Preservation Of Privacy In Mining Using Association Rule Technique

Prof. Geetika. Narang, Anjum Shaikh, Arti Sonawane, Kanchan Shegar, Madhuri Andhale

Abstract: - Privacy has become an important issue in Data Mining. Many methods have been brought out to solve this problem. This paper deals with the problem of association rule mining which preserves the confidentiality of each database. In order to find the association rule, each participant has to share their own data. Thus, much privacy information may be broadcasted or been illegal used. These issues can be divided into two categories: data hiding and knowledge hiding. This paper reviews the major method of privacy preserving on each category and chooses some of them to complete our system. At the end, an improvement of sensitive rule hiding is proposed to make it more accurate and secured.

Index Terms: - *Association rule; data mining; privacy preserving; data hiding; knowledge hiding, SMEs (Small and Medium Enterprise)*

I INTRODUCTION

Association rule mining, as a very important technique, has already been applied in a wide range of areas. However, it breaks out many privacy issues. From a general point of view, it may classify privacy issues into two broad categories. The first is related to the data perse and is known as data hiding, while the second concerns the information ,or else the knowledge, that a data mining method may discover after having analyzed the data , and is known as knowledge hiding. Data hiding tries to remove confidential or private information from the data before its disclosure. In this case, many randomization methods have been addressed. The randomization method has been traditionally used in the context of distorting data by probability distribution. The miner uses the perturbation data to get the association rule. In this case, the data miner does not know the raw data and also can get the similar result. Which the key point for data miner is how to reconstruct the raw data distribution. R.Agrawal[5] gave out a method bayes reconstruction method. Randomization method will change the raw database, there will be many negative impacts, just can be classify into two parts:

- Useful rules have been lost;
- New rules have been produced artificially.

There will be a contradiction between accuracy and security, by perturbing the raw data so much that the mining result will turn to poor. In the same, by perturbing it so little that the hacker will get the data they want more easily. To solve this problem, Secure multi-party computation is addressed, it use security communication protocols and cryptography, make it possible to mining rules on the raw data, this can reduce the negative impact. But this will cause huge communication costs and complex cryptography, which will make the algorithm low efficiency.

- Co-Author- Prof. Geetika Narang
- Department of Computer Engineering, Sinhgad Institute Of Technology, Lonavla

Knowledge hiding, on the other hand, is concerned with the sanitization of confidential knowledge from the data. As a result of association rule mining, many useful association rules will be discovered, but at the same time, many privacy rules will also be exposed which do not want others to know. To solve this, it need to limit the mining process, in order to keep these sensitive rules being hidden. There are so many methods to solve this problem. Which we used is just one kind of them, called support-based and confidence-based blocking schemes. We are developing a Shopping Mall Management Suite SMMS for SMEs (Small and Medium Enterprise) comprising observation, privacy preservation, prediction and strategic analysis. The Suite shall encompass utilities for handling, streamlining and standardizing day to day operations, planning, analysis and prediction by means of an indigenous correlation mechanism based on Indian psychology, behavior, preferences, history, social structure and other parameters. Also as previously stated, the lack of a software to ‘predict’ the demands and needs of Indian middle and lower class people determined us to develop a software which would integrate both the features - of SMMS and of 'prediction'. All the existing ones had only a single feature and mostly are not user friendly. We propose a plan which would give the SMEs the much needed power of SMMS along with predicting the people's choices and also the company's growth. It would take inputs of various factors - demographic, social, professional, personal etc. - which would be used to record the trends, read the psychology and present them with the product that will best suit them. It would also be highly user friendly for them to understand it.

II. APPROACHES

In this section, the main approaches used in our system are presented. Apriori algorithm is adopted to find the frequent item sets, then in order to get the global support and confidence without privacy disclosure, secure computation is used. For the knowledge hiding, an improved algorithm is mentioned to get satisfy result.

A. Secure computation

The basic idea of secure multiparty computation (SMC) is that a computation is secure if at the end of the computation, no party knows anything except its own input and the results. Security sum is a very simple and useful method which base on SMC. It is used for get the sum of data from the different site.

