Sensitive Association Rule Hiding using Hybrid Algorithm in Incremental Environment

Ankit Kharwar
CE/IT department
Uka Tarsadia University
Bardoli, Surat, India

Chandni Naik
CE/IT department
Uka Tarsadia University
Bardoli, Surat, India

Niyanta Desai
CE/IT department
Uka Tarsadia University
Bardoli, Surat, India

Nikita Mistree
CE/IT department
Uka Tarsadia University
Bardoli, Surat, India

ABSTRACT
Security of the huge database which includes certain sensitive information will become a vital issue when the data is released to outside world. Privacy preserving data mining is a new research area to protect privacy for sensitive information from exposé. PPDM include various association rules hiding method. The existing approach follows the concept of hiding the rule by fine tune the support of the LHS and RHS item of the rule. So the proposed approach is to combine the concept of ISL and DSR algorithms by manipulating the support of the LHS and RHS item of the rule and RHID is used to hide the rules for incremental environment. The advantage of combining this is to hide the rules from both sides in incremental environment. A novel approach for ARM using pattern generation is used instead of traditional apriori which reduce multiple database scan and require less memory space.

Keywords
Association rule hiding; Hybrid algorithm; Incremental Association rule hiding; Privacy preserving data mining.

1. INTRODUCTION
Data mining is a well-known analysis field to discover valuable pattern from huge amount of data. These patterns give valuable information which is depict in terms of clusters decision trees and association rules. So the exposure risks of sensitive information are increased when the data is released to the anonymous parties. Finding unknown patterns while not revealing crucial information is one of the biggest challenges of data mining. Considering this, it becomes essential to hide sensitive knowledge in database. Privacy preserving data mining technique provides novel way to solve this problem. Association rule hiding is one of the methods of PPDM to protect the association rules which is produce by association rule mining. Association rule hiding is the methodology of modifying the original databases in such how that certain sensitive association rule vanish without affecting the data and the non-sensitive rules[1][2].

1.1 Association Rule Hiding
To hide sensitive association rules serveral privacy preserving techniques is used, Association rule hiding is one of them. Traditional association rule hiding algorithm intend to improve the original database such that no sensitive association rule is procure from it. Association rule hiding method incorporate decreasing the support or confidence of rules and decreasing the support of frequent itemsets which contain sensitive rules. The support of X in transactions which is not supporting Y will be increased by reducing confidence of rule and decreasing the support of Y in transactions supporting both X and Y.[1] The problem can be stated as follows: Given a transactional database D, a set R of rules mined from database D, minimum confidence and minimum support. R has a subset RH, it is a set of sensitive association rules which are to be hidden. The purpose is to transform D into a database D in such a way that all non-sensitive rules in R could still be mined from D but no association rule in RH will be mined. Fig.1 shows the general framework for association rule hiding.

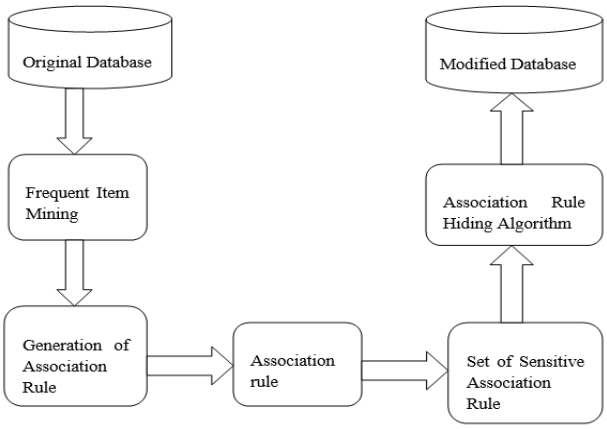

Fig 1: General framework for association rule hiding [1]

2. LITRATURE SURVEY
The aim of all association rules hiding algorithm is to minimally change the original database, deriving no sensitive knowledge and no sensitive association rule. We study all basic algorithms such as ISL and DSR[3][5], DSRRC[6][7], ADSRRC and RRLR[8], MDSRRC[9][10], RHID and ARM [11].

