# Algorithms Feasibility Inquiry Based on Data Mining in Privacy

**Linhai Tan**

*School of Artificial Intelligence, Hunan Vocational College of Science & Technology, Changsha, 416488246@qq.com*

***Abstract:*** *This paper firstly summarizes the current research status of privacy protection data mining algorithms and the significance of researching privacy protection data mining; and then according to the different distribution of data objects, this paper discusses the corresponding privacy protection mining methods of integrated data and distributed data respectively, and then it analyses and studies association rule mining algorithms and SVM classification mining algorithms; And focusing on distributed database system classification data mining which is horizontal distribution, privacy protection classification algorithm based on the SVM is proposed. The mathematical model has been established, and experimented with the method of computer simulation. The results show that the algorithm has certain stability under the circumstances of distributed node increases, and the algorithm is feasible and has a practical guiding significance.*

***Keywords:*** *privacy protection, integrated data, Distributed, association rule mining, SVM classification mining*

## 1. Introduction

In the current information age, the wide application of database technology has brought great convenience to the storage, management, daily query of the mass data. While at the same time some new problems has been produced. On the one hand, the continuous emerging mass data are collected and stored in vast large database, and people could not understand and apply them under the situation of lacking a strong analysis tool[1-3]. As a result, the utilization rate of mass data is extremely low, and some data become difficult to access afterwards. On the other hand, the mined original data imply some strong privacy sensitive data, or the mining results have certain confidential, so we must take corresponding measures to protect them, ensuring data mining be done in the case of privacy protection. Using the privacy protection data mining technology, the original data and the mining results can be prevented from leaking out effectively and thus reducing the threat of personal privacy by data mining[4,5].

Before building the privacy protection method of data mining, the first is to learn the data object form and the distribution features of the data, and then on that basis combine data mining method and the safety factor to design and form a privacy protection method.

## 2. Privacy protection data mining method

### *2.1 The privacy protection data mining of data concentrated distribution*

The privacy protection data mining basic flow of the integrated data is shown in Figure 1:

At present, in the privacy protection classification mining of data concentrated distribution and it mainly adopts the following two types of mining method.

(1) Random response method: using a model from statistics—“Warner model”, the data is transformed randomly, and on the basis of converted data, the value probability of original data is derived, and finally classification mining algorithm is carried out.

(2) Add a random offset method: a method that add random offset in the original data, and then reconstruct the distribution of data.

Combine random response technology and the classification algorithm to apply them to privacy protection data mining. Due to its simple and easy to use, high efficiency, it has a good effect. But there are also some limitations about random response technology: it can only process attribute values for a Boolean data type (binary data)[6-8].

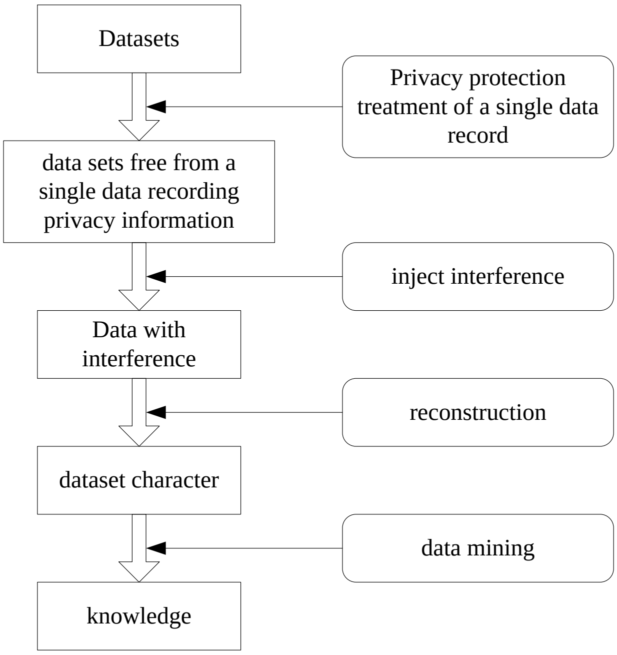
Datasets
Privacy protection treatment of a single data record
data sets free from a single data recording privacy information
inject interference
Data with interference
reconstruction
dataset character
data mining
knowledge

*Figure 1 The privacy protection data mining basic flow of the integrated data*

As for the method of adding random quantity to original data, the original value must be known, and this has brought the possibility of divulge the information. In addition, this method also requires the original data meeting a certain probability distribution and not suitable for classification properties, and has large amount of calculation.

