ISSN:2229-6093

Jayashree Patil et al ,Int.J.Computer Technology & Applications,Vol 3 (6), 1967-1970

Association Rule for Privacy Preserving in Data Mining

Jayashree Patil
BVDUCOE, Pune, India

Y.C.Kulkarni
BVDUCOE, Pune, India

Abstract

Data mining is the computer oriented process to evaluate vast database and extract the meaning of the data. Privacy is inter-related with secrecy and anonymity. Thus privacy preserving in data mining means maintaining secrecy which is concerned with the information that the others can retrieve from us. Then the loss of privacy means leakage of that information. So to mine the information from huge database number of algorithms have developed that assures mining with privacy preserving. These algorithms are used to secure the sensitive items while extracting the appropriate knowledge from database. In this paper, the methodology used to preserve the privacy in data mining using association rule is used because association rule mining is one of the important aspect in data mining. In this algorithm, secure multiparty computation is used which assures security using cryptograpy is also discussed in brief. This algorithm ensures better privacy preserving with high efficiency.

.

1. Introduction

The process of extracting significant information from the very large amount of database is called data mining. In many business organizations, data mining has emerged as one of the key feature. So the privacy has become an important issue in data mining. A very important technique has been applied in a wide range of areas called as association rule mining. Association rule mining greatly helps in protecting the secrecy and confidentiality of each database. From a general point of view, privacy information may be transmitted and illegally used. These problems and issues can be divided into two separate categories: one is data hiding and the second one is knowledge hiding.

Data hiding tries to elimiate secret, confidential , private information from the huge data before its explorsure. Many methods are addressed, based on randomization, which is used to the data using probability distribution. In this randomization, data miner should focus on key point of reconstruction of raw data distribution. Bayes reconstruction method is mentioned by Agrawal R[1].

Effect of randomization method is that it modifies the raw database which results in negative impact such as useful rules may be misplaced or omitted and artifically new rules may be constructed. So this results in contradiction of accuracy and secrecy. Solution to this problem is to use secure multiparty computation(SMC) . Secure multiparty computation is communication protocol and cryptography approach. Following figure shows SMC vs Randomization used for privacy and secrecy.

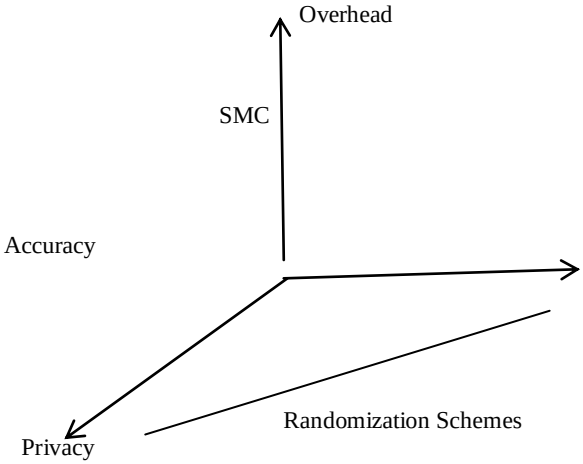

Figure 1: SMC vs Randomization

On the other side, knowledge hiding is concerned with the sanitization of confidential knowledge from data. For this many association rule mining algorithms are useful but it is quite possible that few privacy rules can be exposed which the user do not need to know. To solve this, mining process can be limited in order to safe the sensitive rule. Many methods have been addressed for limiting minig process in order to keep sensitive rules hidden. This methods are based on support-based and confidence base blocking schemes. The main disadvantage of a blocking algorithm is that adversary can know the hidden rules by identifying those itemsets that

contains question marks and lead to rules with maximum confidence that is above the minimum confidence threshold. This is because blocking algorithm blocked values is not modified. [3][4].

2. Techniques

The main approach used in this system is presented in this section. Frequenct item sets are detected using apriori algorithm. For finding global support and confidence without privacy leakage secure computation is used. For satisfying result, knowledge hiding algorithm has been improved..

2.1 Secure Computation

Yao has introduced history of the multi-party computation problem[4]. Goldreich et al has extend[5] and many others also.

