to quantify data in traffic accident analysis, which is going to be helpful to accident experts and traffic administrative agencies to clarify how much of role different factors play in investigations of traffic severity.

Acknowledgment

This project is supported by NSFC (Grant no. 50808093).

References

[1] H. Nabi, L. R. Salmi, S. Lafont, M. Chiron, M. Zins, and E. Lagarde, "Attitudes associated with behavioral predictors of serious road traffic crashes: results from the GAZEL cohort," *Injury Prevention*, vol. 13, no. 1, pp. 26–31, 2007.

[2] E. R. Green, K. R. Agent, and J. G. Pigman, "Evaluation of auto incident recording system (AIRS)," Tech. Rep. KTC-05-09/SPR277-04-1F, Kentucky Transportation Center, 2005.

[3] M. Hirasawa, "Development of traffic accident analysis system using GIS," *Proceedings of the Eastern Asia Society for Transportation Studies*, vol. 10, no. 4, pp. 1193–1198, 2005.

[4] B. Yu, W. H. K. Lam, and M. L. Tam, "Bus arrival time prediction at bus stop with multiple routes," *Transportation Research Part C*, vol. 19, no. 6, pp. 1157–1170, 2011.

[5] B. Yu, Z. Yang, and S. Li, "Real-time partway deadheading strategy based on transit service reliability assessment," *Transportation Research Part A*, vol. 46, no. 8, pp. 1265–1279, 2012.

[6] B. Yu, Z. Z. Yang, P.-H. Jin, S.-H. Wu, and B. Z. Yao, "Transit route network design using ant colony optimization," *Transportation Research Part C*, vol. 22, pp. 58–75, 2012.

[7] K. Geurts, G. Wets, T. Brijs, and K. Vanhoof, "Profiling of high-frequency accident locations by use of association rules," *Journal of the Transportation Research Board*, no. 1840, pp. 123–130, 2003.

[8] T. Tesema, A. Abraham, and C. Grosan, "Rule mining and classification of road traffic accidents using adaptive regression trees. I," *Journal of Simulation*, vol. 6, no. 10, pp. 80–94, 2005.

[9] R. Marukatat, "Structure-based rule selection framework for association rule mining of traffic accident data," in *Computational Intelligence and Security*, vol. 4456, pp. 231–239, 2007.

[10] L.-Y. Dong, G.-Y. Liu, S.-M. Yuan, Y.-L. Li, and Z.-H. Wu, "Applicaition of data mining to traffic accidents analysis," *Journal of Jilin University Science Edition*, vol. 44, no. 6, pp. 951–955, 2006.

[11] D.-H. Lee, S.-T. Jeng, and P. Chandrasekar, "Applying data mining techniques for traffic incident analysis," *Journal of the Institution of Engineers*, vol. 44, no. 2, pp. 90–101, 2004.

[12] Y. Hassan and E. Tazaki, "Emergence decision using hybrid rough sets/cellular automata," *Kybernetes*, vol. 35, no. 6, pp. 797–813, 2006.

[13] Y. Zhang, F. Wang, J. Dai, W. Huang, and Y. Chen, "New comprehensive evaluation algorithm based on fuzzy clustering and information entropy," *Journal of Changchun Post and Telecommunication Institute*, vol. 22, no. 6, pp. 643–647, 2004.

[14] Y. Shi and R. Eberhart, "A modified particle swarm optimizer," in *Proceedings of the IEEE International Conference on Evolutionary Computation (ICEC '98)*, pp. 69–73, May 1998.

[15] X. Wang, X. Liu, and F. Dai, "Method and application of mining association rules based on PSO algorithm," *Information Technology and Informatization*, vol. 3, pp. 85–87, 2009.

[16] Q. K. Pan, M. F. Tasgetiren, and Y. C. Liang, "A discrete particle swarm optimization algorithm for the no-wait flowshop scheduling problem," *Computers and Operations Research*, vol. 35, no. 9, pp. 2807–2839, 2008.

[17] T. J. Ai and V. Kachitvichyanukul, "A particle swarm optimization for the vehicle routing problem with simultaneous pickup and delivery," *Computers and Operations Research*, vol. 36, no. 5, pp. 1693–1702, 2009.

[18] P. Pongchairerks and V. Kachitvichyanukul, "A particle swarm optimization algorithm on job-shop scheduling problems with multi-purpose machines," *Asia-Pacific Journal of Operational Research*, vol. 26, no. 2, pp. 161–184, 2009.

[19] A. W. Mohemmed, N. C. Sahoo, and T. K. Geok, "Solving shortest path problem using particle swarm optimization," *Applied Soft Computing Journal*, vol. 8, no. 4, pp. 1643–1653, 2008.

[20] M. Lovbierg, T. K. Rasmussen, and T. Krink, "Hybrid particle swarm optimizer with breeding and subpopulations," *Institute of Electrical and Electronics Engineers*, no. 7, pp. 115–118, 2001.

[21] I. B. Gundogdu, F. Sari, and O. Esen, "A new approach for geographical information system-supported mapping of traffic accident data," TS1B-GIS in Urban Planning and Management, 6:03~11, 2008.

[22] ESTC, *EU Transport Accident, Incident and Casualty Databases: Current Status and Future Needs*, European Transport Safety Council, 2001.

[23] A. Savasere, E. Omiecinski, and S. Navathe, "An efficient algorithm for mining association rules in large databases," in *Proceedings of the 21th International Conference on Very Large Database (VLDB '95)*, pp. 432–443, Zurieh, Switzerland, 1995.