### *2.2 The privacy protection data mining of distributed data*

The privacy protection data mining basic flow of distributed data is shown in Figure 2:

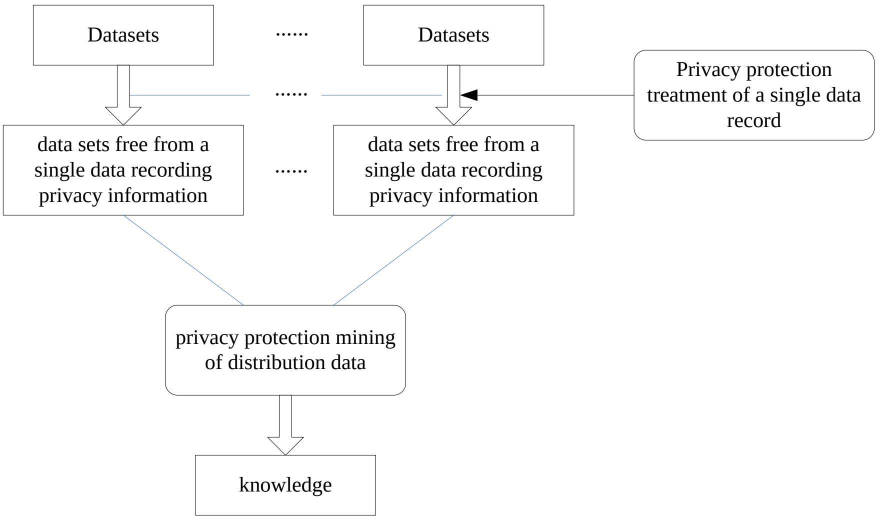
Datasets
......
Datasets
Privacy protection treatment of a single data record
......
data sets free from a single data recording privacy information
......
data sets free from a single data recording privacy information
privacy protection mining of distribution data
knowledge

*Figure 2 The privacy protection data mining basic flow of distributed data*

Data mining can extract important knowledge from a large number of datasets, but sometimes these datasets is scattered across many sites. The data warehouse can get data from many data sources in certain permission, but it will increase the risk of privacy being violated. In addition, the factors such as privacy, law and commercial secret has carried on the limits to the centralized access of data, impeding the implement of data mining. However, the privacy protection data mining technology of distributed data has resolved this problem. In the process of mining, each site only knows the final results, and can not speculate data information of any other site according to the mining results, so any data privacy will not be revealed[6].

In a distributed environment, secure multi-party computation (SMC) protocol is usually used to solve the problem of privacy protection. When several parties participate in data mining, all parties wish to accomplish data mining privacy task under the premise of its own data privacy, and ensure that any private data information of the site cannot be infer from the mining results. In the privacy protection data mining of distributed data, according to different forms of data partition, the privacy protection data mining of distributed data can be divided into two kinds of horizontal distribution and vertical distribution.

## 3. Privacy protection mining algorithm aiming at two kinds of distribution

### *3.1 Association rule mining algorithm*

According to the different protection objects, privacy protection association rule mining algorithm of data centralized distribution can be also divided into protecting rules algorithm and protecting the original data algorithm. At present most algorithms are based on the protection rules, so next the paper mainly introduces the algorithm of protection rules.

The proposal of the protection rules is from the data transfer. The protection rules refers to ensure the data miner can only dig out preset rules in the data, and the rules that data owners want to protect or hide cannot be exposed. For example, D represents the original datasets, and R stands for rules that can be digged out from datasets D, however, as for rule R, some rules R that the data owners do not want to be mined[7]. And how to transformed the original datasets D into a public datasets exit, and ensure the data miners on the datasets D to dig out rules except rule R' successfully. Figure 3 shows mining algorithm model of protection rules.

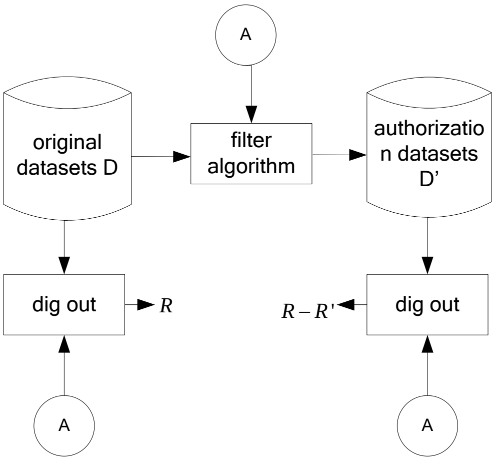

*Figure 3 The mining algorithm model of protection rules.*

For the problem of protecting rules, one method is to select a data item, and replace the value of 1 with 0, in this way, the support degree of protecting rule will be reduced, and meanwhile ensure that the support degree of non-protection rules that hidden cannot be affected by too much in the released changed database. Another method is to make the confidence level of the important rules less than the specified minimum confidence level of users, so as to prevent the rule needed to protect from digging out. However, this is a NP problem, and it is difficult to establish a balance between the time complexity and side effects of algorithm. At the same time, due to modifications of the non-sensitive rules and infrequent item sets into the frequent item sets, some side effects have been produced, reducing the accuracy of released database.