Secure multiparty computation enables privacy preserving without trusted third party. The basic idea behind secure multiparty computation (SMC) is computation is secure if no party can revel anything except its own input and results at the ned of the computation. SMC uses security sum which is very simple and useful. It is used to find the summation of different data from the different site. SMC is great achievement in modern cryptography. It enables a set of untrusted parties to compute any function of their private inputs releaving only the output of any function and not sensitive information. For making computation secure , cryptograghy is used while sending the data.

2.2 Associaton rule mining.

Description of the rule hiding problem is given. Suppose transactional database is denoted as D and set of literal which are frequently called as item are denoted as I = {i1,. . . , in} then each transaction is an item set that is included in I. To sort the items in a an item set or in any transaction lexicograaphic order can be used. Using bitmap notation transaction t in the any database D can be represented as a triple ie <TID ,values of items, size>, where TID is the identifier of the transaction t. If an item has value in the 'values of item' in triple as one, then an item gets the support of transaction t and vice-versa ie if an item has value in the 'values of items' in triple is zero, then it is not supported by t . Size is the number the number of items supported by the transaction.

An association rule, is an effect of the form X => Y between two disjoint item sets X and Y in I. Both a support and a confidence value is assigned to each rule. Support is calculated as the measure of a rule frequency and more precisely it is the probability to find in the database transactions containing all the items in union of X and Y. The confidence is calcuated as a measure of the strength of the relation between the antecedent X and the consequent Y of the rule, and more precisely is, the probability to find transactions containing all the items in in union of X and Y. Data mining process using association rule is divide in to two steps: first step is the detection of all the frequent item set which means that all the item sets, whose supports are greather or more than minsupp, that is pre-determined minimum support threshold and second step constists using frequent item sets to generate strong association rules that is, finding frequent rules with confidence values greater than a minimum confidence threshold, minconf. For the rules which are strong and frequently used sensitivity levels are assigned along with confidence and support. If a strong, frequent rule is on top of a certain sensitivity level then the hiding process should be applied in such a way that either the strength of the rule or the frequency of the rule is decreased. The frequency of rule should be decreased below the min_supp and strength of the rule should be decreased below the min_conf. The problem of association rule hiding can be stated as follows: Given a small or large database D, a set of relevant rules denoted as R that is to be mined from D and a subset of those sensitive rules included in R,denoted by Rh, then if we want to retrive D' from database D then retriveal should be in in such a way that the rules in R can still be mined, except for the rules in Rh.

3. Proposed Algorithm

Three steps are discussed below to defien the general structure of the system.

First step is that global support and confidence is calculated using security sum as disscussed earlier. In this way participant is kept away from loocal support from each site.

Second step is to all the item sets are sent to different sites from the data center. SMC is used for this pupose. All this is done without any data leakage.

Third step is very important step in which sensitive rule database is build. This database contains all the rules that are to be hidden. In this step the system will hide the sensitive rules.

When the item sets are sent to different sites RSA encrupt algorithm can be used. Data miner will produce the center encrypt key denoted by C_{-e}. and the key pair is denoted as (e_i,d_i), where e_i is encrption key and d_i is decryption key. Data miner will send d_i to center and send ei and C-e to enach site. Each site will use the ID, e_i and C_{-e} to encyrypt their own frequent itmes thereby sending the result to center. In the decrypt step, each site send their data to data center. Data center will use decrypt the data from each site and remove the ID. Data center will change

the order of the data, and will send to data miner. Data miner will decrypt the dat, getting frequent item sets.

S.L. Wang[3] proposes two data mining algorithms using association rule for hiding sensitive information. Two algorithms are ISL ie Increase support of left hand side and DSR,ie decrease support for right hand side. ISL is used to increase support of left hand side of the rule while DSR is used to decrease support of right hand side of the rule. Using above methods the sensitive rules will be hidden but along with it some insensitive rule may also be hidden. It is possible that many new rules may also be produced artificially. To solve this problem, check the mining result for modification to chooose other item as sacrifice item, in order to get a better and good results. Privacy quantify is used to measure the result. If it is not satisfied return to change the choosing policy. Checking of entire database is done with their suport. Classify the items using their support values. Those items with support values more close to min_sup are classified as unsettled itemsets and the remaining are classified as settled itemsets. For blocking algorithm, the fact is that dataset apart from the blocked values is not distrorted, some of the noise rules are necessary, to make dataset distortion, which can be deleted in trim step. Flowchart for the algorithm discussed above is as follows.