2.1 ISL and DSR
ISL algorithm decreases the confidence of rule by increasing the support value of LHS. It works only for modification of LHS; it doesn't work for both side of rule. In DSR algorithm, reducing the support value of RHS, confidence of the rule can be decrease. It works for the modification of RHS [8][10]. The benefit of ISL and DSR algorithm is that they required fewer numbers of databases scanning and trim large number of hidden rules. The disadvantages of these algorithms are that they do not hide the entire rule and goes for number of modification to hide certain rule. It does not work for both side of rule. If we want to hide right hand side rule we have to go for DSR algorithm and if we want to hide left hand side rule we have use ISL algorithm. Now let us consider an example of ISL and DSR algorithm. Suppose there is a

database which shown in table 2 [9]. Four association rules can be found: A → B (60%, 75%), B → A (60%, 75%), A → D (60%, 75%), D → A (60%, 100%) by giving MST=60% and a MCT=70%.
From the above generated association rules D and B items will be hiding from below transaction database.

Table 1. Original Database [9]

TID	ITEMS
T1	ABD
T2	B
T3	ACD
T4	AB
T5	ABD

In ISL algorithm, by modifying the transaction T2 from B to BD we can hide D and B. But, still ISL cannot hide the rule D → A because support and confidence of rule D → A will became 60% and 75% respectively if we modifyT2 from B to BD. We can hide rule D → A by DSR approach because its support and confidence is now 40% and 66% respectively, but as a side effect the rule A → D is also hidden.

TID	Items	Bitmap
T1	ABD	1101
T2	B	0100
T3	ACD	1011
T4	AB	1100
T5	ABD	1101

Hiding D → A by ISL

TID	Items	Bitmap
T1	ABD	1101
T2	BD	0101
T3	ACD	1011
T4	AB	1100
T5	ABD	1101

Fig 2: ISL example [9]

TID	Items	Bitmap
T1	ABD	1101
T2	B	0100
T3	ACD	1011
T4	AB	1100
T5	ABD	1101

Hiding D → A by DSR

TID	Items	Bitmap
T1	BD	0101
T2	B	0100
T3	ACD	1011
T4	AB	1100
T5	ABD	1101

Fig 3: DSR example [9]

2.2 RHID Algorithm

Existing algorithms are designed for static database so they cannot handle incremental datasets effectively and efficiently. To deal with this problem, RHID (Rule Hiding for Incremental Dataset) algorithm is used. It is used to conceal the rules for incremental environment. This algorithm is an extension of MDSRRC algorithm. The goal of RHID algorithm is to maintaining dataset quality, preserving privacy and to achieve low computational cost [17]. The framework of RHID can be divided into two parts: the first is rule hiding in original database that uses the MDSRRC algorithm and the second is rule hiding in incremental database that use RHID Algorithm. Fig. 4 shows general framework of RHID algorithm. First mine the association rules AR from original database D using Apriori Algorithm, then from this association rules user will manually select the sensitive association rules SR. After selection of sensitive association rules SR apply MDSRRC. Once the hiding is done in original database D sanitized database D is produced. From this sanitized database D generate template table which will help for hiding the rules in the newly added database D1.

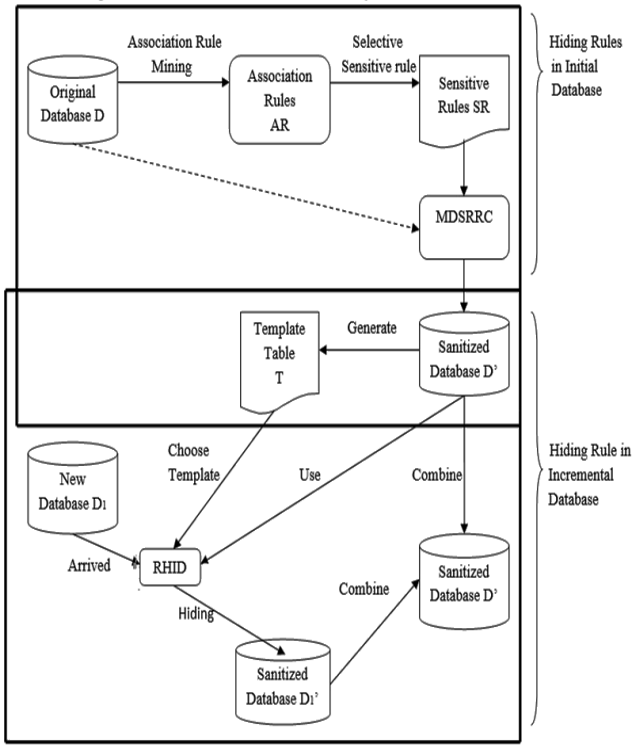

Fig 4: General framework for RHID algorithm [11]

For the second part, hiding is done using template table T by choosing the templates and calculating the support count of these templates from the new database D1. If the support count > MST (Minimum Support Threshold) then delete this template from new database with the help of RHID algorithm. This process is repeated till support court of templates is less than the defined MST. After this hiding process we get new sanitized database D1. Then finally combine new sanitized database D1 and sanitized database D to get the final Sanitized Database D, hence this sanitized database D can be realized to the outside world. In this way n numbers of incremental files can be added to this system and can generate the final Sanitized Database D which will be the combination of all Sanitized Database (D1+ D2+ D3..... Dn = D).

