Integration of information security and network data mining technology in the era of big data

Lu Li[1]

Abstract. The purpose of this paper is to apply data mining technology to effectively analyze and process these data and find valuable information that can help the decision and understanding. In the paper, privacy preserving algorithm to data mining is studied, and the methods of classification of privacy preserving algorithm for existing data mining are introduced. In addition, the existing algorithms are summarized from the perspective of data processing technology, and the algorithm is evaluated and analyzed with the given criteria. Moreover, a privacy preserving mining algorithm for frequent patterns based on the increase of noise is put forward, which solves two key problems: the noise increase and transaction noise in the way of the noise generated. At last, an experiment is designed to verify the effectiveness of privacy preserving algorithm facing frequent pattern mining proposed by this paper. And in the same experimental platform, the time and space efficiency of privacy preserving based on frequent pattern mining in data cleaning are compared. The experimental results and comparative analysis show that the the privacy preserving system has good performance. In conclusion, the effectiveness of the system is verified and it can be used in the protection of information security.

Key words. Data mining, privacy preserving, frequent pattern mining.

1. Introduction

Various privacy preserving techniques are gradually applied to various branches of data mining, including classification, clustering, association rule mining and so on. A lot of ways to protect the sensitive information have been put forward. But no matter which way is used to protect the privacy, it will produce different degrees of damage to the quality of data [1]. As a result, in the privacy preserving, we must consider the impact of mining results. To ensure the correctness and effectiveness of the mining results is the ultimate goal of the mining, and the privacy preserving is the needs of the data provider [2], which should be considered.

This paper introduces several classification standards of privacy preserving algorithm for data mining. In addition, from the point of view of data processing

[1] Nanjing Audit University Jinshen College, Jiangsu Nanjing 210000, China

technology, the typical algorithms of data mining privacy preserving are introduced in detail, the advantages and disadvantages of each algorithm are analyzed, and the evaluation standard of privacy preserving algorithm for data mining is put forward. What is more, the privacy preserving algorithm for frequent pattern mining is discussed with great attention. On this basis, the method for increasing the noise is introduced into the frequent pattern mining privacy preserving, noise data generation is illustrated in details, and the experimental verification is carried out. The experimental results showed that this algorithm can effectively improve the efficiency of privacy preserving algorithm for frequent pattern model mining.

2. Materials and methods

First of all, the classification of privacy preserving algorithms for data mining is introduced, and then, from the data processing technology, the privacy preserving algorithm facing data mining of the current typical collective data set is analyzed.

2.1. Privacy preserving techniques and algorithm analysis for data mining

The existing privacy preserving algorithms for data mining can be classified from the aspects of data distribution, data processing technology, data mining algorithm, privacy preserving object and so on angles.

According to the different data storage methods, the data set used for data mining can be divided into centralized data and distributed data. The distributed data can be divided into two levels: horizontal distribution and vertical distribution. The horizontal distribution refers to the distribution of data in different sites in accordance with records, and the vertical distribution indicates the distribution of data n different sites according to the properties [3]. The privacy preserving algorithm based on data mining can be classified according to the different data storage methods. In this paper, the studied algorithm is the centralized data privacy preserving algorithm.

The privacy preserving algorithm based on data mining can also be classified according to different data processing methods, such as data cleaning, data conversion, data block, data encryption, data anonymity and so on.

Different privacy preserving algorithms are suitable for different data mining tasks. For instance, some privacy preserving algorithms are suitable for classification, some are suitable for clustering, while some are generally suitable for both classification and clustering [4]. The privacy preserving algorithm based on data mining can be classified according to the application problems.

According to the privacy preserving objects, privacy preserving algorithms for data mining can be divided into the following two categories. One category is the objects of privacy preserving with the sensitive data in the data source, the other is the objects of privacy preserving with the implied sensitive knowledge in the data source.

2.2. Data mining privacy preserving based on data cleaning

Based on the idea of data cleaning, Oliveria and so on scholars proposed a series of privacy preserving association rule mining algorithms. The SWA algorithm realizes the purpose of privacy preserving by deleting some data. Specifically, the solution of the problem can be described as follows: if there are the two sides of cooperation A and B [5]. A has the transaction data set, and B desires to mine the association rules. The problem is that if A doesn't want B to dig out some rules, then A needs to implement a number of techniques, and the preserved rules are called sensitive rules.

The SWA algorithm involves three basic concepts: sensitive rules, sensitive models and sensitive transactions. Sensitive rules refer to a collection of association rules that are necessary to be hidden in the original data set. Sensitive patterns can also be called sensitive frequent item sets, which refer to all of the frequent item sets set mined out from it, supporting the frequent item sets of the sensitive rules. Sensitive transactions are transaction records that contain sensitive patterns in the original database [6].

