Volume 4, Issue 2, February 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Privacy Preservation by Hiding Highly Sensitive Rules with Fewer Side Effects

Dharmendra Thakur
M.Tech. Scholar, P.C.S.T., BHOPAL
India

Prof. Hitesh Gupta
C.S Dept, P.C.S.T., BHOPAL
India

Abstract- In this paper, we propose a heuristic based algorithm named AMDSRRC (Advanced Modified Decrease Support of R.H.S. item of Rule Clusters) to hide the highly sensitive association rules with multiple items in consequent (R.H.S) and antecedent (L.H.S). The highly sensitive rule is the one having support greater than or equal to MST (Minimum Support Threshold) and confidence greater than or equal to MCT (Minimum Confidence Threshold) & near to 100%. The Efficiency and scalability of the proposed algorithm is better than existing rule hiding algorithm MDSSRC. Experimental result shows that proposed algorithm is highly efficient and maintains database quality.

Keyword: Frequent Itemsets, Highly Sensitive Rule, Sensitivity, Association rule, Sanitization, Performance Parameters.

I. INTRODUCTION

Today, with the volatile growth in Internet, storage and data processing technologies, privacy preserving is the key concern in business fields. If a publicly available system is to be made secure, we must guarantee that private sensitive data have been spruced out and that certain inference channels have been blocked as well. The privacy preserving association rule mining has been a widely used approach concerning hiding private data by sanitizing the original database. Association rule hiding algorithms can be divided into three main classes, namely heuristic approach, border-based approach, and exact approach [6]. Heuristic approach has been the hot topic of research in recent years. Heuristic approaches hide sensitive association rules by directly modifying, or we say, sanitizing the original data D, and get the released database D' directly from D. Heuristic Based Approaches can be further divided into two groups based on data modification techniques: data distortion techniques and data blocking techniques. The proposed algorithm mainly based on distortion technique to decrease support or confidence in order to hide the sensitive rules. The proposed algorithm is the improved version of MDSRRC. Following are the limitations of MDSRRC:

(i) Database owner specifies the sensitive rules.
(ii) Do not mention the number of sensitive rules to be taken into consideration.
(iii) What are those sensitive rule

The proposed algorithm overcomes all the above limitations and increases the efficiency of MDSSRC by applying a novel idea of selecting highly sensitive rules to hide sensitive data.

II. PROBLEM DEFINITION

An association rule is an implication of the form $X \rightarrow Y$, where X, Y are Itemsets, and $X \cap Y = \emptyset$. We say the rule $X \rightarrow Y$ holds in the database D with confidence c if $|XUY|/|X| \geq c$. It can also be said that the rule $X \rightarrow Y$ has support s if $|XUY|/|D| \geq s$. Note while the support is a measure of the frequency of a rule, the confidence is a measure of the strength of the relation between sets of items. The well-known association rule mining problem aims to find all significant association rules. A rule is significant if its support and confidence is no less than the user specified Minimum Support Threshold (MST) and Minimum Confidence Threshold (MCT). To find the significant rules, an association rule mining algorithm first finds all the frequent itemsets and then derives the association rules from them. On the contrary, the association rule hiding problem aims to prevent some of these rules, which is referred as "sensitive rules", from being mined.

Given a data set D to be released, a set of rules R mined from D, and a set of Highly Sensitive Rules $HS_R \subseteq R$ to be hided, how can we get a new data set D', such that the rules in HS_R cannot be mined from D', while the rules in R- HS_R can still be mined as many as possible.

III. LITERATURE REVIEW AND THEORETICAL BACKGROUND

Table 1: Notation and definition

D	Original database
R	Association rules generated from D
HS_R	Highly Sensitive Rules $HS_R \subseteq R$
D'	Sanitized database

TR	Set of transactions
T_i	i_{th} transaction in TR
IT	Set of distinct items in D
I_i	i_{th} item in IT
ICS	ICS= {ics_0,ics_1,ics_2,….,ics_k } k≤n, n=total items in D Set of items present in R.H.S. of sensitive rules with decreasing order of their frequency in R.H.S. of sensitive rules
ics_0	Item with highest count in consequent of sensitive rules
Min_sup	Minimum support threshold
Min_conf	Minimum confidence threshold
L.H.S.	Antecedent of an association rule
R.H.S.	Consequent of an association rule
MSS	Maximum support among highly sensitive rules
Diff	Difference of MSS and Min_sup

In Table1, we show the notations used in this paper. Mining an association rule with support confidence is defined as follow:

