IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 5, Ver. II (Sep – Oct. 2014), PP 25-31
www.iosrjournals.org

Privacy Preserving by Hiding Association Rule Mining from Transaction Database

Mr. Pravin R. Ponde[1] and Dr. S. M. Jagade (Ph.D) [2]
[1]M.E, Department of Computer Science and Engineering,
TPCT's College of Engineering, Osmanabad, Maharashtra, India
[2]Principal, TPCT's College of Engineering, Osmanabad, Maharashtra, India

Abstract: *For making the decision of data mining process some expertise are required, some organization have their own expertise, but many organization doesn't have their own expertise, so the organization helps with some external advisor for the process of data mining. But risk is occurred at the time of getting advice from the external advisor; the question arises regarding the privacy of the customer data and loss of business intelligence. The Security and Privacy of the data are main challenging issues. The owner of the data has some private property like the outsourced database which contains the association rules. However, if the service provider is not trustworthy then integrity of mining results can affect badly. The proposed scheme for privacy preserving mining on databases to protect association rule means the corporate privacy. As per our study, in our paper we are proposing the heuristic based algorithm for hiding the sensitive association rules the algorithm is named as MDSRRC , owner hide sensitive association rule and place transform rules to the server for outsourcing purpose. In this algorithm we are providing an incremental association rule for mining. The recent study concludes that the problem of the incremental association rule mining task's importance was observed, when data is updated. The Matrix Apriori algorithm is proposed which is based on analysis of two association algorithm named as Apriori algorithm and FP-growth algorithm. The matrix Apriori algorithm has a simple structure similar as a matrices and vectors, the algorithm generates frequent patterns and minimizes the number of sets, as compared to previous algorithm. The matrix algorithm is simple and efficient way to generate association rule than the previous algorithm. For hiding the sensitive information of the database proposed algorithm MDSSRC selects the transactions and items by using certain criteria which transform. As per comparing with the previous algorithm the proposed algorithm is much better in performance which can be concluded with the results of the implementation.*

Keywords: *MDSRRC, AES, Hiding association rule, privacy preserving policy***.**

I. Introduction

All the corporate as well as government sector used the concept of data mining. We know that the concept of data mining is used for preserving privacy to the data, reveals their data or information for manual benefit for searching out some useful data for some decision making purpose and for improving their database scheme. Any organization needs security regarding the confidential database of the organization which can be used at the time of revealing data or information for the use of manual work. At the time of sharing data for external work the private data also visible to the user which can affect the security and privacy of the confidential data, for resolving this issue that is before revealing the data sensitive pattern should which the organization doesn't want to disclose the technique is used named as PPDM. The technique is used for safety of database or information. The importance of hiding the sensitive pattern can be explained with the help of following example: let a grocery mall that purchase shampoo from two companies PQR and XYZ and let both the companies can access customer's data store. If now PQR applies data mining techniques and mines association rules related to XYZ's products. PQR has found that almost all customers who buy XYZ's shampoo also buy conditioner. So no PQR offers some discount on purchases of conditioner if they buy PQR's shampoo. As a result the business of XYZ‘s goes down. So providing access to sensitive information along with the database caused the problem.

These PPDM approaches have in general the advantage to require a minimum amount of input and then a low effort is required to the user in order to apply them. For performing association rules for hiding databases there are two approaches. The first one is transaction and second one is about the concept of restriction pattern. In the first approach of transaction, it is used for hiding a rule at a one time. For performing this operation the steps are like first the transaction is selected on the basis of item in a given rule. After that try to modify transaction one after the another, it can be continue till the confidence of the rule is below the minimum level, the cant be support for transaction. The second approach proposed to work for hiding sensitive data or information, for this we concern with hiding association rule. The sensitive information present in the both sides of the rule, i.e. right or the left hand side. For this reason the rule contains the confidential data which can’t be disclosed or open by anyone. The main of this paper is to modify the database by using association rules does this with increasing or decreasing the value of both sides of the hiding rule i.e. right or left hand side rule.

The proposed work represents the working of two algorithms which are Matrix Apriori algorithm and MDSRRC algorithm. The Matrix Apriori algorithm gets some properties from the previous algorithm for generating the association rule. The matrix algorithm gives simple and efficient process for association rule. Second one is MDSRRC algorithm this algorithm use for hiding the sensitive rules which are generated in the database. The sanitized database is generated by the minimum confidence threshold and minimum support threshold algorithm. the sanitized database which are generated from this algorithm successful to hide all sensitive rules. This database maintains the quality of the database.