The frequent item sets mining algorithm of privacy protection is to filter data record and protect sensitive rules and reduce the influence of data cleaning stage through a transaction retrieval methods, achieving a certain degree of balance between finding as many rules and protecting privacy data. The main ideas of the algorithm are as follows: suppose D stands for the original datasets, P stands for frequent item set digged out from datasets D, $R_1$ stands for a group of rule that needed to hide, and a set of item sets $P_1$ can be able to deduce rule $R_1$, $P_1$ is a set of item sets that needed to hide, while $P_2$ is a set of item sets that needed not to hide, of which $P_1 \cup P_2 = P$.

The process of algorithm is divided into the following four parts:

(1) Find out all the frequent item set $P$ ;

(2) According to the protecting rules, $P$ can be divided into $P_1$ and $P_2$.

(3) According to the transaction retrieval method, find out the sensitive record from the original datasets by $P_1$ .

(4) Using a kind of data processing algorithms which can delete limit model to produce database $D'$ without sensitive rules.

### *3.2 SVM classification mining algorithm of privacy protection*

In a distributed environment, each node are data holders, so each node must ensure their data privacy before summarizing data to the data center, and meanwhile when collaboration calculation of distribution node has been carried out, each node should also prevent the mutual information leakage. Therefore, from the direction of data flow, each holder must take effective measures to ensure data privacy. Privacy protection algorithm steps are as follows:

(1) The primary node 1 produces a same size random matrix with local matrix.

(2) The primary node adds the random matrix with local matrix together, and sends the sum to the next node.

(3) Each slave node receives the interference matrix, and adds the matrix with local matrix, and then sent to the next slave node (the last slave node sends the data to the primary node).

(4) The primary node minus the random matrix after receiving the data sent by the last slave node, and the data obtain is the sum of all the matrices, and in the process the data of each node hasn’t been revealed.

Algorithm adopts the adding mechanism, and it has been proved to be safe and effective. In this algorithm, only you can see the accurate original data. Each node calculates its own local data, namely Gram matrix. In order to be able to realize privacy protection, the algorithm needs three or more nodes at least. Each node can get global data. The global data actually is the sum of all nodes. This algorithm can achieve better privacy protection, but when the system exist collusion attack, the algorithm is more fragile. If there are two nodes with collaborative attack, then it is likely that the accurate data of the third party will be obtained.

## 4. Algorithm analysis and experiment

### *4.1 Model building*

In the distributed database system, in order to reduce the number of variable in function, we use kemd function, so the main problem is converted into the following problem:

$$\min_{\alpha} \frac{1}{2} \alpha' Q \alpha - e' \alpha$$

$$s.t. \quad 0 \le \alpha_i \le v \quad and \quad \sum_i d_i \alpha_i = 0, i = 0, \cdot 1; m$$

In the case of linear inseparable, a straight line which can separate these classes cannot be found. Linear SVM, in this case, can be extended into data creation non-linear SVM of linear inseparable data (also known as nonlinear non-linear separable data, or the nonlinear data for short).This kind of SVM can find nonlinear decision boundary input in the space (i.e., nonlinear hypersurface).

This method has the following two steps:

Firstly, the nonlinear mapping has been used in order to transform the original input data to a high dimensional space. This step many kinds of nonlinear mapping can be used, making the data transform to a new high dimensional space.

Secondly, search the linear separating hyperplane in the new space. This problem has become into a quadratic optimization problem again, the formula of linear SVM can be used to solve maximal margin hyperplane in the new space corresponding to nonlinear separating hyperplane in the original space.

In order to solve the problems of the coordinated attack in the algorithm, using the method presented in reference to dispose Gram matrix of each node. First the data held by each node is Xi, m represents node number.

(1) The data $x_i$ of each node is split into m number, of which $x_i = \sum_{j=1}^{m} x_{ij}$ , $x_{ij}$ and $x_i$ belongs to the same range.

(2) $x_{i1}$ of each node should be kept, and $x_{i2}, \cdots, x_{im}$ be sent to the other m-1nodes randomly.

(3) Each node has received the other m-1 different number.

(4) Each node add its own $x_{i1}$ and received m-1 different number together, and send the data to data center of the second party (DC).

