Application of Association Rule Hiding on Privacy Preserving Data Mining

Mrs.D.Deva Hema, Karan Saxena, Saurabh Satpute, Aditya Gupta

Abstract— **Privacy preserving data mining is an area of research concerned with the issues of privacy thus providing a solution to minimize privacy threats in data mining. PPDM also helps in maximizing our analysis outcome and also helps in minimizing the disclosure of individual or organizational private data. The main objective of PPDM is to develop algorithms for modifying the original data in some way, so that even after the mining process the private data and private knowledge remains undisclosed. Association rule mining is an efficient data mining technique that recognizes the frequent items sets and associative rule for large set of transactional databases. Association rule hiding emerge as the best possible solutions for providing confidentiality and improving the performance of user. The process starts with the application of association rules across the data and thus leads to the creation of two sets of rules which are either strong or sensitive. Our aim is to minimize and maintain the privacy of each individual or organization sharing their details. One way of preserving the data is by hiding the sensitive rules generated earlier in the process so that no one other than the original handler of the data has access to these sensitive rules. Thus simultaneously all the sensitive rules are hidden and there are no side effects related to this technique rather it is technique which provides us with maximum data utility .Association rule hiding is a technique of generating sensitive association rule which are hidden within the data to prevent its unlawful disclosure.**

Index Terms—**Association rules, Association Rule Hiding, Data mining, Privacy preserving data mining (PPDM).**

I. Introduction

Data which belongs to a person or an organization contains different levels of sensitivity. Such data are made available only for authorized persons. Data mining operations are done in a scenario where two or more parties maintaining private database which contain sensitive information decide to cooperate by computing a data mining algorithm on the union of their databases. Since the parties involved try to keep their databases confidential, neither party will be ready to reveal any of their sensitive contents to the other. So to protect privileged information and enable its use for research and other purposes key algorithms are proposed.

Similar problems which are faced during data mining process can be solved using secure multi-party computation by known generic protocols. But firstly, to motivate the data provider to give unmasked data it is made sure that the loss in privacy faced by the provider is matched by giving useful information which would help the process of decision making easier and save time for the provider.

A. Association Rule Hiding

The association rule hiding technique is a process to remove the sensitive rules from the transactional database during the overall process of association rule mining.

B. Association Rule Mining

An association rule is an implication of the form XY, where X and Y are subsets of I and X∩Y= Ø. The support of rule XY can be computed by the following equation: Support(XY) = |XY| / |D|, where |XY| denotes the number of transactions in the database that contains the item set XY, and |D| denotes the number of the transactions in the database D. [1]

C. Data Mining

It is the computational process of discovering patterns in large data involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. [2]

D. Privacy preserving data mining

Privacy preserving data mining (PPDM) is a novel research direction in Data Mining (DM), where DM algorithms are analysed for the side-effects they incur in data privacy. The main objective of PPDM is to develop algorithms for modifying the original data in some way, so that the private data and private knowledge remain private even after the mining process. [3]

II. LITERATURE SURVEY

Based on the roles and permissions concept in the market, there are a number of existing systems on which we will take a brief look on. Here the two existing approaches of privacy preserving data mining are discussed. In the first one the aim is to preserve the underlying data without jeopardizing the similarity between two objects. The other approach is of the perturbation method where a random noise from a known distribution is added to the sensitive data and hence through this method another reconstructed distribution can be created using the perturbed data. The advantages and the disadvantages of these existing approach are being described.

A. Privacy Preserving Clustering By Data Transformation [4]

Clustering is mechanism where the data sets are partitioned into sub-classes. The challenge was to protect the underlying data values which are subject to cluster analysis without risking the similarity between objects. Privacy preserving clustering was designed to meet privacy requirements as well as guarantee valid clustering results.

Advantages

• The geometric data transformation methods (GDTMs) are used through which numerical attributes are suppressed so that secure cluster analysis takes place.