3. PROPOSED WORK

The proposed work is same like RHID algorithm used for incremental environment with some changes like using hybrid privacy preserving algorithm (combination of ISL and DSR) instead of MDSRRC and ARM for generating association rules. Now in the first part, the original dataset D is given to ARM the advanced Apriori algorithm to generate interesting patterns like correlation, association rules etc. From original dataset D as it takes less time to execute than traditional Apriori. Here association rules are generated using ARM. The generated association rules are than given to hybrid privacy approach which is a combination of ISL and DSR is used for hiding rule from both side (LHS and RHS) to generate the sanitized dataset D' and sensitive rules of that dataset D

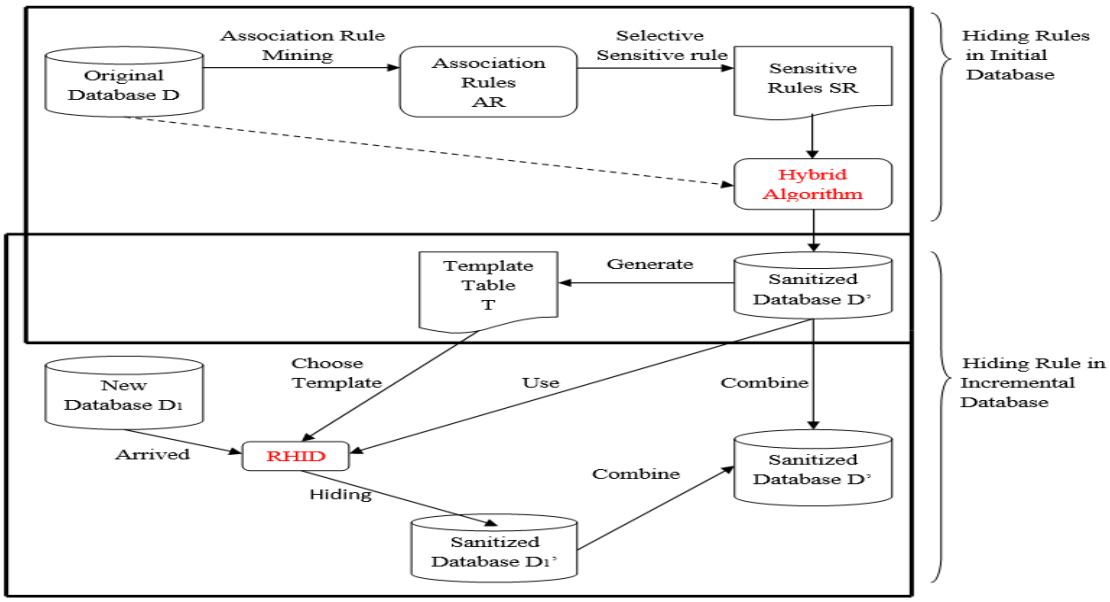

Fig 5: Proposed Work

Now, template table T is generated from the sanitized dataset D' which will help us for hiding the rules in the newly added database D1 for the hiding the sensitive rules. Now for the second part, hiding is done using template table T by choosing the templates and calculating the support count from the new database D1. If the support count > MST (Minimum Support Threshold) then we will delete this template from new database with the help of RHID algorithm. This process is repeated till support court of templates is less than the defined MST. After this hiding process we get new sanitized database D1'. Then finally combine new sanitized database D1' and sanitized database D' to get the final Sanitized Database D", hence this sanitized database D" can be realized to the outside world. In this way n numbers of incremental files can be added to this system and can generate the final Sanitized Database D" which will be the combination of all Sanitized Database (D1'+ D2'+ D3'..... Dn' = D").

3.1 Algorithm

Input: Sanitized Database D Template Table T, New Dataset D1, Sensitive Rules SR
Output: Sanitized Database D with both side sensitive rules hidden.