Let $IT = \{I_1, I_2, \ldots, I_m\}$ be a set of items [2]. Let D be a database of transactions where each transaction T_i is a set of items such that $T_i \subseteq IT$. Each transaction is associated to an identifier, call *TID*. A transaction T_i is said to contain A if and only if $A \subseteq T_i$. An association rule is an implication of the form $A \Rightarrow B$, where $A \subseteq IT$, $B \subseteq IT$, and $A \cap B = \emptyset$. The rule $A \Rightarrow B$ holds in the transaction set D with support s, where s is the percentage of transactions in D that contain $A \cup B$. The rule $A \Rightarrow B$ has confidence c, where c is the percentage of transactions in D that contain A also contain B. That is,

$$Sup\ (A \Rightarrow B) = P\ (A \cup B) = \frac{|A \cup B|}{|D|} \qquad (1)$$

$$Conf\ (A \Rightarrow B) = P\ (B/A) = \frac{|A \cup B|}{|A|} \qquad (2)$$

where A is named as the support count of the set of items A in the set of transactions D, as denoted by *sup_ count (A)*. A occurs in a transaction T_i, if and only if $A \subseteq T_i$. Rules that satisfy both a minimum support threshold (*Min_ sup)* and a minimum confidence threshold (*Min_conf)* are called strong. A set of items referred to as an itemset. An itemset that contains k items is a k -itemset. Itemsets that satisfy *Min_ sup* is named as frequent itemsets. All strong association rules result from frequent itemsets.

