RESEARCH ARTICLE OPEN ACCESS

Hiding Sensitive items using MDSRRC to Maintain Privacy in Database

Pratiksha Sapkal, Minakshi Panchal, Manisha Pol, Madhumita Mane
[1,2,3,4,](Department of Information Technology, Pune University, India)

ABSTRACT
This paper focuses on the research in hiding sensitive association rules to maintain privacy and data quality in database. In this paper we have proposed heuristic based algorithm named MDSRRC (Modified Decrease Support of R.H.S. item of Rule Clusters) to hide the sensitive association rules with multiple items in consequent (R.H.S) and antecedent (L.H.S). The algorithm selects the items and transactions based on certain criteria which modify transactions to hide the sensitive information. Our evaluation shows that the proposed technique is highly valuable and maintains dataset quality.
Keywords-Association rules, Privacy preserving data mining (PPDM), MDSRRC

I. INTRODUCTION

Mining association rule techniques are wide employed in data mining to and relationship between item sets. The corporate and many government organizations reveals their data or information for mutual benefit to search out some useful data for some decision making purpose and improve their business schemes. But this database may contain some confidential information and which the organization does not need to reveal.

The problem of concealment plays important role once the corporate share their data for mutual profit however there is no one need to leak their private data. So before revealing the information, sensitive patterns should be hidden and to resolve this issue PPDM (Privacy preserving data mining) techniques are helpful to boost the safety of database. These approaches have in general the advantage to require a minimum amount of input (usually the database, the information to protect and few other parameters) and then a low effort is required to the user in order to apply them. The selection of rules would require data mining process to beexecuted first. For association rules hiding, two basic approaches have been proposed. The first approach hides one rule at a time.

First selects transactions that contain the items in a give rule. It then tries to modify the transaction by transaction until the confidence or support of the rule fall below minimum confidence or minimum support. The modification is done by either removing the items from the transaction or inserting new items to the transactions. The second approach deals with groups of restricted patterns or association rules at a time. In our work we are concern of hiding certain association rules which contain some sensitive information which are on the Right hand side or left hand side of the rule, so that rules containing sensitive item can't be reveal. Our approached is based on modifying the database in a way that confidence of the association rule can bereduce with the help increase or decrease the support value of R.H.S or L.H.S correspondingly. As the confidence of the rule is reduce below a given threshold, it is hidden or we can say it will not be disclosed.

DSRRC is not able to hide association rules with multiple items in the antecedent (L.H.S) and consequent (R.H.S). To overcome this limitation, the work proposed by Nikunj H. Domadiya and Udai Pratap Rao elt al. [1] is the improved version of DSRRC i.e. MDSRRC, which uses count of things in consequent of the sensitive rules. It modifies the minimum number of transactions to cover most sensitive rules and maintain data quality.

II. RELATED WORK

There have been several methods proposed for hiding sensitive patterns in dataset. In 1999, M. Attalah and E. bernito[13] proposed the idea of Disclosure limitation of sensitive rules. It discusses security risks on the database when reveals it in public. They introduce algorithm for hiding sensitive items with little impact on database. V.S. Verykios et al, A. K. Elmagarmid et al. [12] present five different algorithm to hide the sensitive rules. These algorithm use hiding strategies which are based on decrease support and confidence of the sensitive rule. Furthermore, C. N. Modi and U. P. Rao et al.[3], presented the algorithm for Maintaining privacy and data quality in privacy preserving association rule mining. It improves the quality of DSRRC. Next, Motivation example shows importance of sensitive patterns in business applications.

Let an Mobile store purchase mobiles from two companies X and Y. Now X applies data mining techniques and mines association rules applied to related to B's product. X found that most of the customers who buy mobile of Y also buy camera. Now X offers some discount on camera if customer purchases X's mobile. As a result the business of Y goes down. Therefore discharging the database with sensitive information cause the problem. This scheme provides the order on sensitive rules hiding in the database.

The proposed algorithm customizes least possible number of transactions to hide supreme sensitive rules and preserving data quality.

III. PROBLEM STATEMENT

Association rule activity problem is defined as: converting the original database into sanitized database so that data mining techniques will not be ready to mine sensitive rules from the database while all non sensitive rules remain visible. Given transactional database D,Minimum confidence, Minimum support, and generated set of association rules R from D, a subset SR of R as sensitive rules, which database owner want to hide. Problem is to and the sanitized database D' such that when mining technique is applied on the D', all sensitive rules in set SR will be hidden while all non sensitive rules can be mined. The aim of association rule hiding is to satisfy the following conditions:

- Database must not disclose any sensitive rules.
- Sanitized database must facilitate mining of all non-sensitive rules.
- It must not generate any new rules which are not present in the database.

