IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 16, 2019, accepted October 30, 2019, date of publication November 4, 2019, date of current version November 14, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2951195

Particle Swarm Optimization-Based Association Rule Mining in Big Data Environment

TONG SU, HAITAO XU, (Member, IEEE), AND XIANWEI ZHOU
School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China

Corresponding author: Haitao Xu (xuhaitao@ustb.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1003905, in part by the Natural Science Foundation of China under Grant 61971023, and in part by the Fundamental Research Funds for the Central Universities under Grant FRF-TP-18-008A3.

ABSTRACT With the explosive growth of information data in today's society, the continuous accumulation and increase of data in recent years make it difficult to extract useful information from it, so data mining comes into being. Association rule mining is an important part of data mining technology. Association rule mining is the discovery of frequent item sets in a large amount of data and the mining of strong association relations between them. Traditional association rule algorithms need to set minimum support and minimum confidence in advance. However, these two values are largely influenced by human subjectivity. Many scholars use average and weight to set these two values, but the effect is still not very good. In order to solve this problem, this paper proposed an improved algorithm of association rules - PSOFP growth algorithm, this algorithm is introduced into intelligent algorithm, particle swarm optimization algorithm, it can find the global optimal solution, we use this fact to find the optimal support, then using FP - growth algorithm for mining association rules, and finally put forward by information entropy to measure effectiveness in association rules mining, and the improved algorithm was applied to the social security event correlation analysis, the improved algorithm proved to our expectations.

INDEX TERMS Big data, association rule algorithm, particle swarm optimization, global optimal solution.

I. INTRODUCTION

In today's world, the application of data mining [1] is more and more extensive. Data mining refers to applying algorithms to a large number of data, analyzing the data and mining the hidden information. With great applications, data mining has been popularly recognized as an important research field that has drawn lots of research interests [2]. Data mining mainly includes classification, clustering, regression and correlation, such as SVM support vector machine [3], neural network [4], and so on. In the field of data mining research, association analysis [5] is a very important branch, which is based on support and confidence to mine whether there is correlation between data. One of the highly used methods for association analysis is association rule mining (ARM) [6]. Based on the association rule mining, we can find the possible associations among items in a large transaction-based dataset.

The associate editor coordinating the review of this manuscript and approving it for publication was Yuedong Xu.

Association rule mining aims to find out the association rules that satisfy predefined minimum support and confidence from a given database [7]. The basic algorithms of association rules include Apriori, FP-growth and other algorithms [8]. The principle is to first find frequent terms according to support degree by scanning data set, and then obtain association rules according to confidence degree [8]. Association rule is the probability of occurrence of Y in the case of occurrence of X, which is a conclusion summarized from a large amount of data and widely applied in real life. For example, many aspects such as the display of supermarket items, the promotion of financial products, news recommendation, warehouse planning, and the analysis of network faults [9] are of high practical value [10].

However, the minimum support and minimum confidence of traditional association rule mining algorithms are set by users themselves. Although some of them refer to the opinions of relevant experts, they are not objective enough and have no theoretical support. If the value is smaller or bigger, it will affect the results. As shown in

TABLE 1. The influence of these two values on the results.

support	confidence	results
smaller	smaller	More rules, less effective, and longer running time.
bigger	smaller	Missing important item sets.
smaller	bigger	Better rules, better effective.
bigger	bigger	Less rules, better effective, and shorter running time.

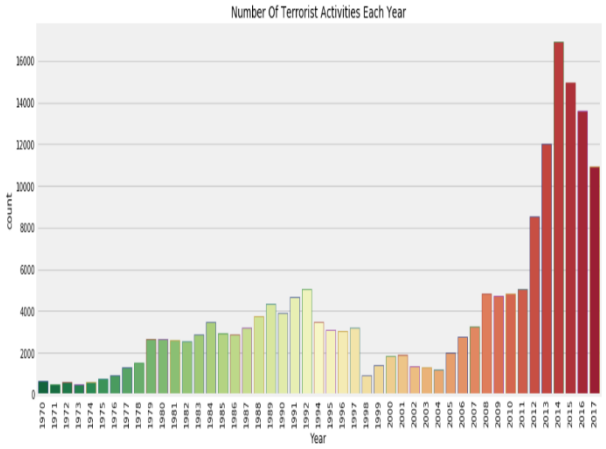

FIGURE 1. Number of terrorist activities each year.

Table 1, we analyze the influence of these two values on the results.

In order to compensate for this defect, this paper based on the existing methods, puts forward an improved algorithm of FP-growth that is PSOFP-growth algorithm. The new algorithm combines particle swarm optimization algorithm [11] with the traditional association rule algorithm. And the characteristics of particle swarm optimization algorithm is to search the optimal solution in the whole world, we use it to find the optimal support, for the association rules algorithm of scanning data set filter data to provide the best support. Then uses FP - growth association rules mining. Finally, this paper proposes to use information entropy [12] as the degree of interest to measure the effectiveness of association rules and output accurate association rules that meet the requirements.