In this paper, knowledge about the existing work is mentioned in the literature survey section, in Section II. Some background details and formulation is mentioned in Section no III. Proposed system is mention in section IV. Experimental result is in Section V. final Conclusion and the future work is mentioned in the Section VI.

II. Literature Survey

Methods designed or implemented for association rule mining is as follows:

1. Heuristic Based Approaches:

They are divided into two techniques:

A. Data Distortion Technique

In this type we replace the values from 1 to 0 or 0 to 1. Again this has two basic approaches for rule hiding. First it reduces the support of rules and second reduces the confidence of rules. Verykios et al. studied this concept and implement and proposed new five algorithms, this algorithms used for hiding the sensitive knowledge of database, this can be possible by reducing the support or confidence of the sensitive rules. Hiding of association rues is done by first three algorithms and hiding of large item sets are related to algorithms 2.b and 2.c. Oliveira was responsible for improving the balance between the two, sensitive knowledge and discovered pattern, which provided better privacy. A method that reduced the side effects on sanitized database introduced by Y H Wu in this method two algorithms are described, in the first if the item is present in the left side then algorithm increases the support of the sensitive item. In the second algorithm if the sensitive item present in the right hand side then algorithm decreases the support of the sensitive item.

1) Pros:
- It can be scaled to very large data sets.
- It is easy to utilized the new distance function rather than matching the old one.

2) Cons:
- The main problem we are facing is on with transactional sets of binary data, flipping entry values.
- These values have the number of side effects on the non-sensitive rules.

B. Data Blocking Technique:

First Y. Saygin in purpose blocking technique in order to decrease or increase items support, replaced 0’s or 1’s by the sign “?”. So it is difficult for anyone to finding the value which is stored behind the “?”. This technique provides some privacy, the more efficient approaches were proposed by Wang and Jafari. At the time of hiding many rules at a time, they require few numbers of scans for databases and cut more number of rules.

1) Pros
- One of the best attractive things about the blocking approach is that it maintains the truthfulness of the data.
- With the use of hiding process it reduce the disclosure of sensitive entries.

2) Cons
- For the non-sensitive rule they have the number of side effects.
- This approach is restricted up to low dimensional data sets and binary based dataset.

2. Border based Approaches:

Border based approach was proposed by Sun and Yu, this approach used to modify the borders of the original database, it modify the lattice of the frequent and infrequent item sets. Hides sensitive association rules by a border are formed to separating the infrequent and frequent item sets. In this approach by using border value of non sensitive item, separation of positive and negative value of the border from the item set. Then the value which is minimum affected is selected. For minimum side effect purpose modification is done by greedy selection.

In this approach by using the opposite values, i.e. by using opposite negative and positive borders values of the database, after that try to cut down the item set with sensitive data and with the negative border. In this approach the positive border value with highest support and maximum distance from the border is selected.

1) Pros:

- From the result database the quality of database can be well maintained by the controlling modification.
- For selecting the modification with the minimum side effect is one of the application of the border-based approached.

2) Cons :

Border based approach used to separated data along border have bad support.

3. Cryptographic based techniques:

Most of the times multiple organizations want to share the private data but without losing their sensitive data.

a. Vertically partitioned distributed data

For the secure calculation of the idea of secure sum is the use of this technique. This technique also includes secure calculation of the union set and size of the scalar product and the interactions sets. These techniques used the vertical and horizontal partitioning technique. This technique describes the use of scalar dot product, for counting the frequent item sets.

b. Horizontal partitioned distributed data.

Finds global frequent item sets while ensuring no loss of inter-site information. It calculates support degree inter-sites secure sum.

4. Recent work

a. Ling Qiu for outsourcing association rule mining at the time of protecting BI and the privacy of the customer the approached is proposed. They proposed Bloom filter based approach. It can outsource the mining task for protecting business intelligence and the customer data privacy, and simultaneously maintain the result for precision mining which save storage space requirement without any running time.

b. Mohammad A. Ouda represents the PPDM method for horizontally partitioned of the data. The proposed algorithm used RSA encryption and homomorphism technology which is same time secured. No any global computation carrying the data at the centralized site but the algorithm named as KNN has need to be conduct locally for every site.

c. C N Modi proposed an algorithm named as DSRRC. This algorithm maintain privacy and quality of database. This algorithm used to improve the quality of database.

d. Laks V. S. Lakshmanan proposed the model for association rule for privacy preserving from the outsourced Database Transaction. This method solves the problem for preserving the mining of frequent pattern on an encrypted outsourced transaction database placed at cloud. Where they assume a traditional model from which the advisor knows the exact frequency of the item and the domain of the item that where it is located. For identifying the cipher items they can used this knowledge.