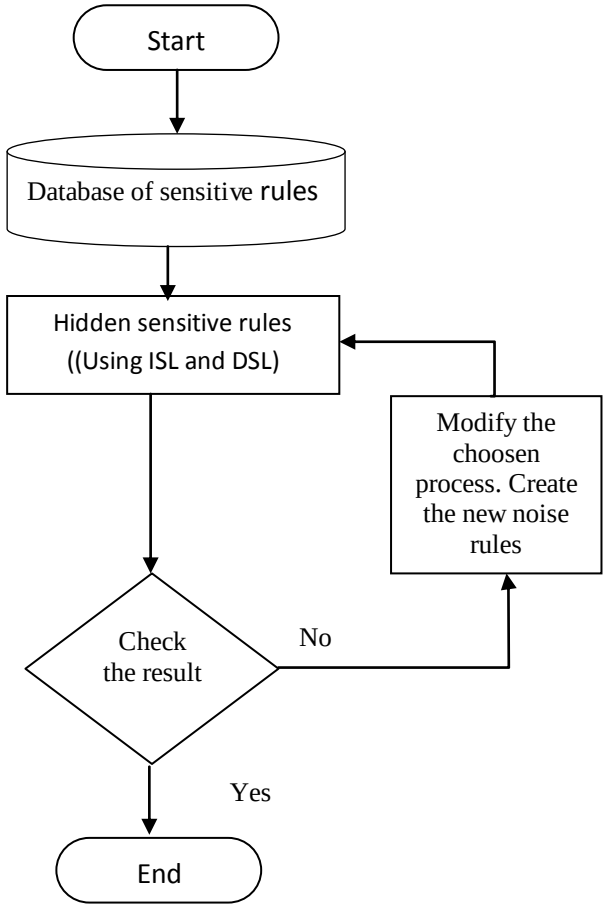

Figure 2 : Flowchart for Association rule hiding

The main goals of any privacy preserving association rule mining algorithm should insist on following factors : An algorithm using association rule mining for privacy preserving should prevent the finding of sensible information; The algorithm should not restrict the use and access of non sensitive data items/information; The algorithm should not have large computational complexity; It should be challenging to the various data mining techniques.; The algorithm should be equally efficient for very large database. This is very important factor

All mentioned algorithms does not satisfy all the goals only some of them satisfy these goals. Considering above mentioned goals ,the algorithms can be evaluated using

- Efficiency : The efficiency of algorithm is measured with its ability to execute with good performance using all required resources.
- Scalability : The algorithm should work with good performance even when storage requriment is very large along with communication cost s of the distributed system when data sizes is increased.
- data quality : If the data quality is not relevant, the knowledge extraction is of no use.
- hiding failure: Privacy preserving algorithms should be developed for zero hiding failure.
- and privacy level offered by algorithm : is the degree of uncertainty, according to which hidden sensitive data can still be calculated.

4. Conclusion and Future Research

In this paper, a methodology for evaluating privacy preserving association rule mining algorithms is proposed. To avoid the data leakage while sharing the data secure multiparty computation and RSA encryption technique is used. Sensitive rules can be hidden using DSR and ISL algorithms. To get good quality result some modification is also suggested. Drawback is that secure computation will cause high communication cost for huge database. The proposed evaluation methodologies can be applied in new set of privacy preservation like cryptography-based algorithms.

References

[1] Agrawal R., Srikant R, "Privacy Preserving data mining". *Proceedings of the ACM SIGMOD Conference, 2000*.

[2] S.L Wang and A. Jafari, " Hiding informative association rule sets." *Expert Systems with Applications 33 (2007)* pp 316-323

[3] Y Saygin, V. S. Verykios, and C. Clifton, "Using unknowns to prevent discovery of association rules." *ACM SIGMOD Record, 30(4) 2001* pp 45-54,

[4] A.C. Yao,"How to generate and exchange secrets", *Proceedings of the 27th IEEE Symposium on Foundations of Computer Science, 1986*, pp. 162-167.

[5] O. Goldreich, S. Micali, A. Wigderson,"How to play any mental game--a completeness theorem for protocols with honest majority", *Proceedings, 19th ACM Symposium on the Theory of Computing, 1987*, pp. 218-229.