A Heuristic Based Approach for Hiding Sensitive Data to Achieve Privacy

Gajendra Prasad K C[1] Prof. Sathish Kumar S[2]
[1]M.Tech Student [2]Associate Professor
[1,2]Department of Computer Science Engineering
[1,2]RNS Institute of Technology Bangalore, Karnataka, India

Abstract— Hiding sensitive data to ensure confidentiality and privacy of the information has been a long term goal in the field of database security, so it is very important for the organizations to ensure that before disclosing data to the public access sensitive data must be hidden. In this paper, we are implementing a algorithm called MDSRRC (modified decrease support of RHS item of Rule Cluster) to hide the sensitive patterns ,which is the advanced version of Decrease support algorithm, which could applied over multiple itemsets whereas decreased support algorithm cannot be applied over multiple itemsets.

Keywords: MDSRRC, sensitive data, sensitivity levels.

I. INTRODUCTION

Recent technologies in the field of data mining enabled very efficient extraction of data from large repositories, this increases the risk of disclosing data when the website ,databases are to be released to public access. Information such as passwords, credit information, type of diseases, income, customer purchases, atm passwords are sensitive in nature. To address this confidentiality and privacy preserving issue, the original database is sanitized in such a way that sensitive patterns are hidden. Association rule mining enables us to find relationship between the items [2]. Many companies disclose their database for the mutual benefit, but before doing so they got to ensure their private data is hidden.

So it is very much important for the organizations which having very huge amount of data to ensure the confidentiality and privacy of the data from the unauthorized access, it is equally important to share the data among other organizations for the mutual benefit but it is very crucial to ensure that the sensitive information got to be hidden before disclosure, or else the other organizations may take advantage and mines sensitive data by using different data mining techniques and could bring down the business, so any organizations companies before disclosing their database they must have ensured that sensitive patterns are well protected and could not be mined.

Hiding or masking the sensitive information is the key task in order to achieve privacy and confidentiality of the information it is the major goal in the field of database security so it is very much needed to ensure the privacy factor in the websites and database, and safeguard the sensitive data from unauthorized access. The proposed algorithm is the improved version of DSRRC which could be applied over multiple itemsets unlike DSRRC.

II. RELATED WORK

In this section we study notations used in this paper and terms like support and confidence.

TID	Items
T1	XYZ
T2	XYZ
T3	XYZ
T4	XY
T5	X
T6	XZ

Table. 1: Sample Database

Itemset	Support
X	100%
Y	66%
Z	66%
XY	66%
XZ	66%
YZ	50%
XYZ	50%

Table. 2: Support for Items

Rules	Confidence	Support
Y ⇒ X	100%	66%
Y ⇒ Z	75%	50%
Z ⇒ X	100%	66%
Z ⇒ Z	75%	50%
Y ⇒ XZ	75%	50%
Z ⇒ XY	75%	50%
XY ⇒ Z	75%	50%
XZ ⇒ Y	75%	50%
YZ ⇒ X	100%	50%

Table. 3: The Support And Confidence For The Rules

The table 2.1 represents the sample database [2] which has six transactions with items X, Y, Z, as item X is present in all the six transactions its support is 100% table 2.2 , Y and Z are present in four transactions out of six transactions so its support is 66% table 2.2 ,similarly we could calculate support of different combination of itemsets as shown in table 2.2.

The table 2.3 represents support and confidence of different rule Y→X has confidence 100% because item Y is associated with item X in all the four transactions in which it is present table 2.1,the support of the rule Y→X is 66% because the XY combination is present in four transactions out of six transactions in sample database table 2.1.

Consider the rule Z→Y, item B is present in four transactions out of six transactions in sample database table 2.1, item is present in four transactions out of six transactions in sample database table 2.1,Y is associated with Z in three transactions out of its presence in four transactions so confidence for the rule is 75% and combination YZ is present in three transactions out of six

transactions in sample database table 2.1 ,so for the combination is support is 50%.
Similarly we could calculate support and confidence for all the rules in table 2.3.

D	Original database
D'	Sanitized database
R	Association rules generated from original database
SR	Sensitive association rules SR $\subset$ R
Ti	The i^{th} transaction of database
I	Set of distinct items in database
IS	IS={is0,is1,...,isk} k $\leq$ n, Set of items present in consequent of sensitive rules with decreasing order of their frequency in consequent of sensitive rules
is0	Item with highest count in consequent of sensitive rules.
MST	Minimum support threshold
MCT	Minimum confidence threshold
L.H.S	Antecedent of an association rule
R.H.S	Consequent of an association rule

Table. 4: Notations Used.

Table 2.4 represents the notations used in this paper. In this paper we are mainly concerned with hiding sensitive patterns that must not be disclosed. In order to hide the itemsets we are using modified version of decreased support algorithm because it could be applied on multiple itemsets, whereas it is not possible in decreased support algorithm. The strategy we are following is selecting transactions from database, on the selected transactions apriori algorithm is applied to generate rules. From the generated rules some of the rules are selected as sensitive rules, for the selected sensitive rules sensitivity levels are calculated and sorted in decreasing order of their sensitivity levels, items with maximum sensitivity levels are hidden first likewise the process continues till all the sensitive items are hidden. In the original database sensitive items is not hidden, whereas in sanitized database sensitive items are hidden.

III. PROPOSED WORK

In this section we study about the working procedure of MDSRRC algorithm.