We apply the improved algorithm to the correlation analysis of social security events, which have always been a major problem to be solved. Although the development of the world is now based on peace, social security events still exist, among which the most destructive terrorist activities are emerging in an endless stream, bringing huge disasters to our compatriots. We use records from the global terrorism database to plot Figure 1. We can see from it, since 2012, the number of terrorist incidents around the world has increased dramatically. Many of our compatriots are suffering from terrorist attacks. The occurrence of these terrorist events has the attributes of time, place, cause, use means, casualties and losses caused, etc., and only by digging out their internal correlation can we help prevent and make decisions.

The main work of this paper includes the following,

1) The particle swarm optimization algorithm is applied to the search for optimal support. This paper creatively proposes a fitness function for association mining.

2) An improved FP-growth algorithm, PSOFP-growth algorithm, is proposed, which uses the optimal support obtained by particle swarm optimization algorithm to measure the effectiveness of association rules by taking information entropy as interestingness and reduce the occurrence of invalid rules.

3) We selected 5000 social security events with attributes of time, place, attack target and casualty number. By using our improved algorithm, the association rules between events are mined and the validity of the rules is improved.

The rest of this paper is arranged as follows: in section 2, we briefly introduce the related work of some existing association relation algorithms. In section 3, we will design the mining model of association rules according to the FP-growth and particle swarm optimization algorithms and searched for the optimal support. In section 4, association rule mining is carried out, information entropy is introduced, and the effectiveness of association rules is measured by taking information entropy as degree of interest. In section 5, we compare traditional association algorithms with new ones in terms of rule effectiveness and algorithm running time. Finally, we present a conclusion in Section 6.

II. RELTED WORK

As far as we know, in the era of big data, mastering the connections between data and mining the knowledge hidden in data can play a very important role in prevention and decision-making. Association rule mining algorithms are widely used in real life, but with the amount of data is very large, the memory consumption of the algorithm is too high, the operation is too slow, and even too many invalid rules are mined, which may backfire and hinder people's judgment. Based on this defect, many scholars began to study how to improve the efficiency of association rule mining algorithm, enhance the effectiveness of rules, and made many improvements.

On association rule application, Silvaa *et al.* [13] conducted a metrological review of the literature from 2010 to 2018 and used data association rules to analyze confidential information privacy issues. Liu [14] applied the association algorithm to the university library management, and together with the collaborative filtering algorithm, realized the data mining of the library and generated the book recommendation model. Siswanto and Thariqa [15] used the cooking rules video on YouTube to analyze which cooking ingredients are most easily used in cooking recipes. They also combined the IST-EFP algorithm to reduce the dimensionality of the data set without reducing the mining association rules. Zhao *et al.* [16] used the FP-growth algorithm in keyword extraction to mine Chinese words that are always used by the author at the same time, eliminate interfering words and synonymous words. This method improves the accuracy of

keyword extraction. Kalaskar and Barkade [17] applied the association algorithm in the field of access control to try to generalize each candidate rule by changing the connection in the feature expression with constraint to overwrite the extra tuple in the consumer authority relationship.

In terms of algorithm optimization, there are also many improvements in algorithm optimization.

In order to solve the problem of low efficiency of association algorithm caused by frequent dataset replacement, Zhang and Chen [18] proposed a genetic algorithm combined with immune optimization is proposed to mine association rules. The results show that the optimized algorithm can build redundant rules and improve efficiency. Mguiris *et al.* [19] introduce an algorithm based on fuzzy Formal concept analysis and prime number coding. By constructing a frequent fuzzy minimum generator mesh, it can effectively solve the problem that the extraction fuzzy association rules run for a long time. In order to solve the problem that the association rules are inefficient in the recommendation model, Li *et al.* [20] designed a tree structure of ordered forests to compress and store frequent patterns. Experiments showed that the efficiency is significantly improved. Dea Delvia Arifin combined FP-growth with naive Bayes to mine frequent patterns of SMS and classify SMS spam [21].

The above work gives us an in-depth understanding of the contributions of other scholars in association rule mining, but there are few studies on how to improve the effectiveness of rules. Invalid association rules will mislead our judgment, so it is necessary to study this. In this paper, we will focus on how to improve the effectiveness of the mined association rules and reduce the running time of the algorithm.

III. SYSTEM MODEL

At present, association rule algorithms include Apriori algorithm, FP-growth algorithm and Eclat algorithm, etc. These algorithms generally have problems such as frequent database scanning times, too much I/O overhead and too many miscellaneous rules, which are not suitable for big data environment. Therefore, in this paper, in order to improve the above problems, an association rule mining method for large and complex discrete data is proposed in the big data environment. Firstly, the discrete data are classified and marked by letters or Numbers. Then, according to the median and average of the data, a minimum support degree is given and optimized by PSO algorithm to obtain the optimal support degree. Finally, FP-growth algorithm is mined, and the effectiveness of association rules is measured by information entropy as interestingness. This model can improve the effectiveness of relational rules, reduce memory consumption, and greatly improve user experience.

A. MODEL OVERVIEW

Let's first look at what particle swarm optimization is. Particle swarm optimization was first proposed by Eberhart and Kennedy and originated from the study of bird foraging behavior [22].