2.3. Data mining privacy preserving based on data encryption

Medical researchers may not have the expertise to analyze the data, and its software and hardware facilities are not complete. For a large number of medical data, they can only not analyze or looking for professionals to analyze. No analysis of data will result in a waste of data resources, and professional analysis has a potential security problem. Using increasing noise method or data cleaning method, it will reduce the accuracy of results, so these two kinds of methods are not suitable for medical data sets. The reason is that the medical field has a high requirement on security of data and accuracy of results. While the use of data encryption method can just avoid this problem.

In this paper, a data set encryption algorithm is proposed to protect the privacy information. Medical data can be generally divided into digital data, character data, time data, and image data these four kinds. As a result, it is possible to make encryption of data of different types by defining a set of reversible conversion rules. Symbol f1 refers to the conversion function of digital data, f2 indicates the conversion function of character data, f3 suggests the conversion function of time data, and f4 represents the conversion function of image data. Then, the new table is built according to the old table structure by using the conversion rules, and then the content of the old table is converted into the new table according to the transformation rule, and the new data set is generated. Medical research institutions only transfer the encrypted data set for the data analysis professionals. The data analysis professionals make data mining of the encrypted data table, and then the mined results are returned back to the medical scientific research institutions. Medical research institutions only need to convert the results into the initial stage in accordance with reverse conversion, and then they can get the meaningful results.

2.4. Privacy preserving algorithm for frequent pattern mining

The knowledge discovered by data mining can be used not only for derivation of sensitive information from non-sensitive information, but the knowledge itself may be the sensitive information related to national security, business secrets, personal privacy and so on. If data mining technology is abused by some malicious users, it will pose a threat to privacy and information security in the data sharing.

Based on some existing research ideas, this paper focuses on how to protect the hidden knowledge information in the original data set. The basic research idea is: before the release of the data set, in advance, make change processing of the original data set, so as to prevent the leakage of these sensitive knowledge. However, transforming the original data set will inevitably distort some non-sensitive information, and the true extent of them will be influenced. Thus, it will make misleading for the data receiver, especially for some important non-sensitive information. The purpose of this paper is to reduce the side effects of the knowledge information as much as possible in the change processing of the original data set, and focus on the privacy preserving in the mining of frequent patterns.

As a basic research in data mining, frequent pattern mining has its application value in data mining, such as association rule mining, feature extraction, data classification and clustering. At the same time, frequent pattern itself is an effective means of knowledge expression. The objective for studying the frequent pattern privacy preserving is, without disclosing sensitive patterns, to preserve non-sensitive patterns as much as possible in the data retention, especially some non-sensitive patterns containing important information. And it aims at improving the usability of the data set, which can identify the non-sensitive patterns containing important information by using frequent pattern mining.

Figure 1 shows the flow of processing the original data set to get the result data set.

The original data set contains all the frequent patterns, and the result data set contains only non-sensitive patterns.

First of all, the frequent pattern mining is carried out in the original data set to get the corresponding set of frequent pattern mining. Data owners, through the analysis, determine which contains sensitive information in the frequent pattern mining (called sensitive pattern or private pattern). Then, according to the frequent pattern mining (some of which are frequent patterns that have been identified as sensitive pattern by data owners), apply the privacy preserving algorithm to process the original data set and get a new results data set. As a result, when the results data set is mined, it will not lead to the leakage of the sensitive pattern in the original data set. That is to say, even if the attackers do data mining on the results data set, they cannot find the sensitive patterns in the original data set. At the same time, taking into account the increase of the usability of the data set, it also requires reducing the impact of data change on those non sensitive patterns in the original data set as much as possible, especially some non-sensitive patterns containing important information, so as to improve the usability of the results data set.