The Proposed algorithm named MDSRRC hides sensitive association rules with fewer modifications on database to maintain data quality and to reduce the side effect of database. In this paper we have apply the MDSRRC algorithm for the Gojee transaction dataset

IV. METHODOLOGY

The proposed methodology is categorized into five major modules viz; Binarization, Apriori, Sensitive rules generation, Rule Hiding and sanitized database creation.

The Fig1 shows general architecture of the proposed method.

- Binarization

We have integrated a pre-binarization step in order to enhance the input dataset quality. Let put 1 to the items present in the transaction and 0 for the items in the transaction.

- Applying Apriori

This algorithm starts with mining the association rule from the original dataset using association rule mining algorithm. Apriori is an influential algorithm for generating association rules. Let $I=\{i_1 \ldots i_n\}$ be distinct literals called items. Given a database $D=\{T_1.....T_m\}$is a set of transaction where each transaction T is a set of items as $T \subset I(1 \leq i \leq m)$.The association rule is define as $X \rightarrow Y$,where $Y \subset I$,$X \subset I$ and $Y \cap X = \phi$.X is called rule's antecedent(L.H.S) and Y is called rule's consequent (R.H.S).Association rules generation is usually split up into two separate steps.

1. Minimum support is applied to find frequentitems sets in a dataset. The support of rule $X \rightarrow Y$ is calculated using the following formula: $Support(X \rightarrow Y) = |X \cup Y|/|D|$, where |D| define the total number of the transactions in the database D and $|X \cup Y|$ is the number of transactions which support item set XY. A rule $X \rightarrow Y$ is mined from database if support $(X \rightarrow Y) \geq$ MST (minimum support threshold).

2. Secondly, comparing those frequent item sets with the minimum confidence constraints to form mining association rules. The confidence of rule iscalculated using following formula: Confidence $(X \rightarrow Y) = |X \cup Y| / |X|$, where |X| is number of transactions which supportitem set X. A rule $X \rightarrow Y$ is mined from database if confidence$(X \rightarrow Y) \geq$MCT.

3. While the second step is straight forward, the first step needs more attention. Finding all frequent item sets in a database is difficult since it involves finding all item sets. The standard definition of association rule is given in [4].

- Sensitive Rule Generation :

The user specifies some rules as sensitivity as sensitive rules (SR) from the rules generated by the association rule mining algorithm. Then algorithm counts occurrences of each item in R.H.S of sensitive rules (SR) from the rules generated by the association rule mining algorithm. The algorithm counts occurrence of each item in R.H.S of sensitivity rules. The algorithm finds Is= $(is_0, is_1 \ldots\ldots is_k)$ k<_n, by arranging those items in decreasing order of their counts. After that sensitivity of each item is calculated. Occurrence of each item in R.H.S of sensitivity rules. The algorithm finds Is= $(is_0, is_1 \ldots\ldots is_k)$k<_n, by arranging those items in decreasing order of their counts. After that sensitivity of each item is calculated. Then

transactions which support is 0 are sorted in descending order of their sensitivities.
The algorithm counts occurrence of each item in R.H.S of sensitivity rules. The algorithm finds Is= (is_0, is_1.....is_k) k<_n, by arranging those items in decreasing order of their counts. After that sensitivity of each item is calculated. Occurrence of each item in R.H.S of sensitivity rules. The algorithm finds Is= (is_0, is_1.....is_k)k<_n, by arranging those items in decreasing order of their counts. After that sensitivity of each item is calculated. Then transactions which support is 0 are sorted in descending order of their sensitivities.

- Rule Hiding

This paper uses heuristic based approaches for hiding the sensitive rule. The data distortion mechanism changes the item value by a new value in dataset. It alters '0' to '1' or '1' to '0' for selected items in selected transactions to decrease the confidence, by decreasing or increasing support of items in sensitive rules. Sensitivity of Item is the number of sensitivity rules which contain this item. The sensitivity of transactions is the total of sensitivities of all sensitive items which are presented in this transaction.

The Rule hiding process starts by selecting first transaction form the sorted transaction with higher sensitivity; delete item is_0 from that transaction. Then update support and confidence of all sensitive rules and if any rules have support and confidence below MST and MCT respectively then delete it from SR. Finally update sensitivity of each item, transaction and Is. Again select transaction with higher sensitivity and delete is0 from it. This process continues until all sensitive rules are hidden.

- Sanitized Database Generation

The modified transactions are updated in the original database and new database is generated which is called sanitized database D', which preserves the privacy of sensitive information and maintains database quality.