Let us imagine a scene where you put a piece of food on a clearing, surrounded by a group of birds foraging. Suppose the birds don't know where the food is, but they know how far away the current location is from the food. How do they find food? The most effective way is to find the bird closest to the food and slowly approach it. The PSO algorithm is a bio-intelligence algorithm derived from this behavior of bird foraging. In this algorithm, we call birds in the search space particles. A possible solution to this problem is given. According to the different problems to be solved, different fitness functions are set and the fitness value is calculated. Each particle has a velocity to change the direction and distance from the next position, and then changes the position of the particle.

The current optimal particle is constantly changing position in the solution space until the optimal solution is found. In the particle swarm optimization algorithm, a set of random particles, that is, a random solution, is first initialized, and then the optimal solution is obtained by iteration. In each iteration, the particles update their position. There are two important solutions in the algorithm. The first one is the optimal solution found by the particle itself, called the individual solution. The second is the optimal solution of the whole particle swarm, called the global solution. Both of the two solutions optimize the initial random solution.

We first use particle swarm optimization to find the optimal support, and then apply it to mining association rules of events.

As we all know, the occurrence of an event basically has the attributes of time, place, person, method and result. We want to explore the relationship between these attributes. Encodes each property value of an event into an event set, where each property value is called an item. The event set is scanned to get the item set, and multiple items are selected from the item set as particles of PSO optimization algorithm. Particle swarm optimization was used to search and calculate the particle, and the particle with the maximum fitness and its location were obtained as the optimal support. The frequency of major activities is greater than or equal to the optimal support project. The frequency is arranged in descending order according to the project, and the centralized event frequency pattern tree (FP-tree) is constructed according to the project sequencing order. The construction of FP-tree is used to collect all frequent projects, and the association rules between commodities are determined by the confidence and interest formula.

B. MODEL IMPLEMENTATION PROCESS

S1, Scan the event set, get the item set, calculate the occurrence times of each item in the item set, and get the item frequency. Set a minimum support value in advance;

S2, Choose N/2 term near the minimum support and randomly select N/20 as the initial particle of PSO algorithm, where N is the maximum of preset iteration times;

S3, Calculate the fitness value of each initial particle through our fitness function;

S4, start the iteration. We select the next batch of particles according to the step size, and use the selected next batch of particles to update the current particles. Then calculate the fitness value of the updated particles.

Determine whether the current iteration number reaches the preset maximum iteration number, if so, terminate the iteration, obtain the particle with the maximum fitness value, and take its position as the optimal support degree; Otherwise, execution S4 is returned.

We add information entropy into the fitness formula. The higher the information entropy is, the greater the uncertainty of the variable it represents [23]. However, when mining association rules, we want to find the certain occurrence, especially when these variables appear simultaneously. So we use the reciprocal of the information entropy. We also added the number of item of support degrees in the formula, for example, there are 5 items with a degree of support of 10 and 1 item with a degree of support of 15, which is also a factor we should consider.

The fitness function is as follows,

$$F(x) = aSupport(x) + \frac{b}{H(x)} + cCount(x) \tag{1}$$

$$H(x) = -p(x)\log_2 p(x) \tag{2}$$

where, $F(x)$ represents the fitness value of particle x, $Support(x)$ represents the Support degree of particle x, and $Count(x)$ represents the number of items that support is x, $p(x)$ and $H(x)$ represents the information entropy, a, b and c are three constants, represents the probability of occurrence of item particle x.

The current particle update formula is,

$$V_i = V_i + c_1 \times rand() \times (pbest_i - x_i) + c_2 \times rand() \times (gbest_i - x_i) \tag{3}$$

$$x_i = x_i + V_i \tag{4}$$

Among them, V_i represent the step length of particles i, and c_1, c_2 represent learning factor, $rand()$ is a random function used to generate random number between (0, 1). x_i represent the location of the particle i. The $pbest_i$ represents the optimal position of the particle itself, and the $gbest_i$ represents the optimal position of the whole particle swarm, The entire block diagram of the new algorithm is shown in figure 2

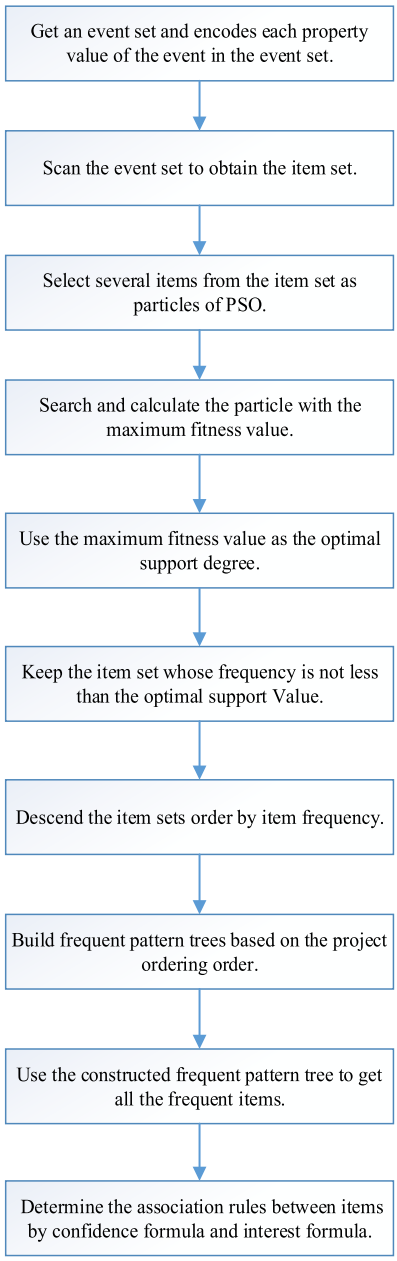

FIGURE 2. The overall flow chart of the algorithm.