Mrs. D. Deva Hema, Assistant Professor, Department of CSE, SRM University Chennai, Tamil Nadu

Karan Saxena, B.Tech Student, Department Of CSE, SRM University Chennai, Tamil Nadu

Saurabh Satpute, B.Tech Student, Department of CSE, SRM University Chennai, Tamil Nadu

Aditya Gupta, B.Tech Student, Department Of CSE, SRM University Chennai, Tamil Nadu

- After the cluster analysis end users may use the tools of their own so that constraint on privacy can be applied before mining process by data transformation.

Disadvantages

- The data shared after privacy cluster analysis is very complex.
- Clustering the similarity between objects which are under analysis for privacy is hard to achieve.

B. *Perturbation Based Privacy Preserving Data Mining For Real World Data [5]*

Perturbation based method is a technique where random noise from a known distribution is added to the sensitive data before the data is sent for the mining process. Consequently, the data miner rebuilds an approximation to the original data distribution from the perturbed data and the reconstructed distribution is being used for data mining purposes. Individuals or organizations of different type may have different approaches towards privacy preserving. Unfortunately, recent privacy preserving data mining techniques based on perturbation do not allow the individuals with full freedom to choose the desired levels of privacy.

Advantages

- Efficient and simple data mining models can be made from perturbed data.
- The data miner could rebuild the original distribution using various statistical models and mine the rebuilt data once the distribution of added noise is known.

Disadvantages

- Individuals are unable to choose their desired privacy levels due to data mining techniques based on perturbation.
- In perturbation based approaches there is generally a trade off in information loss versus preservation of privacy after the addition of noise

III. ASSOCIATION RULE HIDING

The association rule hiding methodologies sanitizes the database in such a way so that at least one of the following goals is accomplished [6].

A. The rule which is considered as sensitive from the owner's perspective and cannot be mined from the database at pre-specified thresholds of support and confidence and also data can be revealed from the sanitized database, when the database is mined at the same or at higher thresholds.

B. All the rules that appeared non-sensitive during the mining the original database at pre-specified thresholds of support and confidence can be successfully mined successfully from the sanitized database at the same thresholds or higher.

C. The rules which were not derived from the original database when mined at pre-specified thresholds of confidence and support cannot be derived from its sanitized database when it is mined at the same or at higher threshold.

D. Association rule hiding is a process that is totally dependent on the support or confidence of the rule, there are two way to hide any rule, either decrease the support up to certain threshold or decrease confidence up to certain threshold, so the mining algorithm, that works on support will not able to mine sensitive rules with the same efficiency.

However with the modification performed on the database it may lead to some side effect that may results to disturbance in association rule mining, following are the side effects that may occur in the rule hiding process:

Lost Rules: these are the non-sensitive association rules which are present in original database and are mined by applying the mining algorithm but they cannot be mined after applying hiding algorithm from modified database.

False rules: these are the sensitive association rules which cannot be hidden by hiding algorithm and can be mined by applying mining algorithm on modified database.

Ghost rules: the rules which are not at all present in the original database but are generated after the application of hiding algorithm.

IV. ASSOCIATION RULE HIDING STRATEGIES

The association rule hiding strategies are classified into the following:

A. Heuristic based approach

This approach is further classified into two methods for sensitive association rules.

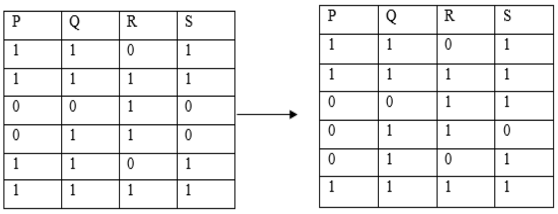
i. A technique based on distortion of data

This technique was first used in hiding association rules. Proof for optimal sanitization of NP hard problems can be derived by using this technique [7]. In this process the modification of the database matrix is used to hide the rules, which is done by replacing the values of some items in the database matrix by 0 or 1. This technique of data distortion contains two basic methods which are used for rule hiding.

