International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 3, Issue 7, July 2014 **ISSN 2319 - 4847**

Maintaining Privacy and Data Quality to Hide Sensitive items from Database

Mr. Pravin R. Ponde[1] , 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

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

1. 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 be executed 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 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 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 confidential item can't be reveal. Our approached is based on modifying the database in a way that confidence of the association rule can be reduce with the help increase or decrease the support value of RHS or LHS correspondingly. As the confidence of the rule is reduce below a specified threshold, it is hidden or we can say it will not be disclosed.The proposed formula is that the improved version of DSRRC. DSRRC could not hide association rules with multiple items in antecedent (L.H.S) and resultant (R.H.S.). To overcome this limitation, we proposed an algorithmic rule MDSRRC which uses count of things in resultant of the sensitive rules. It modifies the minimum number of transactions to cover most sensitive rules and maintain data quality.

2. RELATED WORK

There have been several methods proposed for hiding sensitive patterns in dataset. In 1999, M. Attalah and E. bernito proposed the idea of Disclosure limitation of sensitive rules. It discusses security risks of database when reveals it in public. They introduce algorithm for hiding sensitive items with little impact on database. V.S. Verykios, A. K. Elmagarmid presents 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, 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 that purchase mobiles from two companies X and Y. Now X applies data mining techniques and mines association rules applied to related to Y'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 discharge the database with sensitive information cause the problem. This scheme provides the order on sensitive rules hiding in the database.The proposed algorithm customize least possible number of transactions to hide supreme sensitive rules and preserving data quality.

3. PROBLEM STATEMENT

Association rule activity problem is defined as: convert 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:

1. Database must not disclose any sensitive rules.
2. Sanitized database must facilitate mining of all non-sensitive rules.
3. 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.

4. METHODOLOGY

The proposed methodology is categorized into five major modules viz; Binarization, Apriori, Sensitive rules generation, MDSRRC algorithm and sanitized database creation. The Fig1 shows general architecture of the proposed method.

- We have integrated a pre-binarization step in order to enhance the input dataset quality. Let a, b, c, d, e, f be the items in transcational dataset. Let T be the number of transactions. Therefore during the binarization we get the following binarized output dataset.

Table.1: Binary Table

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

4.1 Applying Apriori

Apriori is an influencial algorithm for generating association rules. Association rules generation is usually split up into two separate steps.

1. Minimum support is applied to find frequent items sets in a dataset.
2. Secondly, those frequent item sets and the minimum confidence constraints to form mining association rules.

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.

4.2 Sensitive Rule Generation

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 transactions.

4.3 MDSRRC

This algorithm starts with mining the association rule from the original dataset using association rule mining algorithm. 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= ($is_{0,}$ $is_{1.....}is_{k}$) k<_n, by arranging those items in decreasing order of their counts. After that sensitivity of each item is calculated.

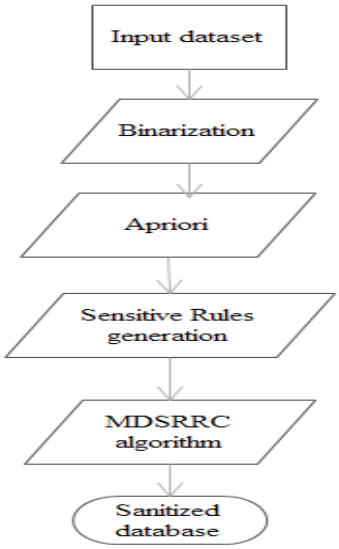

Fig1 . General Architecture

Then transactions which support is0 are sorted in descending order of their sensitivities. This tends to the initialization of Rule hiding process 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. 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.

4.5 Sanitized Database Generation

The possible generated association rules by Apriori algorithm

a→b , b→ a, a→ c, c → a, a→d,
d→ a, b→c, c→b, b→d, d→b, c→ d,
d→ c, c → e
e→c, d → e, e → d, a→ cd, c→ad,
ac → d, d→ac, ad→c, cd→a

Let the database owner specify rule a→bd, a→cd and d→ac as sensitive rules. Then select transaction with highest sensitivity and delete is_0 item from that transaction. Update confidence and support of all the sensitive rules. Sort transactions which support is_0 and delete the is_0 from transaction with highest sensitivity, then delete the is_0 from transaction with highest sensitivity. Finally all the sensitive rules are hidden.

5. 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] 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.

[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.
[14] 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
[15] V. Verykios and A. Gkoulalas-Divanis, A Survey of Association Rule Hiding Methods for Privacy, ser. Advances in Database Systems, C.

AUTHORS

Mr. Pravin R. Ponde , M.E, Department of Computer Science and Engineering, TPCT's College of Engineering, Osmanabad - 413501, Maharashtra, India

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