IV. MINING ASSOCIATION RULES BY PSOFP-GROWTH ALGORITHM

In this section, we begin the real mining of association rules. Compared with traditional algorithms, we have made the following improvements: (1) after the minimum support is set, PSO algorithm is firstly carried out to find the optimal support, so as to avoid too small support setting, which will cause a lot of computational work, drag down the algorithm's running speed, dig out many useless rules, and waste analysis time; Avoid setting support too high. Filter out important items when scanning the event set for the first time, and finally extract no relevant association rules. (2) after obtaining the association rule according to the confidence formula, we judge whether the degree of interest is satisfied or not. We propose to measure the degree of interest by information entropy, and the association rule that satisfies the degree of interest is the one we finally need.

A. THE BASIC CONCEPT

Definition 1 (Itemset Support) [19]: Given an itemset I(set of attribute values in an event set)and an event set D, the support of an item $I_1 \subseteq I$ on D is the percentage of the events I_1 contained in D, t is an event consisting of several item particles.

$$support(I_1) = \frac{\|\{t \in D \,|I_1 \subseteq t\}\|}{\|D\|} \tag{5}$$

Tid	Items
1	I1、I10、I13
2	I3、I10、I15
3	I2、I9、 I16
4	I3、I10、I13
5	I4、I11、I14
6	I2、I11、I16
7	I5、I11、I14
8	I2、I9、 I16
9	I1、I12、I16

FIGURE 3. Nine event sets.

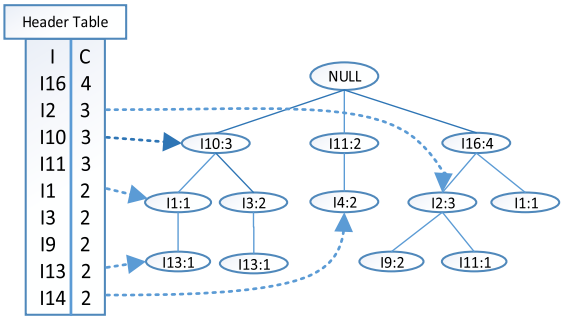

FIGURE 4. Insert the data and construct FP-tree.

Definition 2 (Confidence of the Rule) [19]: An association rule that defines patterns on I and D such as $I_1 \rightarrow I_2$ is given with Confidence. The so-called credibility of the rule refers to the ratio of transactions containing I_1 and I_2 to those containing I_1,

$$confident(I_1 \Rightarrow I_2) = \frac{support(I_1 \cup I_2)}{support(I_1)} \tag{6}$$

where, $I_1, I_2 \subseteq I$, $I_1 \cap I_2 = \emptyset$, I_1, I_2 represent items in the item set; $confident(I_1 \Rightarrow I_2)$ represents the probability of project I_2 in the case of project I_1, $\cup$ represents union set; $support(I_1 \cup I_2)$, $support(I_1)$ respectively represent the support degree of project I_1, I_2 union and project I_1.

Determine whether $confident(I_1 \Rightarrow I_2)$ is greater than the preset minimum confidence; if so, output the association rule between items I_1, I_2, that is, the probability of item I_2 occurs when item I_1 occurs.

B. BUILD FREQUENT PATTERN TREES

In this algorithm, we only scan the itemset twice and do not generate candidate sets. Instead, we directly compress the itemset into FP-tree and generate association rules through the trees.

We need to first traverse the item set, calculate the frequency of all the items, delete the items with less than the optimal support frequency, and then arrange the remaining items in descending order of frequency.

There are three main steps,

1) Generate one-time item 1-frequent item set by scanning item set, sort in descending frequency, and put it in the list of header table;

2) Create the root node, mark it as null, scan the item set again, insert the events into FP-tree one by one, and then achieve the growth of FP-tree through recursion;

3) Start from the item at the end of FP-tree, recursively up to get the conditional basis and n-frequent item set.

As shown in Figure 3, we selected 9 events, each event has three attributes, and each column represents one attribute, a total of three columns. I1-I16 respectively represents different values of the three attributes, that is, the item we said, we set the minimum support degree to 2 and the minimum confidence degree to 50%. Perform the first scan to count the occurrence frequency of items, remove the items with occurrence frequency less than 2, and put the remaining items in descending order into the Header Table in Figure 4, which is a Table containing the head pointer of the frequent 1-item set. The dotted line in Figure 4 connects the nodes with the same name in the tree. Sort each event attribute in descending order of frequency and insert it into FP-tree. First, create an empty root node, insert the first event with a frequency of 1, and then insert the second event. If the item of the second event already exists in the tree, add 1 to the frequency. If the item of the second event does not exist in the tree, create a branch with a frequency of 1. According to this approach, all the events are inserted into the tree and the FP-tree is finally built as shown in figure 4.