III. Background And Problem Formulation

There are some important concepts which are used for the designing or implementing the MDSRRC algorithm, the concept is as follows:

1. Item Sensitivity: the item sensitivity was used to measure rule sensitivity. The item sensitivity is defined as the number of frequency of the data items present in the number of sensitive association rule.
2. Rule Sensitivity: the rule sensitivity is defined as the total of all items which contain the association rule.
3. Cluster Sensitivity: cluster sensitivity is defined as the association rules present in the cluster, the sum of all the sensitive association rule.
4. Sensitive Transaction: the transaction of the item which contains the sensitive items is called sensitive transaction.
5. Transaction sensitivity: it is the total of all sensitivities of the intensive item which are present in the transaction.

By using the MST (minimum support threshold) and the MCT (minimum confidence threshold) the given algorithm is implemented, first the algorithm generates the number of association rule from the database D. with the help of database owner some generated association rules are selected as the sensitive rule set. The rule which contain only single right hand side are specified as a sensitive. After that the C cluster based on the right hand side item is calculated. After that the sorting of all cluster in the descending order is done. Sorting is depending upon the decreasing order of their sensitivities.

For converting the original database into sanitized database so that it is not possible using data mining technique for mining the sensitive rules from the original database while all non sensitive rules remain visible. The whole process is called as the association rule activity problem. Let's take one example for explaining this definition. The transactional database with D, with minimum confidence, minimum support, and generated set of association rule R from D, the owner of the database want to hide some data from SR which is a subset of R. For this we want to create sanitized database D', such that when mining technique applied to the sanitized database D, all sensitive rules in set the SR will be hidden while all non sensitive rules can be mined. There are some conditions which satisfy the aim of association rule hiding. The conditions are:

1. Any sensitive rule must not be disclosed by database
2. The sanitized database must facilitate the mining of all non-sensitive rules
3. It only generates the rules which are present in the database.

With some modification on the database for maintaining data quality and for reducing the side effect of database the proposed algorithm named as MDSRRC is implemented.

IV. Proposed Approach

The implementation of this method is divided into five different modules. The modules are Binarization, Matrix Apriori algorithm, sensitive rule generation, MDSRRC algorithm and creation of sanitized database. The general architecture of method is shown in Fig 1

T	a	b	c	d	E
1	1	1	1	0	0
2	0	1	1	1	1
3	1	0	1	1	0
4	0	0	0	1	1

Table.1 Binary Table.

In the above table, let a, b, c, d, e are the items in the transactional dataset. T is the number of transactional. During the process of the Binarization we get the output as shown in the following table. The output is called binarized output dataset.

A. Matrix Apriori Algorithm

Apriori algorithm is a classic algorithm basically this algorithm is used in the data mining for learning the association rules. Matrix Apriori algorithm is the result of analysis of basic two association algorithm which are the result of analysis of two algorithms named as Apriori algorithm and FP-growth algorithm. The algorithm has the successfully generates the frequent patterns and after that from the generated patterns it created the association rules this are the two steps of the matrix Apriori algorithm. the algorithm performs same as the Apriori algorithm in the second step but different in the first step.

The basic definition of association rule us that learning about association rules means finding the items that are purchased together in comparison to others. In our proposed work, Matrix Apriori algorithm is implemented for hiding the association rules. The Matrix Apriori algorithm is mainly use for generating data association rules. The generation of association rule is generally categorized into two steps. The steps are:

I. Generating the frequent item sets from a dataset the minimum support is applied.
II. In these step, for the mining association rules the frequent item sets from set 1 and the minimum confidence constraints is required

While the second step is straight forward, the first step needs more attention. It is difficult to finding all frequent pattern in a database hence it involves finding all item sets.