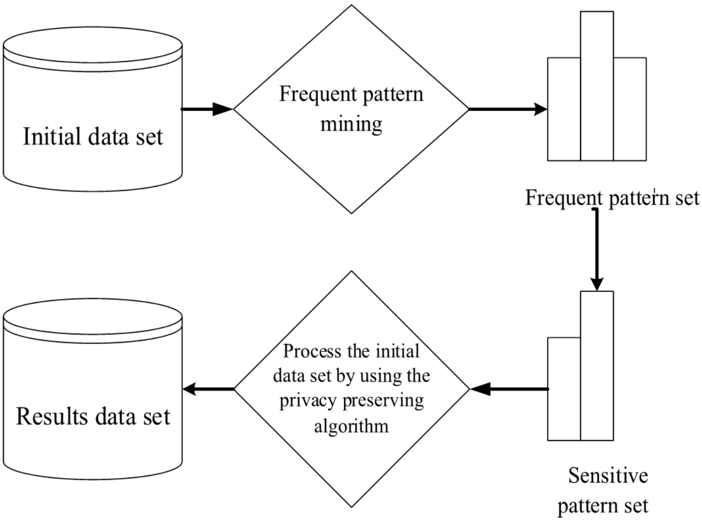

Fig. 1. Schematic diagram of privacy preserving flow in frequent pattern mining

In order to solve the problem of privacy preserving algorithm for frequent pattern mining based on data mining, this paper takes the method of increasing noise to transform the original data set, so as to realize the preserving of sensitive patterns. That is, by increasing some transaction records to the original data set D, to obtain the results data set D'.

Because the transaction mode support equals to the number of transactions containing patterns divided by the total number of transactions, it is necessary to hide the sensitive patterns in the results data set D'. That is, to reduce the support of the sensitive pattern to be less than the specified support threshold. The following two ways can be applied: one is to reduce the number of transactions containing sensitive pattern. The preserving of sensitive patterns based on data cleaning is, by deleting sensitive items in the transaction containing sensitive patterns, to reduce the number of transactions that contain sensitive patterns, thereby reducing the support of sensitive patterns in D'. The second is to increase the total number of transactions. That is to say, increasing noise transactions in the original data set, to reduce the support of sensitive patterns, so as to hide the sensitive patterns.

The basic idea of increasing noise sensitive mode preserving is shown below in Fig. 2.

Firstly, the original data set is performed with frequent pattern mining, obtaining the sensitive pattern set and release mode set. According to the above pattern set, appropriate noise transaction is generated. Then, the original data set and noise data are fused to produce the final results data set D'. Among them, the principle of increasing the noise transaction should be: to reduce the support of sensitive patterns in the sensitive pattern set and to reduce the support of the release pattern in the release pattern set.

In this paper, the core problems that the noise increasing method need to solve

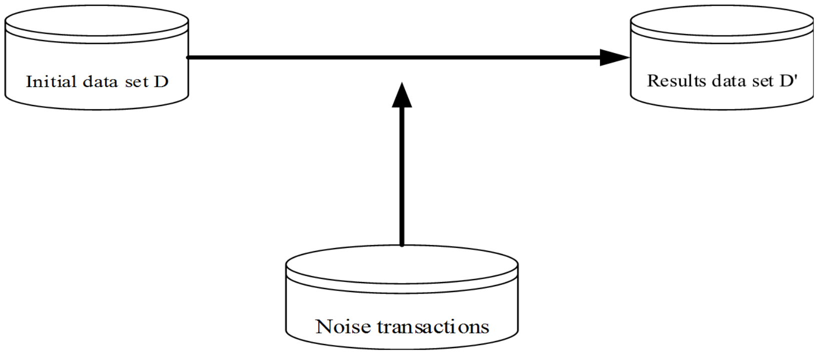

Fig. 2. Increased noise sensitive mode preserving

are summed up in two aspects. One is to increase the number of noise transactions, and the other is to generate the noise transaction, that is, which is included in the noise transactions. The idea to solve the problem of this article is: through the support of sensitive pattern and privacy preserving threshold set in advance, to calculate the number of noise transactions needed to increase, and to generate the noise transactions through the decomposition of the existing non-sensitive patterns.

3. Results

In this chapter, we design and implement a privacy preserving mining algorithm for frequent pattern mining, which is used to verify the effectiveness and performance of the new privacy preserving algorithm based on increased noise designed in the paper. And the privacy preserving algorithm for frequent pattern mining based on data cleaning on the same platform is realized, the privacy preserving algorithm for frequent pattern mining based on noise increase is compared in a number of indicators, and the conclusion is given.

3.1. Experimental flow framework of privacy preserving for frequent pattern mining

The first stage is to preprocess the original data set, convert the data in the original data set for the form that the frequent pattern mining needs, and then apply the typical frequent pattern mining algorithm, according to the given support threshold, to find all the frequent patterns (frequent item sets), and then randomly select sensitive pattern set and released pattern set.

The second stage is, based on the results of the first stage (sensitive pattern set and released pattern set), to apply the privacy preserving algorithm for mining frequent patterns in the original data set. That is, to apply the privacy preserving algorithm for frequent patterns based on data cleaning and the privacy preserving algorithm for frequent pattern based on noise increase for the processing, and get the final results data set.