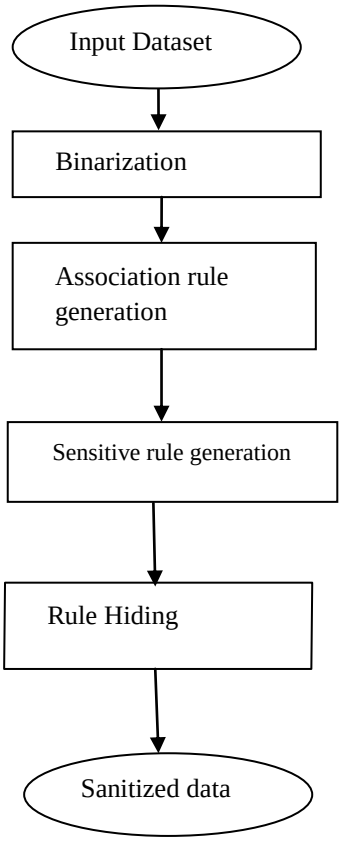

Fig1 .General Architecture

V. Example

To understand MDSRRC following example is illustrated In Table I transactional database D is shown. With 3 as MST and 40% as MCT, The possible generated association rules by Apriori algorithm[4] are:FishCrab→Salad,FishSalad→Crab,Crab→FishSalad,Salad→FishCrab,Fish→crab,Fish→Salad,Crab→Fish,Salad→Fish,CrabSalad→Fish.Let the Database Owner Specify Crab→FishSalad,Salad→FishCrab as sensitive rules. The sensitivity of Crab=2, Salad=2, Fish=2.

Transaction with itssensitivity is shown in Table II. Now algorithm findsfrequency of each item presents in R.H.S of sensitive rules.Here frequency of Crab=1,Salad=1,Fish=2 IS={Fish,Salad,Crab}.
In this example item 'Fish' is selected as is_0. Then it sorts the transactionswhich supports is_0 in descending order of their sensitivity. Then Select transaction with highest sensitivity and delete is0 item from that transaction. Update confidenceand support of all the sensitive rules. Table III show modified database D1after first deletion of item from first transaction. Now update sensitivity of each item. Updated count of eachitem for IS .Sort transactions which support $is_0$and delete the is_0 from transaction with highest sensitivity, then delete the is0 from transaction with highest sensitivity. Now

all sensitiverules are hidden. Final sanitize database is shown in table III.

Table I. Transaction Database

TID	Items	Binary matrix of Items
TRN001	Tofutti,fish,steak	1110000000
TRN002	Steak,beef	0011000000
TRN003	Fish,steak	0110000000
TRN004	Fish,monkfish,potato	0100110000
TRN005	Fish,steak,halibut	0110001000
TRN006	Fish,crab,salad	0100000110
TRN007	Fish,potato,crab	0100010100
TRN008	Fish,salad	0100000010
TRN009	Fish,potato,crab,salad	0100010110
TRN010	Tofutti,fish,crab	1100000100
TRN011	Tofutti,steak,salad,shell	1010000011
TRN012	Tofutti,fish,crab,salad	1100000110
TRN013	Tofutti,steak,beef,potato	1011010000

Table II.Transaction with Sensitivity

TID	**Sensitivity**
TRN001	2
TRN002	0
TRN003	2
TRN004	2
TRN005	2
TRN006	6
TRN007	4
TRN008	4
TRN009	6
TRN010	4
TRN011	2
TRN012	6
TRN013	0

Table III.Sanitized Database D'

TID	**Items**
TRN001	Tofutti,steak
TRN002	Steak,beef
TRN003	Fish,steak
TRN004	Fish,monkfish,potato
TRN005	Fish,steak,halibut
TRN006	Fish,crab,salad
TRN007	Fish,potato,crab
TRN008	Fish,salad
TRN009	Fish,potato,crab,salad
TRN010	Tofutti,fish,crab
TRN011	Tofutti,steak,salad,shell
TRN012	Tofutti,fish,crab,salad
TRN013	Tofutti,steak,beef,potato

VI. EXPERIMENTAL RESULT AND ANALYSIS OF PROPOSED ALGORITHM

We compare MDSRRC algorithm with DSRRC algorithm. These algorithms are used to hide the Sensitive rules on database. After applying bothAlgorithms on sample database we have done evaluation by considering the performance parameters which are given in [12] viz.

(a) HF (hiding failure): It is the percentage of the sensitive data that remain exposed in the sanitized dataset.
(b) MC (misses cost): It is the percentage of the non-sensitive data that are hidden as a side-effect of the sanitization process.
(c) AP (artificial patterns): It is the percentage of the discovered patterns that are artifacts.
(d) DISS (dissimilarity): It is the difference between the original and the sanitized datasets.
(e) SEF (side effect factor): It is the amount of non-sensitive association rules that are removed as an effect of the sanitization process.