B. Association rule hiding

There are two modification schemes that incorporate unknowns and aim at the hiding of predictive association rules, i.e., rules containing the sensitive items on their LHS (left-hand-side).The algorithms presented in require a reduced number of database scans and exhibit an efficient pruning strategy. However, by construction, they are assigned the task of hiding all the rules containing the sensitive items on their LHS, while the algorithms in the work of can hide any specific rule. The first strategy, called ISL, decreases the confidence of a rule by increasing the support of the item sets in its LHS. The second approach, called DSR, reduces the confidence of the rule by decreasing the support of the item sets in its RHS (right-hand-side). Both algorithms experience the item ordering effect under which, based on the order that the sensitive items are hidden, the produced sanitized databases are different.

III. OUR ALGORITHMS

A. The Association rule mining System

Fig.1 gives a general structure of system which can be divided into three steps:

First step: Security sum is used to get the global support and confidence, which keep the participant don't know the local support of each sit.

Second step: SMC is adopted to send all item sets from different sit to one center, without data leakage.

Third step: In this step system the sensitive rules will be hidden. It's a very important step, building up a

Sensitive rules Database, Which contain the rules should be hidden. Then something should do to hide these rules. The first two steps are belonging to the data hiding process and the third step is the knowledge hiding process.

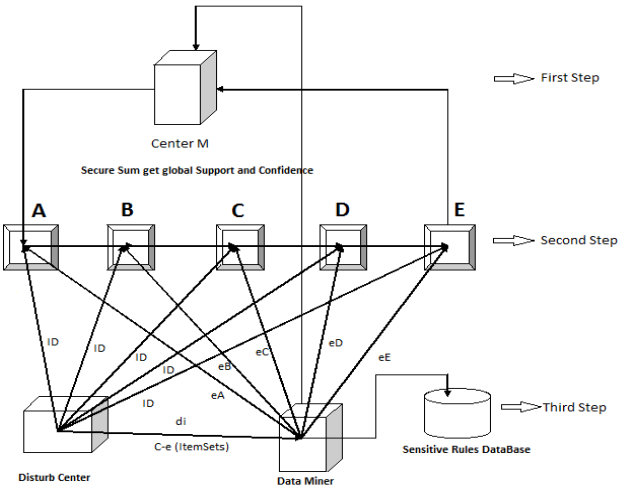

Fig.l.association rule mining system

B. Data hiding

Center M produce a random data matrix, which meet the. 2) Disturb Center use di decrypt the data sit and remove the ID 3) Disturb Center disrupts the order of the data, send C-e(frequent itemset) to Data Miner 4) Data Miner decrypts the data, getting frequent itemsets.

C. Knowledge hiding

In traditional methods, to hide a sensitive rule, some item should be deleted or use an unknown data to change the raw data. In terms of Association rule mining, rule A=>B will be discovered, as long as satisfy these two conditions:

1. **Support(A=>B) = P(A and B)>= Min_sup;**
2. **Confidence(A=>B)=P(A|B) = Support(A U B)/ Support(A) >=Min_conf**

Decrease the support of B or decrease the support of A U B or increase the support of n, all of them can realize hiding the rule. But there is a conflict, when a support of items is changed some other insensitive rule will be hidden at the same time or other new forgery rule will be produced average distribution, then send this matrix to sit A, follow the Security Sum method, so we get the global support and confidence of each item from the local support and confidence on each sit.

The TEA encrypt algorithm is used. Encryption step:

1) Data Miner produce public encryption key: C-e

2) Data Miner produces key-pair of each sit (ei, di)

3) Send ei to each sit, di to Disturb Center

4) Disturb Center send ID to each sit

5) Each sit Encrypt their frequent itemset to ID ei (C-e(frequent itemset))

Decryption step:

1) Each sit send their data ID ei (C-e(frequentitemset)) S.-L.Wang[1] proposes two data mining algorithms for hiding informative association rules, namely Increase Support of LHS (ISL) and Decrease Support of RHS (DSR). first algorithm tries to increase the support of left hand side of the rule. The algorithm is as follow :

Algorithm ISL

Input:

(1)a source database D,

(2)a min_support

(3)a min_confidence

(4)a set of predicating items X

Output: a transformed database D',where rules containing X on LHS will be hidden.