The third stage is to verify the results. For the results data set generated in the second stage, frequent pattern mining algorithm is used to obtain the frequent patterns in D'. It also verifies whether the sensitive patterns in the first stage is hidden, that is, to verify the effectiveness of privacy preserving algorithm in the second stage.

Specific experimental environment is:

Hardware: Microsoft: IBMX86, CPU Intel T5750 2.00G, Memory: 2G

Software: operating system: Windows XP SP2, developing language is JAVA

Data: the data set in the experiment is generated by IBM data generator according to different parameter configuration dynamics.

3.2. Experimental results analysis

We compare the effectiveness and efficiency of the two algorithms through two groups of experiments. In the experiment, in a given minimum support threshold, we first of all use the frequent pattern mining algorithm to get all the frequent patterns in the original data set. And then, from the mining results, we randomly choose 15 frequent patterns to constitute a sensitive pattern set, and take the rest non-sensitive patterns as the release pattern set.

In the first set of experiments, we use the support error to measure the effectiveness of the two algorithms. The support error α is defined as

$$\alpha = \sum_{i=1}^{k} \frac{\mathrm{SUP_{D'}}(\mathrm{P}_i) - \mathrm{SUP_D}(\mathrm{P}_i)}{\mathrm{SUP_D}(\mathrm{P}_i)}\,. \tag{1}$$

Here, P_i refers to the pattern in the release pattern, $\mathrm{SUP_{D'}}(\mathrm{P}_i)$ indicates the support of the pattern P_i in the results data set D', and $\mathrm{SUP_D}(\mathrm{P}_i)$ suggests the support of the pattern P_i in the original data set. From the definition of support error α, the greater the value of α, the greater the impact of the algorithm on the results data set. In the case of initial data set not changing (the number of transactions in the data set is 10 thousand), we gradually increase the privacy preserving threshold to compare the effectiveness of the two algorithms. The experimental results are shown in Fig. 3.

In this set of experiments, we set the minimum support threshold of frequent pattern mining as 2 %, the original data set keeps unchanged, and the privacy preserving threshold is gradually increased from 10 % to 30 %. It can be seen from the figure that, for the preserving algorithm of sensitive pattern based on data cleaning, a smaller threshold means to protect the privacy of sensitive pattern in a higher degree. That is to say, more sensitive transactions are processed. In consequence, the support degree error is higher. And when the privacy preserving threshold increases gradually, since that the impact of the results data set on the original data set is smaller, the support error of two algorithms is gradually approaching.

In the second set of experiments, we compare the efficiency of the two algorithms by increasing the size of the original data set in the case of the same threshold. The experimental results are shown in Fig. 4.

In this set of experiments, we set the threshold of privacy preserving as 20 %, and