The data that the third party obtained is :

$$\sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} = \sum_{i=1}^{n} x_i$$

In the process of the algorithm, each node of the third party always keeps the privacy of $x_{i1}$ . So this kind of algorithm is more difficult for the collusion attack. Without the data center not attending the collusion attack, even the other m-1 nodes carry out collusion attack together, the value of $x_1$ cannot be acquired, while only get $\sum_{i=2}^{m-1} x_i$ . They still cannot get real data value of node data, and this has nothing to do with the other parameters. With the data center participating in the collusion attack, the value of the node will be reveal only after collusion between the receiving the nodes of m-1 data piece and data center.

### *4.2 Algorithm analysis and experiment*

In a horizontal classification mining algorithm of distributed database, suppose there are a total of 200 distributed nodes in each node. And in the process of experiment, assume the data split into m = 3, 5, 10, 30. The node data leakage is shown in Figure 4.

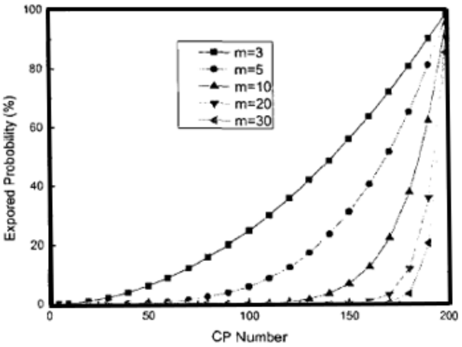

*Figure 4 Algorithm performances under the collusion attack*

As we can see from figure 4, as the number of collusion nodes increases, the probability of data leakage becomes much greater. But with the increase of m, the probability of data leakage becomes much smaller. In the case of fewer collusion nodes, the probability of information leakage is nearly zero. However, in the case of more collusion nodes, the probability of information leakage decreases with the increase of the value m. Thus this algorithm can meet the requirement of security classification mining in the distributed database.

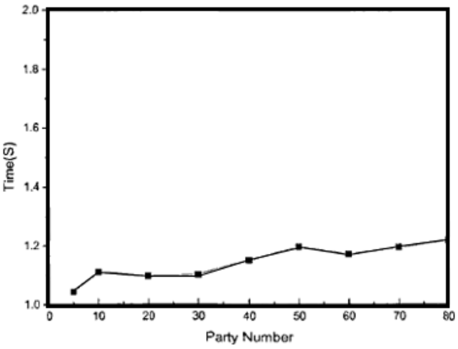

*Figure 5 The running time*

In distributed database system which is the horizontal distributed, using SVM algorithm to the classification of data mining is same with classify mining of the centralized database on the aspect of performance. In the algorithm experiments, it mainly considers the running stability of algorithm, when distributed nodes increases. In the process of experiment, 81 x81 simulated data is used, and this data is vertical distributed in the nodes of 3, 10, 20, 30, 40, 50, 60, 70; the center node deals with the data submitted by other nodes, and then carries out classification mining. The running time under various situations is shown in figure 5. As can be seen from the figure 5, with the increase of number of nodes, the scope of running time increase is not big, and has certain stability for the node.

## 5. Conclusion

The above is related research and discussion process on the subject of privacy protection aiming at data mining. Combined with the basic methods of data mining, the data distribution and construction of the system, the privacy protection data mining of the centralized data and distributed data flow chart is mapped, and also carries out association rule mining algorithm and the SVM classification mining algorithm, and then through the establishment of model and computer simulation experiment, it can be concluded that the algorithm has a certain stability.

## References

*[1] Yao Yao, Ji Genlin. Distributed Clustering Algorithm Based on Privacy Protection [J]. Computer Science, 2009, 36(3):100-102*

*[2] Chen Xiaoming, Li Junhuai, Peng Jun, etc. A Survey of Preserving Data Mining Algorithms [J]. Computer Science, 2007, 34(6):183-186, 19*

*[3] Chen Wenwei, Huang Jincai, Zhao Xinyu. Date Mining technology [M]. Beijing: Beijing University of Technology Press, 2002. 5-6*

*[4] Ma Tinghuai, Tang Meili. Date Mining Based on Privacy Protection [J]. Computer Engineering, 2008. 5*

*[5] Zhang Peng, Tong Yunhai, Tang Shiwei and etc. An effective Method for Privacy Preserving Association Rule Mining [J]. Journey of Software, 2006, 17(8):1765-1774*

*[6] Zhang Yuanping, Zhong Bo. Error Analysis for an Algorithm of Privacy-preserving Rule Mining [J]. Computer Science, 2006, 33(8):82-84*

*[7] Liu Yinghua, Yang Bingru, Ma Na and etc. State of the art in distributed privacy preserving data mining [J]. Application Research of Computers, 2011, 28(10):3606-3610*

*[8] Mi, C., et al., A novel experimental teaching approach for electrical engineering based on semi-physical simulation [J]. World Transactions on Engineering and Technology Education, 2014, 12(4):p. 779-783.*