In the first method rules are hidden by the process of decreasing the support of the rule to a level which is acceptable and on the second method the reduction of the confidence level is done up to a threshold which is agreed. Table 1 shows an example of distortion technique of basic data.

There are five different technique used for association rule hiding. They are based on either reducing support or decreasing the confidence level.

Table 1: Hiding P->Q by Distortion



P	Q	R	S
1	1	0	1
1	1	1	1
0	0	1	0
0	1	1	0
1	1	0	1
1	1	1	1

P	Q	R	S
1	1	0	1
1	1	1	1
0	0	1	1
0	1	1	0
0	1	0	1
1	1	1	1

ii .Technique based on data blocking

Under this process the sensitive association rules are hidden by data blocking technique. In this process the rules are hidden by altering the values of items present in the database matrix to an unknown value [*] from 0 or 1. Hence the support of some of the items present in the database matrix goes down from the existing level and the rule mining algorithm may not be able to accurately mine the sensitive rules. Table 2 shows an example of basic data blocking technique.

Table 2: Hiding P->Q by Blocking

P	Q	R	S
1	1	0	1
1	1	0	1
1	1	1	1
1	0	1	0
0	0	0	1
1	0	1	0

P	Q	R	S
1	1	0	1
1	1	0	1
*	1	1	1
1	0	1	0
0	0	0	1
1	0	*	1

B. An approach using border revision

In this approach the database items are classified into frequently and infrequently used item sets based on which the borders are modified in the lattice to hide the sensitively associated rules [8]. This process finds the items sets which are non-sensitive but frequently used and applies modifications based on the order in which the items sets are tracked and which has minimal impact on the quality to accommodate the hiding of sensitive rules. There are numerous border revision approach based algorithms like Border Based Approach, Min-Max1 and Min-max2 to hide sensitive association rules. These Algorithms uses methods such as deleting a specific sensitive item and reduces the loss in items sets which are non-sensitively associated while storing it in a sanitizes database for protecting sensitive rules.

C. The exact approach

The exact algorithm is an approach based on heuristic function which formulates the process of hiding as a constraint satisfaction problem. This approach is a descendant of the border based methodology and is solved by integer programming.

V. PROPOSED TECHNOLOGY

The PPDM is a process to safeguard sensitive information from unsolicited or unsanctioned disclosure and preserve the utility of data. Our aim is to overcome these privacy issues in data mining using certain privacy preserving data techniques. Association rule hiding emerge as the best possible solutions. We will generate strong associations based on the data which is present in the database. Apart from the strong associations there will be some data which have formed weak associations and our focus mainly relies on hiding those data exhibiting weak associations as they can lead to privacy issues. The aim of a secure multiparty computation task is for the participating parties to securely compute some function of their distributed and private inputs. Each party learns nothing about other parties except its input and the final result of data mining algorithm.

A.Mathematical Model

Association rules by the use of support and confidence can be defined as follows. Let I= {I1, I2,…….Im} be a set of items. Let DB= {Z1,Z2,……...Zn} be a set of transactions. where each transaction Z in DB is a set of items such that $Z \subseteq I$ an association rules of implication in the form of $X \rightarrow Y$, where $X \subset Z$, $Y \subset Z$ and $X \cap Y = \emptyset$.

$$\text{Support } (A\text{->}B) = (A \text{ U } B)/n \geq \text{min \_ support}$$

$$\text{Confidence } (A\text{->}B) = \text{Support } (A \text{ U } B) / \text{ Support } (A) \geq \text{min \_ confidence}$$

Let DB’ be the database after applying a sequence of modification to DB. A strong rule $X \rightarrow Y$ in DB will be hidden in DB’ if one of the following condition holds in DB’.