Experimental results show that MDSRRC increase efficiency and reduce modification of transactions in database. Performance comparison of MDSRRC with algorithm DSRRC is given in Table VI.

Table VI. Performance result

Parameter	DSRRC	MDSRRC
MC	36%	26.66%
DISS(D,D')	6.4%	5.4%
HF	0%	0%
SEF	36.5%	26.66%
AP	0%	0%

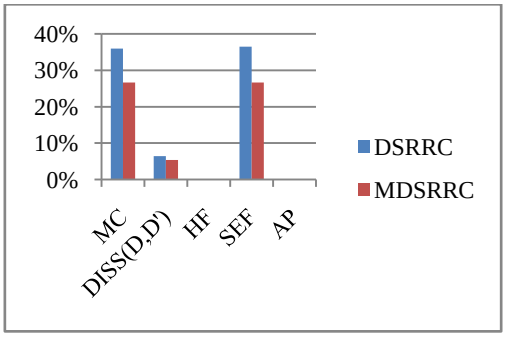

Fig. 2. Performance Comparison between DSRRC, MDSRRC

VI.CONCLUSION

The purpose of the Association rule hiding techniques for privacy preserving data mining is to hide certain crucial information so they cannot discovered through association rule. In this paper, we proposed an algorithm named MDSRRC which hides sensitive association rules with fewer

modifications on database to maintain data quality and to reduce the side effect of database. Functionality of proposed algorithm is shown using sample database with three sensitive rules. Experimental results show that proposed algorithm works better than DSRRC. So MDSRRC hide sensitive rules with minimum modifications on database and maintain data quality.MDSRRC algorithm can be extended to increase the efficiency and reduce the side effects by minimizing the modifications on database.

REFERENCES

[1] Nikunj H. Domadiya and Udai Pratap Rao, "Hiding Sensitive Association Rules to Maintain Privacy and Data Quality in Database"*3rd IEEE International Advance Computing Conference (IACC) 2013.*

[2] V. S. Verykios, A. K. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni, "Association rule hiding," *IEEE Transactions on Knowledge and Data Engineering, vol. 16, pp. 434–447, 2004.*

[3] C. N. Modi, U. P. Rao, and D. R. Patel, "Maintaining privacy and data quality in privacy preserving association rule mining," *2010 Second International conference on Computing, Communication and Networking Technologies, pp. 1–6, Jul. 2010.*

[4] J. Han, Data Mining: Concepts and Techniques. San Francisco, CA, USA*: Morgan Kaufmann Publishers Inc., 2005.*

[5] Y.-H. Wu, C.-M. Chiang, and A. L. Chen, "Hiding sensitive association rules with limited side effects," *IEEE Transactions on Knowledge and Data Engineering, vol. 19, pp. 29–42, 2007.*

[6] S.-L. Wang, B. Parikh, and A. Jafari, "Hiding informative association rule sets," *Expert Systems with Applications, vol. 33, no. 2, pp. 316 – 323, 2007.*

[7] S.-L.Wang, D. Patel, A. Jafari, and T.-P. Hong, "Hiding collaborative recommendation association rules," *Applied Intelligence, vol. 27, pp. 67–77, 2007.*

[8] D.F. Gleich and L.-H. Lim. Rank aggregation via nuclear norm minimization. *In KDD, 2011.*

[9] Y. Saygin, V. S. Verykios, and A. K. Elmagarmid, "Privacy preserving association rule mining." *in RIDE. IEEE Computer Society, 2002, pp 151–158.*

[10] C. N. Modi, U. P. Rao, and D. R. Patel, "An Efficient Solution for Privacy Preserving Association Rule Mining," *(IJCNS) International Journal of Computer and Network Security, vol. 2, no. 5, pp. 79–85, 2010.*

[11] Wu and H. Wang, "Research on the privacy preserving algorithm of association rule mining in centralized database," *in Proceedings of the 2008 International Symposiums on Information Processing, ser. ISIP'08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 131–134*

[12] V. Verykios and A. Gkoulalas-Divanis, A Survey of Association Rule Hiding Methods for Privacy, ser. Advances in Database Systems, C.

[13] M. Atallah, A. Elmagarmid, M. Ibrahim, E. Bertino, and V. Verykios, "*Disclosure limitation of sensitive rules," in Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange, ser. KDEX '99. Washington, DC, USA: IEEE Computer Society, 1999, pp. 45–52.*