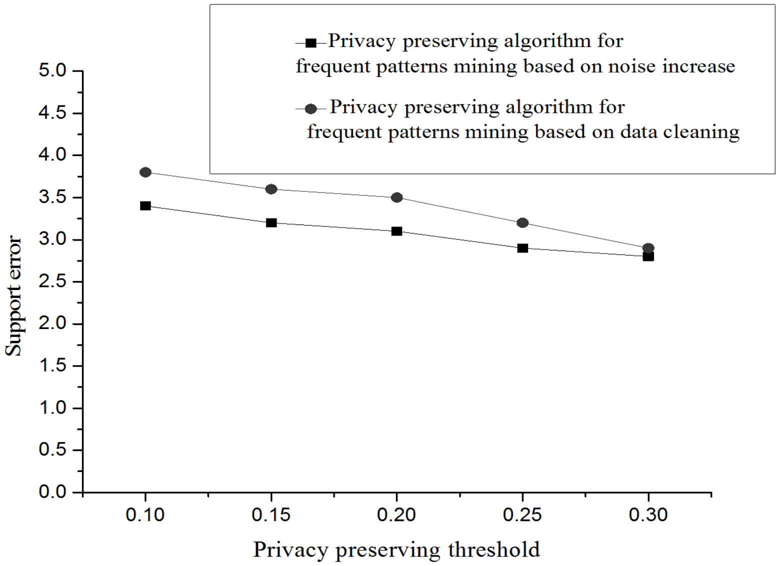

Fig. 3. Comparison of support error under different privacy preserving threshold

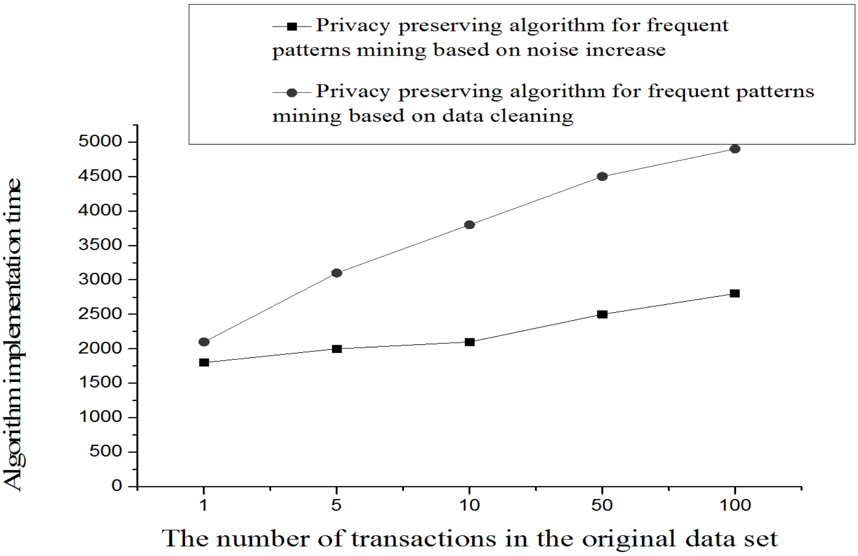

Fig. 4. Comparison of algorithm execution time under different original data sets

the minimum support threshold of frequent pattern mining as 1.5 %, but the initial data set is gradually increasing. As can be seen from the figure that, when the original data set is the same, the implementation time of privacy preserving algorithm for frequent patterns mining based on noise increase is less than that of the privacy preserving algorithm for frequent patterns mining based on data cleaning. With the gradual increasing of the original data set, the algorithm implementation time of privacy preserving algorithm for frequent patterns mining based on data cleaning is much greater than that of privacy preserving algorithm for frequent patterns mining based on noise increase.

From the above two experiments, in different experimental conditions, the privacy preserving algorithm for frequent patterns mining based on noise increase, from effectiveness, complexity and scalability, is superior to privacy preserving algorithm for frequent patterns mining based on data cleaning.

4. Conclusion

As a large number of private data or enterprise data are widely collected and analyzed, the application of data mining technology in these data containing sensitive information may be a threat to the privacy of individuals or businesses. Therefore, it is very meaningful to study how to deal with privacy preserving for the data sets applied in different fields of data mining. Based on the existing privacy preserving algorithm for frequent pattern mining based on data cleaning, a privacy preserving algorithm for frequent pattern mining is proposed. Compared with the method of hiding the sensitive patterns by deleting the sensitive items, the algorithm proposed in this paper can achieve the hiding of sensitive patterns by increasing the noise transactions. In addition, we design and implement the privacy preserving experiments for frequent pattern mining, and validate the effectiveness and performance of privacy preserving algorithm for frequent pattern mining based on the noise increase through experiments.

References

[1] Y. C. Chiu, S. Chen, G. J. Wu, Y. H. Lin: *Three-dimensional computer-aided human factors engineering analysis of a grafting robot.* Journal of Agricultural Safety & Health *18* (2012), No. 3, 181–194.

[2] Z. Li, X. G. Han, J. Sheng, S. J. Ma: *Virtual reality for improving balance in patients after stroke: A systematic review and meta-analysis.* Clinical rehabilitation *30* (2016), No. 5, 432–440.

[3] Y. Huang, Q. Huang, S. Ali, X. Zhai, X. Bi, R. Liu: *Rehabilitation using virtual reality technology: A bibliometric analysis, 1996—2015.* Scientometrics *109* (2016), No. 3, 1547–1559.

[4] S. Strangio, P. Palestri, M. Lanuzza, F. Crupi, D. Esseni, L. Selmi: *Assessment of InAs/AlGaSb tunnel-FET virtual technology platform for low-power digital circuits.* IEEE Transactions on Electron Devices *63* (2016), No. 7, 2749–2756.

[5] Y. Chao-Gan, Z. Yu-Feng: *DPARSF: A MATLAB toolbox for "pipeline" data analysis of resting-state fMRI.* Frontiers in Systems Neuroscience *4* (2010), No. 13, 13.

[6] J. A. Chen, M. S. Tutwiler, S. J. Metcalf, A. Kamarainen, T. Grotzer, C. Dede: *A multi-user virtual environment to support students' self-efficacy and interest in science: A latent growth model analysis.* Learning and Instruction *41* (2016), 11–22.

Received May 22, 2017