C. FIND FREQUENT ITEM SETS

The next step is to conduct rule mining, that is, to extract conditional pattern base from the target project. The conditional pattern base concept is interpreted as a collection of paths ending in found items, each of which is a prefix path. The purpose of extracting conditional pattern library is to find the co-occurrence item set of the target item, so that the target item can only be joined with the co-occurrence item to get n frequent item sets.

As can be seen in Figure 4, there is only one path from the root node to I14, so its conditional mode base is {i11:2}, and the further frequent item set is {i11:2, i4:2}. In the next I13, there are two paths from the root node to I3, but since each I13 has a count number of 1, which is less than the minimum support, there are no frequent itemsets in either path; In the next I9, there is a path {i6:2, i2:2, i9:2}, whose conditional mode base is {i6:2, i2:2}, and the further frequent item set is {i6:2, i9:2}{i2:2, i9:2}{i6:2, i9:2}. After searching all the items in the Header Table in turn, all the frequent item sets can be obtained.

D. EXTRACT ASSOCIATION RULE

According to the frequent item set obtained by part C, calculate the information entropy between $confident(I_1 \Rightarrow I_2)$. If the result is greater than 50%, retain it, and then calculate the information entropy according to the following formula,

$$H(x) = -p(x)\log_2 p(x) \tag{7}$$

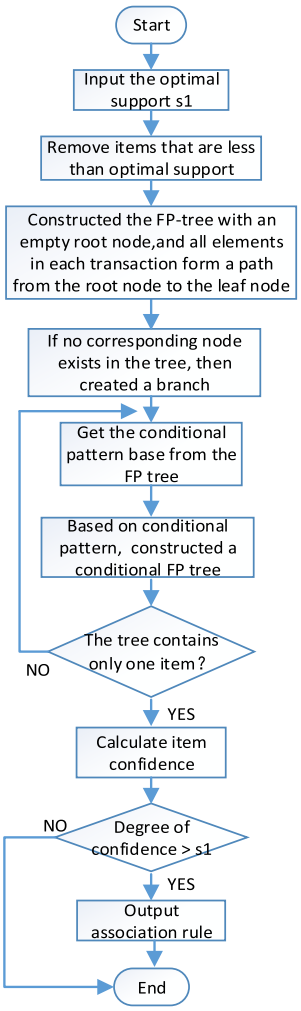

FIGURE 5. The process of FP-growth algorithm.

$$H(I_i \rightarrow I_j) = H(I_j|I_i) - H(I_j) \tag{8}$$

The higher the probability of occurrence of an event, the lower the amount of information it carries, so,

1) If $H(I_i \rightarrow I_j) < 0$, represent the value of $H(I_j|I_i)$ is less than the value of $H(I_j)$, therefore, the occurrence of I_i will facilitate the occurrence of I_j, so we think it is an effective association rule.

2) If $H(I_i \rightarrow I_j) > 0$, represent the value of $H(I_j|I_i)$ is less than the value of $H(I_j)$, therefore, the occurrence of I_i will hold the occurrence of I_j back, so we think it is an effective association rule.

E. ALGORITHM DESIGN STEPS

Through the introduction of the idea and specific design of the improved algorithm, the flow of the PSOFP-growth algorithm proposed in this paper is shown in Figure 5,

1) Code project particles and select initialization particles.

2) Run PSO algorithm to calculate the fitness value of each particle and keep iterating.

3) Judge whether the preset maximum iteration number is reached, if so, take the current particle frequency as the optimal support degree.

4) Select eligible projects according to the optimal support degree, and reorder events according to the frequency of items.

5) Set up the Header Table, where 1-frequent items are stored in descending order of frequency.

6) Build FP-tree and insert events one by one. If the current node exists, add one to the count value; if it does not exist, create a new node branch and set the count value to 1, until all the events are inserted into the FP-tree.

7) Obtain the conditional mode base from the FP-tree, and use the conditional mode base to construct a conditional FP-tree; Each newly constructed conditional frequent pattern tree is repeatedly mined for frequent patterns until all frequent items are obtained, the frequent pattern tree is empty, or the frequent pattern tree contains only a single path.

8) In the case of a single path, generate all possible combinations of sub paths, each of which is a frequent mode.

9) Calculate confidence and information entropy, and extract association rules that satisfy them.

V. NUMERICAL RESULTS

In this part, we mine association rules for social security events to verify our proposed PSOFP-growth algorithm. In order to verify the superiority of our algorithm, we compare it with traditional FP-growth algorithm. Compare the number of valid rules.

A. SIMULATION PROCESS

First, we selected 5000 social security events with attributes of time, place, attack target and casualty number.

1) Delete missing attributes and duplicate events;

2) The attributes of death toll were classified into intervals, and all the attributes in the same interval were set to the same value;

3) Other attributes, such as time, place and attack target, etc. Different attribute dimensions are represented by corresponding letters and Numbers respectively.

As shown in Table 2, we show partial data for ten events. The name of the attribute is formed by letter and abbreviation, and the value of each attribute is represented by letter and number. After mining the rule, we go to the database to find the specific name of the corresponding value and translate it into the language that human beings can understand.