1. Supp $X \cup Y < MS$
2. Confidence $X \rightarrow Y < MC$

VI. SYSTEM ARCHITECTURE

This figure shows the architecture of our overall process which deals with the strategy of association rule hiding. Association rule hiding is a strategy which comes under privacy preserving data mining technique whose aim is to prevent the unlawful disclosure of its data. In this process, the administrator has full access to its database. The changes in the database can be made only by the administrator. The administrator can make changes or access the database only using conditional queries. Once the database has been created the data further goes through a process of data pre-processing which deals with the data selection (to retrieve data from database relevant to us). The data pre-processing consists of four steps.

- Step 1: Data cleaning: to remove noise and inconsistent data, to handle the missing data fields, etc. and data integration (to combine data from multiple sources).
- Step 2: Data integration: to combine data from multiple sources.
- Step 3: Data transformation: transforms the data into forms appropriate for data mining task and also finds useful features in it to represent the data.
- Step 4: Data reduction: the process of removing irrelevant features in the data so as to reduce space and also make it efficient for data mining purpose.

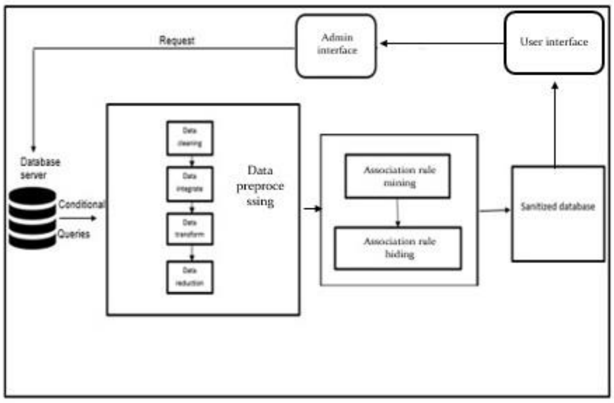


Fig.1 Architecture of Association Rule Hiding

Once the data has been passed through Data pre-processing steps it has transformed into the format of data mining. Association rule mining helps in finding interesting relationships between the data using certain associations between the respective data. Association rules using support and confidence as defined in our mathematical model give us only the strong associations on the data. As the strong associations on the data has been found out there will certain data which are weakly associated with each other. Thus the idea is to use Association rule hiding to those sensitive rules and sensitive items which help in hiding the pattern which is found in the sensitive data which are weakly associated with each other, hence providing a better result since the datasets are not suppressed or falsified much in the process. The rules generated tries to modify transaction until the confidence of the rule fall below minimum confidence or minimum support level agreed by both the parties. The changes to the datasets are done by either removing items from the transaction or inserting falsified new items to the transactions. The process is further improved by restricting patterns found in the data by generating association rules at the time of transaction.

The rule mining process first selects the transactions that contain the intersecting patterns from a group of restricted patterns and depending on the decided threshold value given by users the algorithm sanitizes a percentage of the selected transactions and stores it in a sanitized database in order to hide the restricted patterns. Then the rule hiding approach helps in modifying the database such that confidence level of the association rule can be altered by increase or decrease the support value based on the relationship between the data exchanged between the parties. If the confidence of the rule falls below a specified threshold on a particular side the data is hidden or not disclosed. The association rule hiding discussed is not applied to data which are strongly associated or data which do not contain sensitive information and such numeric or symbolic attributes.

VII. PROPOSED ALGORITHM

A. Design of association rule hiding algorithm

The main objective of association rule hiding algorithm is to hide certain confidential data so that they are not discovered through the data mining techniques. In this research work, it is assumed that only sensitive items are given and one proposed algorithm is used to modify data in database so that sensitive items cannot be deduced through association rule mining algorithms. More specifically, given a transaction database DB, a Min_supp(minimum support), a Min_confid (minimum confidence) and a set of hidden items H , the objective is to modify the database DB such that no association rules containing H will be discovered on the right hand side or left hand side.