- **System Architecture:**

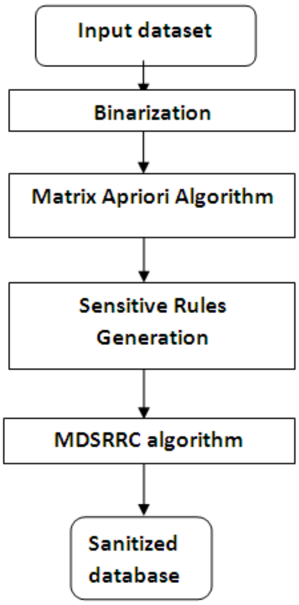


Fig1 .General Architecture

The sorting of the transaction is done in decreasing order of their sensitivity, only if transaction has the value is0. Selecting the first transaction form the sorted transaction with higher sensitivity, deleted item is0 from the transaction it is the process of initialization of rule hiding. After that all the sensitive rule which contain support and confidence update it. If any rule is remaining and it has below the MST and MCT respectively then delete it from SR .Continue this process by selecting transaction with higher sensitivity and deleting is0 from it. When all sensitive rule is hidden this process is terminated, means this process is continue until the entire sensitive rule is hidden. The sanitized database is generated by updating, modifying updated transaction into new database. Sanitized database D' preserves the privacy of sensitive information and maintains database quality.

B. Sanitized Database Generation:

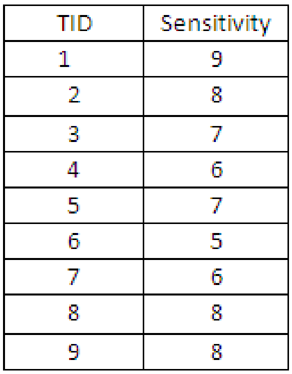
The possible generated association rules by Matrix Apriori algorithm are as follows: Let the database owner specify rule a →bd, a→cd and d→ac as sensitive rules. Then select transaction with the highest sensitivity and delete is0 item from that transaction. Update confidence and support of all the sensitive rules. Sort transactions which support is0, and delete the is0 from transaction with highest sensitivity, then delete the is0 from transaction with highest sensitivity. Finally all the sensitive rules are hidden.

V. Experimental Result With Example

We have tested this application on standard transactional database set which is shown below in Table II. The Binarization technique is applied on it which is shown in table III. IN table III transaction with its sensitivity is shown. In table IV sanitized database is generated with all its input item, after first deletion of item from its first transaction. Table V final sanitize database, with all sensitive rules are hidden.

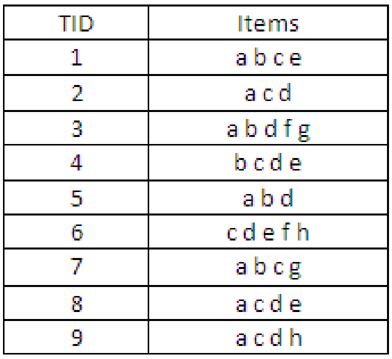
TID	Items	Binary matrix of item
1	a b c d e	1 1 1 1 1 0 0 0
2	a c d	1 0 1 1 0 0 0 0
3	a b d f g	1 1 0 1 0 1 1 0
4	b c d e	0 1 1 1 1 0 0 0
5	a b d	1 1 0 1 0 0 0 0
6	c d e f h	0 0 1 1 1 1 0 1
7	a b c g	1 1 1 0 0 0 1 0
8	a c d e	1 0 1 1 1 0 0 0
9	a c d h	1 0 1 1 0 0 0 1

Table.2 Transactional Database



TID	Sensitivity
1	9
2	8
3	7
4	6
5	7
6	5
7	6
8	8
9	8

Table.3 Transaction with sensitivity



TID	Items
1	a b c e
2	a c d
3	a b d f g
4	b c d e
5	a b d
6	c d e f h
7	a b c g
8	a c d e
9	a c d h

Table.4 Sanitized Database D1

TID	Items
1	a b c e
2	a d
3	a b d f g
4	b c d e
5	a b d
6	c d e f h
7	a b c g
8	a c d e
9	a c d h

Table.5 Finalized Database

Here we are computing the performance of MDSRRC algorithm with Matrix Apriori algorithm. We used algorithm MDSRRC and Matrix Apriori algorithm, for hiding the three rules of sensitive on sample database, as shown in Table 1. After applying algorithm with 3 as MST and 40% as MCT, we select 3 rules as sensitive rules from generated rules. After applying both algorithms on sample database we have done evaluation by considering the performance parameter. MDSRRC increases efficiency and reduce modification of transaction in database. Performance comparison of Matrix Apriori algorithm with Apriori algorithm is shown below. (Here X-axis represent's Support and Y-axis represent's time in second).