IV. PROPOSED FRAMEWORK & ALGORITHM

The framework of the proposed algorithm is depicted in Fig. 4.1

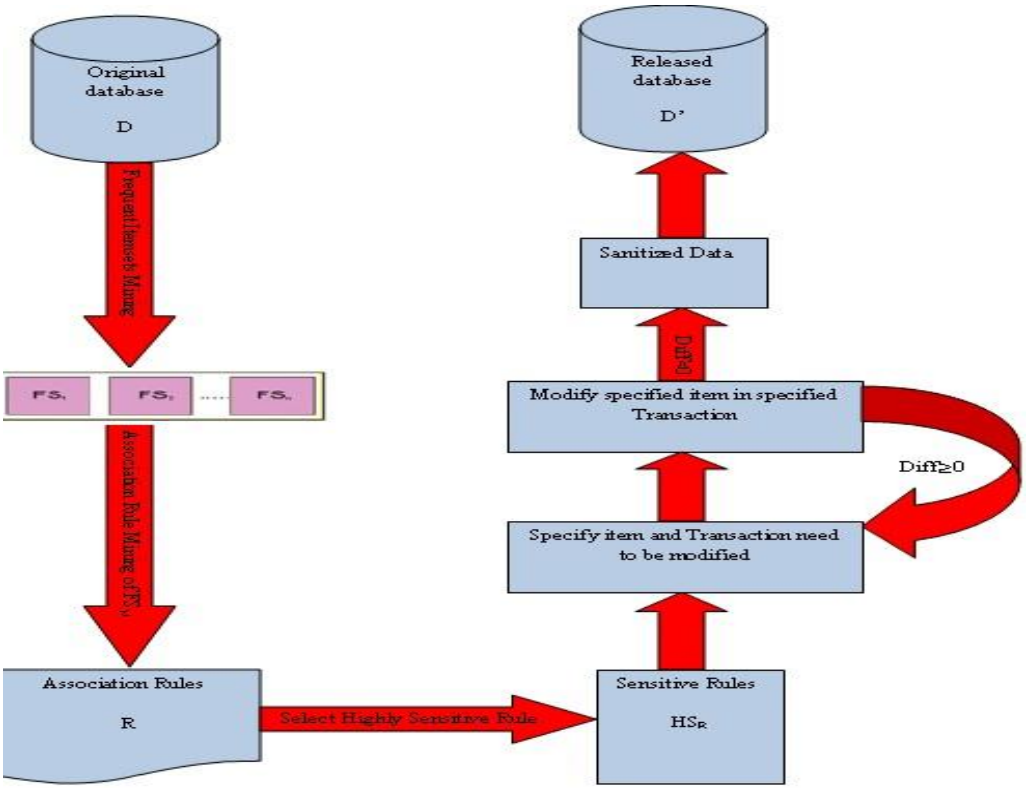

Fig. 4.1 Proposed Framework

The first step is to generate frequent 1-itemset, 2-itemset,…,K-itemset using Apriori algorithm for the set of transactions *TR* of database *D*. Then select the k^{th} level itemsets for association rule mining. The set *R* contains only those association rules whose support≥*Min_sup* and confidence≥*Min_conf* and rest of the rules are trimmed out. The highly sensitive rules are then selected for hiding purpose. The highly sensitive rules are those having support ≥ *MST* (Minimum Support Threshold) and confidence ≥ *MCT* (Minimum Confidence Threshold) & near to 100% (one from each cluster). The set HS_R contains these highly sensitive rules. The algorithm then calculates Diff which indicate the number of iteration to complete the hiding process and is equal to *MSS-Min_sup*. The algorithm counts occurrences of each item in R.H.S of highly sensitive rules. Now algorithm finds *ICS* by arranging those items in decreasing order of their counts. After that sensitivity of each item is calculated then sensitivity of each transaction is calculated.

Now rule hiding process starts by sorting the transactions which support ics_n (where n=0, 1, 2,…) in descending order of their sensitivities. Then select first transaction from the sorted transactions with higher sensitivity, delete item ics_n from that transaction. Repeat the above steps till k ≤ *Diff*, where k = 0, 1, 2,..

As a result, 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.

Proposed algorithm is shown below, which is used to hide the sensitive rules from database. Given algorithm generates sanitized database D'. Sanitized database hides all sensitive rules and maintains data quality.

A. Algorithm

Input: (MST,MCT,D)

Output: D' with all sensitive rules are hidden

1. Apply Apriori algorithm on given database D to generate all frequent itemsets.
2. Select k^{th}level itemsets for association rule mining.
3. Retain rules whose support ≥ MST and confidence ≥ MCT and say the set is R.
4. Select Highly Sensitive Rules from R and say the set is $HS_{R.}$
5. Calculate the sensitivity of each item j where $j \in HS_{R.}$
6. Calculate the sensitivity of each transaction.
7. Count occurrences of each item in R.H.S of sensitive rules.
8. Find ICS= {ics_0, ics_1, ics_2…ics_k} k≤n, by arranging those items in descending order of their count. If two items have same count then sort those in descending order of their actual support count.
9. Set Diff=MSS-Min_sup; k=0
10. If k > Diff goto step 15
11. Select the transactions which supports ics_k then sort them in descending order of their sensitivity. If two transactions have same sensitivity then sort those in increasing order of their length.
12. Delete item ics_k from the first transaction from sorted list.
13. Set k=k+1
14. goto step 10
15. End

V. EXAMPLE

In Fig.5.1, transactional database D is shown. With 3 as MST and 40% as MCT , the possible frequent itemsets using Apriori algorithm are:

Frequent 1-itemsets:
a, b, c, d, e — 1st Level

Frequent 2-itemsets:
ab, ac, ad, bc, bd, cd, ce, de — 2nd Level

Frequent 3-itemsets:
abd, acd, cde — 3rd Level

TID	Items
T_1	abcde
T_2	acd
T_3	abdfg
T_4	bcde
T_5	abd
T_6	cdefh
T_7	abcg
T_8	acde
T_9	acdh

Fig 5.1 Original Database D

Select 3rdLevel itemsets for association rule mining. The possible association rules are shown Fig.5.2. Let R be set of association rules.

R : {a b d}		
Association rules	support	Confidence %
a→bd	3	42
b→ad	3	60
ab→d	3	75
bd→a	3	75
ad→b	3	50

R : {a c d}		
Association rules	support	Confidence %
a→cd	4	57
c→ad	4	57
d→ac	4	50
ac→d	4	80
ad→c	4	66
cd→a	4	66

R : {c d e}		
Association rules	support	Confidence %
c→de	4	57
d→ce	4	50
e→cd	4	100
cd→e	4	66
ce→d	4	100
de→c	4	100

Fig.5.2 Set of Association Rules

Select Highly Sensitive Rules as shown in Fig.5.3. The number of highly sensitive rules is equal to *Min_sup*, and say the set is HS_R . Count the occurrence of each item jϵHS_R

Association rule	Support	Confidence%
e→cd	4	100
ce→d	4	100
de→c	4	100

HS_R={ ce→d ; de→d ; e→cd }

Sensitivity of each item j$\in HS_R$

Item	Sensitivity
c	3
d	3
e	3

Fig.5.3 Highly Sensitive Rules & sensitivity of items

Count of items in R.H.S of sensitive rule: d=2, c=2

ICS = {d, c}

Calculate *Diff* = *MSS- Min_sup* = 4-3 = 1; Set k = 0

Since *Diff* is 1, our rule hiding process completes in two iterations. At first iteration k = 0 and k < *Diff*, rule hiding process starts by sorting the transactions which support ics_0 (i.e. d) in descending order of their sensitivities. If two transactions have same sensitivity then sort those in increasing order of their length. Then select first transaction from the sorted transactions , delete item d from that transaction as shown Fig.5.4.