The two proposed association rule hiding algorithms are used namely ISL (Increase Support of Left hand side) and DSR (Decrease Support of Right hand side) so as to hide useful association rule from transaction data with the use of binary attributes. In ISL method, confidence of a rule is further decreased by increasing the support value of Left Hand Side (LHS) of the rule.

Based on these two concepts, a new association rule hiding algorithm for hiding sensitive items in association rules has been proposed. In the proposed algorithm, a rule $X \rightarrow Y$ is hidden by decreasing the support values of $X \cup Y$ and increasing the support value of H which can increase and decrease the support values of the LHS and RHS item of the rule correspondingly.

This algorithm tries to hide the rules in which item to be hidden i.e., H is in right hand side and then tries to hide the rules in which H is in left hand side. For this algorithm t is a transaction, T is a set of transactions, R is used for rule, RHS (R) is Right Side of rule R, LHS (R) is the Left Side of the rule R, Confid (R) is the confidence of the rule R, a set of items H which are to be hidden.

INPUT: A source database DB, A minimum support min_supp (MS), a minimum confidence min_confid (MC), a set of hidden items H, total transaction Z, first transaction from total z.

OUTPUT: The sanitized database DB, where rules containing X on LHS or RHS will be hidden.

```
1. Begin
2. Generate all possible rule from given items Z;
3. Compute confidence for all the rules hidden item H, compute confidence of each rule R.
4. For each rule R in which H is in RHS
        4.1 If confidence (R) < MC, then
 Go to next 2-itemset;
   Else go to step 5
5. Decrease Support for RHS item H.
5.1 Find Z=z in DB fully support for R;
5.2 While (Z is not empty)
 5.3 Choose the first transaction z from Z;
5.4 Modify z by putting 0 instead of 1 for the RHS item;
5.5 Remove and save the first transaction z from Z; End While
6. Compute the confidence of R;
```

VIII. CONCLUSION

In this paper, a wide survey is conducted for different approaches of privacy preserving data mining, and analyses the major algorithms available for each method and points out the existing drawback. While all the purposed methods are only approximate to our goal of privacy preservation.

To address this issue, we advise that the following problems should be widely studied:

(A) Privacy and accuracy is a pair of contradiction; improving one usually incurs a cost in the other. How to apply various optimizations to achieve a trade-off should be deeply researched.

(B) In distributed privacy preserving data mining areas. We are tried to develop more efficient algorithms and achieve a balance between disclosure cost, computation cost and communication cost.

REFERENCES

[1] Neelkamal Upadhyay,KuldeepTripathi and Prof. Ashish Mishra, “ A Survey of Association Rule Hiding Approaches”, IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555 Vol. 5, No1, February 2015.

[2] https://en.wikipedia.org/wiki/Data_mining

[3] Alberto Trombetta and Wei Jiang (2011), ‘Privacy-Preserving Updates to Anonymous and Confidential Databases’, IEEE Transactions on Knowledge and Data Engineering, Vol. 22, pp. 578-568.

[4] Stanley R. M. Oliveira, and Osmar R. Zaiane, "Revisiting Privacy Preserving Clustering by Data Transformation," Journal of Information and Data Management, vol. 1, no. 1, 2010.

[5] Li Liu, Murat Kantarcioglu and BhavaniThuraisingham, "The applicability of the perturbation based privacy preserving data mining for real-world data," Journal of Data & Knowledge Engineering, vol. 65, pp. 5–21, 2008.

[6] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. S.Verykios “Disclosure limitation of sensitive rules.”In Proc. of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX’99), pp. 45–52, 1999.

[7] V.S. Verykios, A. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni, “Association rule hiding,” IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No.4, 434–447, 2004

[8] A. Gkoulalas-Divanis, and V. S. Verykios, “ Exact knowledge hiding through database extension” IEEE Trans Knowledge Data Eng 2009, pp. 699–713