1. Find large 1-itemsets from D';
2. for each predicating item x € X
3. If x is not large 1-itemsets then X:=X-{x};
4. If X is empty, then EXIT;//no rules contains X in LHS

5. Find large 2-itemsets from D;

6. For each x€X {

7. For each large 2-itemsets containing x {

8. Compute confidence rule U,where U is rule like x-> y

9. If conf(U)<min_conf ,then

10. Go to next Large 2-itemset;

11. Else {

12. Find TL={t in D|t does not support U};

13.Sort TL in ascendind order by number of items;

14.While(conf(U)>=min_conf and TL is empty){

15.Choose the first transaction t from TL;

16.Modify t to support x,the LHS(U);

17.Compute support and confidence of U;

18.Remove and save the first transaction t frorn TL;

19.}// End While

20.}//End if conf(U)< min_conf

21.If TL is empty,then {

22.Can not hide x-> y;

23.Restore D;

24.Go to next large 2-itemset

25.}// end if TL is empty

26.}// end of for each large 2-itemset

27.Remove x from X;

28.}// end for each x € X

29.Output updated D,as transformed D'

In ISL algorithm, the sensitive rules will be hidden, but some insensitive rules may have been hidden also and many new rules may have been produced artificially. To solve this problem, system should use the mining results to restraint the choosing process (which choose item to modify). In sensitive rule hiding process, we choose other item as sacrifice item, in order to get a better results. Then we add some noise rules to increased security. As short itemset is chosen as sacrifice item, because it has less overhead. The first step is the same with ISL or DSR, then check the mining result, in this step, using privacy quantify to measure the result, if it's dissatisfied, return to change the choose policy. Check the whole database, classify the items with their support, the items with support more close to Min_sup as "unsettled itemsets" the other is "settled itemsets", when choose the sacrifice item, the item in "settled itemsets" will be chosen first.

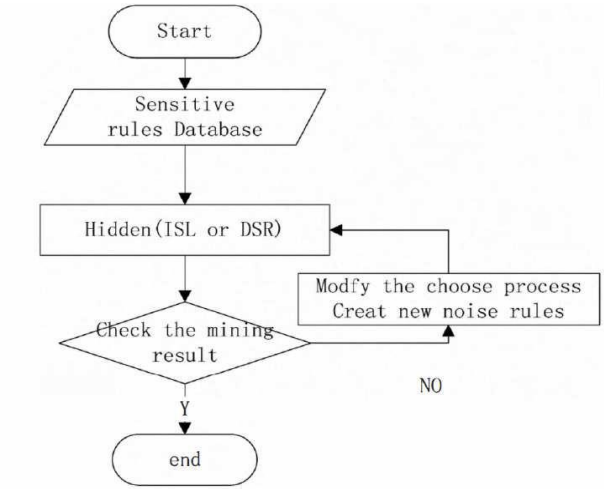

Fig 2 sensitive rules hiding process

The main disadvantage of a blocking algorithm is the fact that the dataset, apart from the blocked values, is not distorted. Thus, an adversary can disclose the hidden rules by identifying those generating item sets that contain question marks and lead to rules with a maximum confidence that lies above the minimum confidence threshold. If the number of these rules is small then the probability of identifying the sensitive ones among them becomes high.

IV. CONCLUSION

The main approach of privacy preservation when doing association rule mining, construction a system for data mining ,by using Secure computation and TEA encryption technology is carried out. It avoids data leakage which cause by data sharing. The knowledge hiding, using ISL achieve sensitive rules hiding, and present an optimization method to get a better result.

ACKNOWLEDGEMENT

This research was supported by Sinhgad Institute of Technology, Lonavala computer department of Pune University.

REFERENCES

[1]. Tinghuai Ma, Sainan Wang, Zhong Liu, "Privacy Preserving Based on Association Rule Mining", in:proceedings of third international conference on advance computer theory and engineering,2010

[2]. Chris Clifton, Murat Kantarcioglou, Xiadong Lin, "Tools For Privacy Preserving Distributed Data Mining" SIGKDD Explorations,2003,4{2}:28-34

[3]. S-L Wang and A.Jafari, "Hiding Informative associative rule sets", Expert systems with Application 33(2007) 316-323

[4]. Shyue-Liang Wang, “Maintenance of sanitizing informative association rules”, Expert systems with Application 33(2009) 4006-4012

[5]. R. Agarwal, R. Srikant, “Fast Algorithms for mining association rules” in:proceedings of 20th international conference on very Large Databases, Santiago, Chile, September 12-15,1994

[6]. Simon Shepherd, ”The Tiny Encryption Algorithm (TEA)”, Professor of Computational Mathematics

[7]. Director of the Cryptography and Computer Security Laboratory,Bradford University, England