Then, we calculate the average value of the occurrence frequency of all items, set it as the minimum support, and get the optimal support through particle swarm optimization algorithm. The optimal support, minimum confidence and interest are taken as input values of FP-growth algorithm to mine association rules of events.

The third step, we began to build FP-tree. First, we establish the root node, set it as null, and then insert the 5000 events left after being deleted into the FP-tree one by one.

TABLE 2. Part of 5000 events.

Num	Ccou	Ddreg	Espe	Fcity	Gm	Hsuc
0	C183	D11	E1	F1	G0	H1
1	C183	D11	E1	F1	G0	H1
2	C230	D11	E3	F1	G1	H1
3	C230	D10	E3	F1	G1	H1
4	C230	D8	E3	F1	G1	H1
5	C230	D11	E3	F1	G1	H1
6	C230	D11	E3	F1	G1	H1
7	C110	D6	E1	F1	G1	H1
8	C213	D11	E1	F1	G0	H1
9	C141	D6	E1	F1	G0	H1
10	C137	D11	E1	F1	G0	H1

```
Null Set  1
  F1  4540
   H1  4069
    G0  3613
     E1  2944
      L6  1417
       I3  1359
        J2  43
         A1  3
          D11  1
           B12  1
          D3  2
           B23  1
            M17  1
           C45  1
            K45  1
             M15  1
              B16  1
```

FIGURE 6. Part of FP-tree.

Partial results are shown in figure 6: F1 appeared the highest frequency, 4540 times, then H1, G0, E1 to B12 branch end, D3 began to appear nodes, D11 and D3 share a branch node.

The purpose of FP-tree construction is that only items on the same tree path may be frequent items, and items on different paths are not frequent items. We do not need to consider the relationship between items and the items on other branches, which can reduce a large part of the workload. The next step is to find the prefix path of the item and associate it with the target item. As shown in figure 7: this is all the conditional basis for A1 terms.

Finally, part of the association rule is shown in figure 8. The result has been to remove the rule that does not satisfy the degree of interest. The minimum confidence we set is 0.8. The first rule, in the case of H2, the probability of J5 is 0.86, which meets the requirements. So we can think of H2 and J5 as an association rule. The second rule, in the case of E10 and J6, the probability of H3 is 0.95, which also meets the requirements. Therefore, we regard E10, J6 and H3 as an association rule. With these association rules, we have a direction when preventing the occurrence of social security incidents.

For example, H2 represents the property of sharp appliances, and J5 represents the property of robbery, which can be

```
{frozenset({'E1', 'F1', 'G0', 'H1', 'I3', 'J2', 'L6'}): 3,
 frozenset({'E1', 'F1', 'G0', 'H1', 'I3', 'J14'}): 1,
 frozenset({'F1', 'H1', 'I2', 'J14', 'L5'}): 1,
 frozenset({'E1', 'F1', 'G0'}): 2,
 frozenset({'D10', 'E1', 'F1', 'G0', 'H1', 'I3', 'L6'}): 8,
 frozenset({'E1', 'F1', 'G0', 'H1', 'I2', 'J14'}): 5,
 frozenset({'F1', 'G0', 'H1', 'I3', 'L6', 'M16'}): 2,
 frozenset({'D8', 'E1', 'F1', 'G0', 'H1', 'I3', 'J2', 'L6', 'M16'}): 3,
 frozenset({'E1', 'F1', 'G0', 'H1', 'J14'}): 6,
 frozenset({'D6', 'E1', 'F1', 'G0', 'H1', 'I3', 'J14', 'L6'}): 2,
 frozenset({'F1', 'G0', 'H1', 'I2', 'J14', 'L5', 'M5'}): 6,
 frozenset({'E1', 'F1', 'G0', 'H1'}): 11,
 frozenset({'E1', 'F1', 'G0', 'H1', 'I3', 'L6'}): 15,
 frozenset({'E1', 'F1', 'G0', 'H1', 'I2', 'L5'}): 3,
 frozenset({'D9', 'E1', 'F1', 'G0', 'H1', 'I3', 'J14', 'L6', 'M16'}): 2,
 frozenset({'D6', 'E1', 'F1', 'G0', 'H1', 'J2', 'L5', 'M5'}): 2,
 frozenset({'D8', 'E1', 'F1', 'G0', 'H1', 'I3', 'J14', 'L6'}): 3,
 frozenset({'E1', 'F1', 'G0', 'H1', 'J14', 'L5', 'M5'}): 4,
 frozenset({'D10', 'E1', 'F1', 'G0', 'H1', 'I3', 'L6', 'M16'}): 9,
 frozenset({'E1', 'F1', 'G0', 'H1', 'I2', 'L5', 'M5'}): 4,
 frozenset({'E1', 'F1', 'G0', 'H1', 'I2', 'J14', 'L5', 'M5'}): 14,
  ......
}
```