TID	Items	Sensitivity
T_4	bcde	9
T_8	acde	9
T_1	abcde	9
T_6	cdefh	9
T_2	acd	6
T_9	acdh	6
T_5	abd	3
T_3	abdfg	3

Delete d from T_1

TID	Items
T_4	bce
T_8	acde
T_1	abcde
T_6	cdefh
T_2	acd
T_9	acdh
T_5	abd
T_3	abdfg

Fig.5.4

At second iteration k = 1 and k = *Diff*, rule hiding process then sort the transactions which support ics_1 (i.e. c) in descending order of their sensitivities. If two transactions have same sensitivity then sort those in increasing order of their length. Then select first transaction from the sorted transactions, delete item c from that transaction as shown Fig.5.5.

TID	Items	Sensitivity
T_8	acde	9
T_1	abcde	9
T_6	cdefh	9
T_2	acd	6
T_4	bce	6
T_9	acdh	6
T_7	abcg	3

Delete c from T_8

TID	Items
T_8	ade
T_1	abcde
T_6	cdefh
T_2	acd
T_4	bce
T_9	acdh
T_7	abcg

Fig.5.5

Now k = 2 and k > *Diff*, so rule hiding process terminates. Final sanitize database is shown Fig.5.6.

TID	Items
T_1	abcde
T_2	acd
T_3	abdfg
T_4	bce
T_5	abd
T_6	cdefh
T_7	abcg
T_8	ade
T_9	acdh

Fig.5.6 Sanitized Database D'

VI. EXPERIMENTAL RESULT AND ANALYSIS OF PROPOSED ALGORITHM

Here we compare our proposed algorithm with MDSRRC algorithm. We used algorithm MDSRRC and our proposed algorithm to hide the specified sensitive rules on the sample database shown in Fig.5.1. The sensitive rules are specified by the owner itself in MDSRRC algorithm whereas our proposed algorithm selects sensitive rule automatically based on

confidence value of the rule. Higher the confidence measure more the sensitive it is. The number of sensitive rules to be selected is equal to the value of *Min_sup*. The result in Fig 6.1 shows that proposed approach is more efficient than MDSRRC considering the performance parameters which are (a) HF (hiding failure), (b) MC (misses cost), (c) AP (artifactual pattern), (d) Diss(dissimilarity) and (e) SEF(side effect factor).

Parameter	MDSRRC	Proposed algorithm (AMDSRRC)
HF	0%	0%
MC	26.66%	10%
AP	0%	0%
DISS(D,D')	5.4%	5.4%
SEF	26.66%	10%

Fig.6.1 Preformance Result

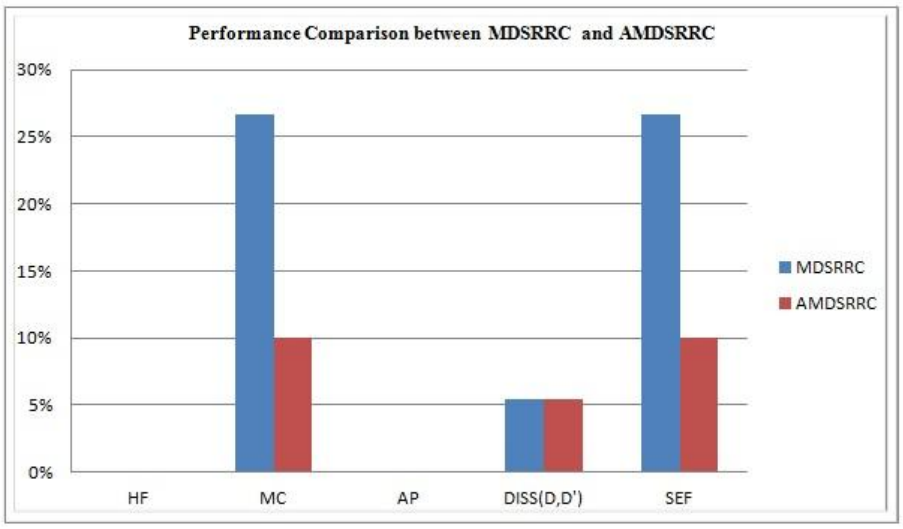

Fig.6.2 Performance Comparison between MDSRRC and AMDSRRC

Fig. 6.2 shows that performance of the AMDSRRC is better than MDSRRC in terms of database quality parameter. So AMDSRRC hide sensitive rules with minimum modifications on database and maintain data quality.