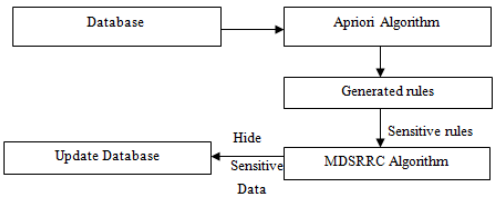

Fig. 1: Framework for MDSRRC

The figure 3.1 represents the system framework of MDSRRC. The database will be having transactions present in it, the transactions will be given as input to the apriori algorithm, the algorithm selects the transactions which meets minimum support threshold and generates rules. From the generated rules some of the rules are selected as sensitive such rules must be hidden. The MDS algorithm, calculates the sensitivity levels for the all the transactions for the corresponding sensitive rules.

The transactions are sorted in the descending order of their sensitivity levels, the items with maximum sensitivity levels are hidden first, likewise all the selected items are hidden. After hiding all the sensitive data the same is updated in the database in such a way that non sensitive data is visible and sensitive part of the data is hidden.

To understand MDSRRC following example is illustrated consider table 3.1 with 3 as MST and 40% as MCT, the possible rules generated by apriori algorithm [3] .p→q, q→p, p→r, r→p, p→s, s→p, q→r, r→q, q→s, s→q ,r→s, s→r, r→t, t→r, s→t, t→s, p→rs, r→ps, pr→s, s→pr, ps→r, rs→p, r→st, s→rt rs→t, t→rs, rt→s, st→r, p→qs, q→ps, pq→s, ps→q, qs→p.

Let the owner of the database specifies p→qs, p→rs and s→pr as the sensitive rules to be hidden. The sensitivity levels of p=3,q=1,r=2,d=3 and the frequency(no of times occurred on RHS side of the rule) of p=1,q=1,r=2 and s=2. IS={s,r,q,p} Select the transactions with highest sensitivity levels and delete is0 item from that transaction and update support and confidence, similarly all the marked sensitive items are hidden after every deletion database is updated.

TRANS ID	ITEMS
1	pqrst
2	prs
3	pqsuv
4	qrst
5	pqs
6	rstuw
7	pqrv
8	prst
9	prsw

Table. 5: Transactional Database

TRANS ID	ITEMS
1	10
2	9
3	8
4	7
5	8
6	6
7	7
8	9
9	9

Table. 6: Transactions with its Sensitivity Levels.

TRANS ID	ITEMS
1	pqrt
2	Prs
3	pqsuv
4	qrst
5	pqs
6	rstuw
7	pqrv

8	prst
9	prsw

Table. 7: Sanitized Database I(after first deletion)

TRANS ID	ITEMS
1	pqrst
2	ps
3	pqsuv
4	qrst
5	pqs
6	rstuw
7	pqrv
8	prst
9	prsw

Table. 8: Final Sanitized Database II(after second deletion)

IV. EXPERIMENTAL RESULTS

The proposed MDSRRC algorithm requires fewer no of modifications to hide the sensitive rules whereas DSRRC require more number of modifications to the database to hide the sensitive items. Fig 4.1 gives the performance comparison between DSRRC and MDSRRC.

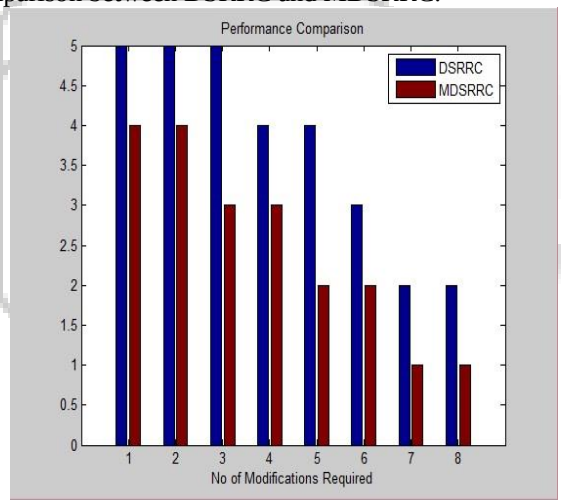

Fig. 2: Performance Comparison between DSRRC and MDSRRC

V. CONCLUSION AND FUTURE SCOPE

This paper addresses the process of hiding sensitive data, which is essential in many applications. The project proposes new algorithm called MDS algorithm which efficiently hides the sensitive data, the proposed algorithm could be applied over multiple item sets to hide sensitive data which is not possible in the existing system, and the proposed MDS algorithm does fewer modifications on database to maintain data quality and of the database.

All the proposed methods and the previous approaches have their own pros and cons. So does this concept. This chapter emphasize on the enhancements that can be made so that better and efficient results can be obtained. The use of the algorithm like FP growth algorithm is very much useful whose computation speed is more compared to apriori algorithm for generating frequent item set, which has better performance when compared to apriori algorithm.

ACKNOWLEDGEMENT

We would like to thank Director Dr. H N Shivashankar, Principal Dr. M K Venkatesha and Dr. G T Raju, Professor and Head, Dept of Computer Science and engineering, RNSIT, for their constant support and encouragement.

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. 45–52.2013 3rd IEEE International Advance Computing Conference (IACC) 1309

[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.

[3] J. Han, Data Mining: Concepts and Techniques. San Francisco, CA, USA Morgan Kaufmann Publishers Inc., 2011 3rd edition

[4] 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.

[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.