FIGURE 7. The condition basis of A1 term.

```
Rules
frozenset({'H2'})  => frozenset({'J5'}) 's conf:  0.86
frozenset({'J6'})  => frozenset({'H3'}) 's conf:  0.94
frozenset({'H3'})  => frozenset({'J6'}) 's conf:  0.97
frozenset({'J6', 'I14'})   =>  frozenset({'H3'}) 's conf:  0.97
frozenset({'H3', 'I14'})   =>  frozenset({'J6'}) 's conf:  0.99
frozenset({'J6', 'E6'})    =>  frozenset({'H3'}) 's conf:  0.93
frozenset({'E6', 'H3'})    =>  frozenset({'J6'}) 's conf:  0.99
frozenset({'J6', 'K0'})    =>  frozenset({'H3'}) 's conf:  0.95
frozenset({'H3', 'K0'})    =>  frozenset({'J6'}) 's conf:  0.97
frozenset({'J6', 'E10'})   =>  frozenset({'H3'}) 's conf:  0.95
frozenset({'H3', 'E10'})   =>  frozenset({'J6'}) 's conf:  0.97
```

FIGURE 8. Association rules.

understood as if the target person buy some sharp appliances, they are likely to rob, so we can give an early warning to the bank and other institutions, and they can take precautions in advance.

B. RESULTS CONTRAST

In this part, we compare the improved PSOPF-growth algorithm with the traditional Apriori and FP-growth algorithm. As can be seen from the Table 3, by changing the preset support value, the number of association rules by PSOFP-growth algorithm is nearly 30% less than that by Apriori and FP-growth algorithm, and many invalid association rules are eliminated.

We conduct experiments on the running time of the improved new algorithm PSOFP-growth and the traditional algorithm Apriori and FP-growth, and the results are shown in figure 9. As can be seen from the figure, the new algorithm takes much less time than the traditional algorithm. Moreover, the higher the support is set, the shorter the running time of

TABLE 3. The number of rules generated is compared with the traditional algorithm.

Support (count)	Apriori (count)	FP-growth (count)	PSOFP-growth (count)
20	2150	2030	1480
40	1385	1289	855
50	898	975	558
80	350	367	105

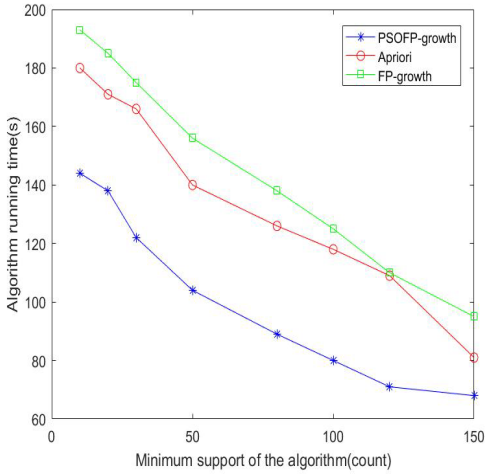

FIGURE 9. Algorithm running time.

the algorithm is, which proves that the value of support has a great influence on the algorithm.

VI. CONCLUSION

The mining of association rules can not only be applied in the social security events as shown in this paper, but also in many aspects such as marketing and social network. Therefore, it is necessary to deeply study association rule algorithms. Due to the explosive growth of data, the applicability of traditional association rule algorithms becomes weak, and it is difficult to find the rules we need directly from a large number of data. Therefore, this paper proposes an association rule mining scheme in the context of big data. Firstly, particle swarm optimization algorithm is used to find the best support and avoid artificial blind setting. Secondly, FP-growth was used to mine association rules. Finally, information entropy was used as interest to measure the effectiveness of association rules, which made it easier for us to extract effective information. In the future, we will further improve our proposals to accommodate more situations.

REFERENCES

[1] D. Fisch, E. Kalkowski, and B. Sick, "Knowledge fusion for probabilistic generative classifiers with data mining applications," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 3, pp. 652–666, Mar. 2014.

[2] H. Hui, C. Zhou, S. Xu, and F. Lin, "A novel secure data transmission scheme in industrial Internet of Things," *China Commun.*, to be published.

[3] W. Wu and H. Zhou, "Data-driven diagnosis of cervical cancer with support vector machine-based approaches," *IEEE Access*, vol. 5, pp. 25189–25195, 2017.

[4] G. Liang, H. Hong, W. Xie, and L. Zheng, "Combining convolutional neural network with recursive neural network for blood cell image classification," *IEEE Access*, vol. 6, pp. 36188–36197, 2018.

[5] X. Liu and H.-W. Shen, "Association analysis for visual exploration of multivariate scientific data sets," *IEEE Trans. Vis. Comput. Graphics*, vol. 22, no. 1, pp. 955–964, Jan. 2016.

[6] J. Han and Y. Fu, "Mining multiple-level association rules in large databases," *IEEE Trans. Knowl. Data Eng.*, vol. 11, no. 5, pp. 798–805, Sep. 1999.

[7] L. H. Son, F. Chiclana, R. Kumar, M. Mittal, M. Khari, J. M. Chatterjee, and S. W. Baik, "ARM–AMO: An efficient association rule mining algorithm based on animal migration optimization," *Knowl.-Based Syst.*, vol. 154, pp. 68–80, Aug. 2018.

[8] F. Coenen, P. Leng, and S. Ahmed, "Data structure for association rule mining: T-trees and P-trees," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 6, pp. 774–778, Jun. 2004.