VII. CONLUSION AND FUTURE SCOPE

Our algorithm hides highly sensitive association rules with very few modifications on database ultimately maintain data quality. The Side Effect Factor (SEF) obtained using proposed algorithm is very much reduced than MDSRRC which means maximum non sensitive rules retain alongwith disappeared highly sensitive rules. In future, AMDSRRC 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,"* IEEE, 2012, *3rd IEEE International Advance Computing Conference (IACC)*, 2013.

[2] Kshitij Pathak, Narendra S Chaudhari, and Aruna Tiwari , *"Privacy Preserving Association Rule Mining by Introducing Concept of Impact Factor,"* IEEE, 2011, *7th IEEE Conference on Industrial Electronics and Applications ICIEA), 2012.*

[3] Hai Quoc Le, and Somjit Arch-int, *"A Conceptual Framework for Privacy Preserving of Association Rule Mining in E- Commerce,"* IEEE, 2011, *7th IEEE Conference on Industrial Electronics and Applications (ICIEA), 2012.*

[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] Pingshui WANG, *"Research on Privacy Preserving Association Rule Mining a Survey"* IEEE, 2010.

[6] S.Vijayarani, Dr.A.Tamilarasi, R.SeethaLakshmi, *"Privacy Preserving Data Mining Based on Association Rule-A Survey"* Proc. International Conference on Communication and Computational Intelligence-2010, kongu Engineering College, Perundurai, Erode, T.N., India.27-29 December, 2010, pp. 99-103

[7] Majid Bashir Malik, M.Asger Ghazi, Rashid Ali, *"Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects,"* IEEE, 2012, *3rd International Conference on Computer and Communication Technologys (ICCCT)*,2012, pp.26-31.

[8] Dharmendra Thakur, Hitesh Gupta, *"An Exemplary Study of Privacy Preserving Association Rule Mining Techniques,"* November-2013, IJARCSSE, Volume 3, Issue 11, pp. 893-900.

[9] Gayatri Nayak, Swagatika Devi, *"A survey on Privacy Preserving Data Mining: Approaches and Techniques"*, International Journal of Engineering Science and Technology, Vol. 3 No. 3, 2127-2133, 2011.

[10] Wang P, *"Survey on Privacy preserving data mining*", International Journal of Digital Content Technology and its Applications, Vol. 4, No. 9, 2010.

[11] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim3, V. Verykios, "*Disclosure Limitation of Sensitive Rules*," Proc. IEEE Knowledge and Data Engineering Workshop, Chicago, Illinois, 1999, pp. 25-52.

[12] E. Dasseni, V. S. Verykios, A. K. Elmagarmid, and E. Bertino, "Hiding *Association Rules by Using Confidence and Support,*" Proc. the 4th Information Hiding Workshop, Pittsburg, PA, 2001, LNCS2137, pp. 369-383.

[13] S. R. M. Oliveira, and O. R. Zaiane, "*Privacy Preserving Frequent Itemset Mining*" Proc. IEEE ICDM Workshop on Privacy, Security, and Data Mining, Maebashi, Japan, 2002, pp. 43-54.

[14] Y. Wu, C.M. Chiang, and A.L.P. Chen, "*Hiding Sensitive Association Rules with Limited Side-effects*," IEEE Transactions on Knowledge and Data Engineering, vol. 19, Jan. 2007, pp. 29-42.

[15] V.S. Verykios, A.K. Elmagarmid, E. Bertino, Y. Saygin, E. Dasseni, "*Association Rule Hiding*," IEEE Transactions on Knowledge and Data Engineering, vol. 16, Apr. 2004, pp, 434-447.

[16] R. Agrawal, and R. Srikant, "*Privacy-preserving data mining*," ACM SIGMOD Record, New York, vol. 29, Feb. 2000, pp.439-450.

[17] R. Agrawal and R. Srikant. "*Privacy Preserving Data Mining*", ACM SIGMOD Conference on Management of Data, pp: 439-450, 2000