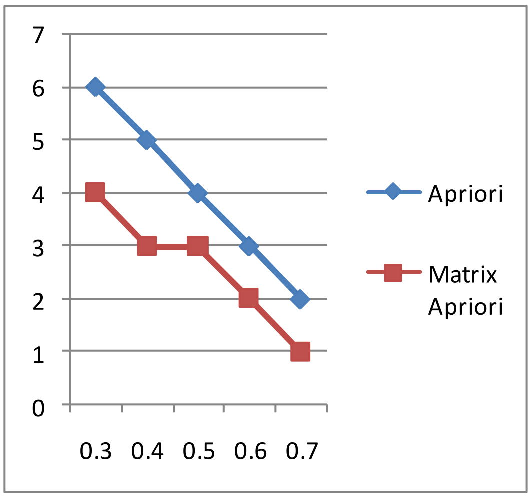


fig2. Performance Graph

Parameter (Support)	Apriori Algorithm	Matrix Apriori Algorithm
0.3	6 Sec	4 Sec
0.4	5 Sec	3 Sec
0.5	4 Sec	3 Sec
0.6	3 Sec	2 Sec
0.7	2 Sec	1 Sec

Table.6 Performance table

VI. Conclusion

The techniques named as association rule hiding technique which was proposed for hiding the sensitive data or information. The man aim of this paper is to propose this technique. For the implementation we proposed algorithm named as MDSRRC. Also we proposed an algorithm for generating the association rule named as Matrix Apriori algorithm. The MDSRRC algorithm hides sensitive association rules with the modification on database for maintaining transaction database quality and the side effect on database reducing. In this model we outsourcing database on server or any service provider and security of sensitive data maintained by encryption policy. The proposed algorithm Matrix Apriori provides an incremental approach for association rule mining. The Matrix Apriori algorithm maintained the efficiency same as the previous algorithm. After implementing the proposed algorithm by taking number of example it conclude that the proposed Matrix Apriori algorithm improves the speed of the mining process than the previously defined algorithm. We had to improve the proposed algorithm in the future like algorithm can help to reduced side effect of modification on datasets also increases the efficiency. In this paper we also discussed about the privacy preserving technique.

References

[1]. X. Sun and P.S. Yu “A Border-Based approach for hiding the frequent item sets” In Proc. Fifth IEEE Int’I conf. data mining (ICDM ‘05), pp. 426-433 Nov 2005.

[2]. V. Verkios and A. Gkoulalas- Divanis, A Survey of association rule hiding method for privacy, ser. Advance in database systems. Springer US, 2008, vol. 34.

[3]. Charu C. Aggrawal, Philip S. Yu, privacy preserving data mining models and algorithm. springer publishing company incorporated, 2008, pp. 267-286.

[4]. Y. Guo, ‘Reconstruction based association rule hiding’, in proc. Of SIG<OD2007 Ph.D. Workshop on innovative database research 2007(IDA2007), 2007.

[5]. J. vaidya and C. Clifton, ‘privacy preserving association rule mining in vertically partitioned data’, In proc. Int’I Conf data mining pp. 639-644 july 2002.

[6]. M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim and V. S. Verkios ‘disclosure limitation of sensitive rules’, in proceedings of the 1999 IEEE knowledge and data engineering exchange workshop (KDEX), pp. 45-52, 1999.

[7]. Han Jiawei and Kamber, Micheline. ‘data mining concepts and tchniques’ 2006. Morgon Kaufmann sanfransisco, C.A.

[8]. Z. Zheng, R. Kohavi, L. Mason, real world performance of association rule algorithms. In procedding of the seventh ACM-SIGKDD International Conference on Knowledge Discovery and Data Mining, 2001, 401-406.

[9]. K. Wang, Y. He, J. Han, Pushing Support Constraints In: Association Rule Mining. IEEE Transactions on Knowledge and Data Engineering

AUTHORS

Mr. Pravin R. Ponde, M.E, Department of Computer Science and Engineering,
TPCT’s College of Engineering, Osmanabad, Maharashtra, India

Dr. S. M. Jagade, Ph.D, Principal, TPCT’s College of Engineering,
Osmanabad, Maharashtra, India.