[9] A. Zakari, S. P. Lee, and C. Y. Chong, "Simultaneous localization of software faults based on complex network theory," *IEEE Access*, vol. 6, pp. 23990–24002, 2018.

[10] R. Zhong and H. Wang, "Research of commonly used association rules mining algorithm in data mining," in *Proc. IEEE Inter. Conf. Internet Comput. Inf. Services*, Hong Kong, Sep. 2011, pp. 219–222.

[11] B. Cao, J. Zhao, Z. Lv, X. Liu, S. Yang, X. Kang, and K. Kang, "Distributed parallel particle swarm optimization for multi-objective and many-objective large-scale optimization," *IEEE Access*, vol. 5, pp. 8214–8221, 2017.

[12] F. Zong, X. Chen, J. Tang, P. Yu, and T. Wu, "Analyzing traffic crash severity with combination of information entropy and Bayesian network," *IEEE Access*, vol. 7, pp. 63288–63302, 2019.

[13] J. Silvaa, J. J. Cubillosb, J. V. Villac, L. Romero, D. Solano, and C. Fernández, "Preservation of confidential information privacy and association rule hiding for data mining: A bibliometric review," *Procedia Comput. Sci.*, vol. 151, pp. 1219–1224, Jan. 2019.

[14] Y. Liu, "Data mining of university library management based on improved collaborative filtering association rules algorithm," *Wireless Pers. Commun.*, vol. 102, no. 4, pp. 3781–3790, Oct. 2018.

[15] B. Siswanto and P. Thariqa, "Association rules mining for identifying popular ingredients on YouTube cooking recipes videos," in *Proc. IEEE Indonesian Assoc. Pattern Recognit. Int. Conf. (INAPR)*, Jakarta, Indonesia, Sep. 2018, pp. 95–98.

[16] M. Zhao, W. Yu, W. Lu, Q. Liu, and J. Li, "Chinese document keyword extraction algorithm based on FP-growth," in *Proc. IEEE Int. Conf. Smart City Syst. Eng. (ICSCSE)*, Hunan, China, Nov. 2016, pp. 202–205.

[17] A. Kalaskar and V. Barkade, "FP-growth policy mining for access control policies," in *Proc. IEEE 4th Int. Conf. Comput. Commun. Control Automat. (ICCUBEA)*, Pune, India, Aug. 2018, pp. 1–4.

[18] G. Zhang and H. Chen, "Immune Optimization based Genetic Algorithm for incremental association rules mining," in *Proc. IEEE Int. Conf. Artif. Intell. Comput. Intell.*, Shanghai, China, Nov. 2009, pp. 341–345.

[19] I. Mguiris, H. Amdouni, and M. M. Gammoudi, "An Algorithm for fuzzy association rules extraction based on prime number coding," in *Proc. IEEE 26th Int. Conf. Enabling Technol., Infrastruct. Collaborative Enterprises (WETICE)*, Poznan, Poland, Jun. 2017, pp. 182–184.

[20] C. Li, W. Liang, Z. Wu, and J. Cao, "An efficient distributed-computing framework for association-rule-based recommendation," in *Proc. IEEE Int. Conf. Web Services (ICWS)*, San Francisco, CA, USA, Jul. 2018, pp. 339–342.

[21] D. D. Arifin, Shaufiah, and M. A. Bijaksana, "Enhancing spam detection on mobile phone short message service (SMS) performance using FP-growth and Naive Bayes classifier," in *Proc. IEEE Asia–Pacific Conf. Wireless Mobile (APWiMob)*, Bandung, Indonesia, Sep. 2016, pp. 80–84.

[22] A. Gupta, V. Pattanaik, and M. Singh, "Enhancing K means by unsupervised learning using PSO algorithm," in *Proc. IEEE Int. Conf. Comput., Commun. Automat. (ICCCA)*, Greater Noida, India, May 2017, pp. 228–233.

[23] X. Zhang, Y. Wang, and L. Wu, "Research on cross language text keyword extraction based on information entropy and TextRank," in *Proc. IEEE 3rd Inf. Technol., Netw., Electron. Automat. Control Conf. (ITNEC)*, Chengdu, China, Mar. 2019, pp. 16–19.

TONG SU is currently pursuing the master's degree with the School of Computer and Communication Engineering, University of Science and Technology Beijing, China. Her research interests include data science, data privacy, and security.

HAITAO XU received the B.S. degree in communication engineering from Sun Yat-Sen University, in 2007, the M.S. degree in communication system and signal processing from the University of Bristol, in 2009, and the Ph.D. degree from the University of Science and Technology Beijing (USTB), in 2014. He was engaged in post doctor study with the Department of Software Engineering, USTB, from 2014 to 2016, where he is currently an Associate Professor. He held a visiting professor position with the Electrical and Computer Engineering Department, University of Houston, from October 2016 to April 2017. He has published 50 articles and one book for cyber security. His research interests include wireless communication, game theory, secure communications, cognitive radio, and mobile edge computing.

XIANWEI ZHOU is currently a Professor with the Department of Communication Engineering, School of Computer and Communication Engineering, University of Science and Technology Beijing. His research interests include the security of communication networks, next-generation networks, mobile computing, scheduling theory, and game theory.