1. Generate Template Table from Sanitized Database D
2. Adding of New Dataset D1
3. for all frequent items in Template Table T
4. {
5. Count (items in Template table T) [Count (items of template table in New DatasetD1)]
6. Update MST Count (items in Template table T)
7. If (Count>MST)
8. {
9. Sort those in descending order of their count
10. Delete item with respect to Hybrid algorithm in Template Table T from New Dataset D
11. Go to step 3
12. }
13. }
14. Generate Sanitized Database D1
15. Combine D1 and D
16. Sanitized Database D is generated which can be released.

Let us see this proposed work with an example:

Table 2. Database

ID	Items
1	abcde
2	acd
3	abdfg
4	bcde
5	abd
6	cdefh
7	abcg
8	acde
9	acdh

Table 3. Incremental Database

TID	Items
1	aced
2	efgh
3	degh
4	acd
5	abcd
6	acd

Table 4. Sanitised Database

TID	Items
1	e
2	acd
3	abcdfg
4	e
5	abcd
6	aefh
7	abcdg
8	ae
9	acdh
10	aed
11	efgh
12	degh
13	ad
14	ad
15	ad

As an example of proposed algorithm, for a given database in Table 2, a minimum confidence = 40% and minimum support is of 33%. Table 2 and 3 shows original dataset D and incremental database respectively.There are number of association rule generated. From that we take four association rules as follows: A → B (60%), B → A (80%), C (70%), A → C (70%). Total Hidden Rule = 16 [B → A, C → A, D → A, A → B, A → C, A → D, C → B, D → B, B → C, B →D, D → C, E → C, C → D, C → E, E→ D, D→ E]. Final Sanitized database D which is form by merging sanitized table D of original Dataset D with sanitized table D1 of incremental Dataset D1. Table 4 shows the final sanitised database.

4. IMPLIMENTATION

All the algorithms are implemented in Java language and tested on the Intel(R) core(TM) i5-4210U CPU-2.430 GHz Windows 8.1 system with 4 gigabytes of main memory. In this paper, we take diabetic data. Dataset is composed of 768. Each patient is characterized in data set by 8 attributes. All attributes are numerical values.[12]

Table 5. Sample Dataset[12]

PREGE ONE	PREGE TWO	PLASL ONE	PRESR TWO	-	-	SKINS ONE	INSUI THREE	PEDIP ONE	AGEA ONE
1	0	1	0	-	-	1	0	1	1
1	0	1	0	-	-	1	0	1	1
1	0	1	0	-	-	1	0	1	1
1	0	1	0	-	-	1	0	1	1
-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-
1	0	1	0	-	-	1	0	1	1
1	0	1	0	-	-	1	0	1	1
1	0	1	0	-	-	1	0	1	1
1	0	1	0	-	-	1	0	1	1

First we compare the number of rule hide of algorithms by varying from 468 to 768.

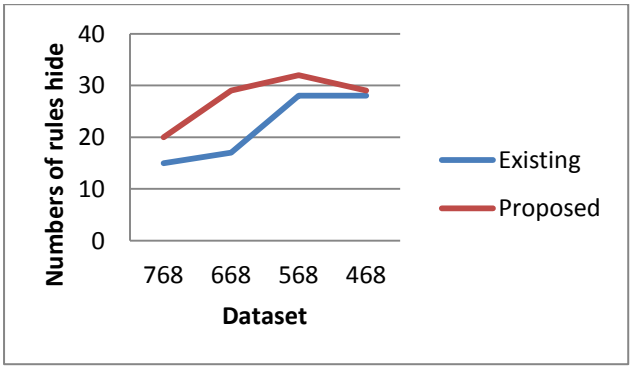

Fig 6: Numbers of rules hides

We use Table 6 compare different number of dataset and number of rule hide produced by proposed (Sensitive Association Rule Hiding Using Hybrid Algorithm in Incremental Environment) algorithm and existing algorithm (RHID algorithm using MDSRRC) under the selected value of rule hide. While support=33% and confidence=40%. From this comparison we can see that, proposed algorithm hide more rule than existing algorithm. This is because our proposed algorithm Hide LHS and RHS side of the rule. We use Table 7 compare different number of dataset and execution time produced by proposed (Sensitive Association Rule Hiding Using Hybrid Algorithm in Incremental Environment) algorithm and existing algorithm using MDSRRC) under the selected value of rule hide.

Table 6. Numbers of Rules hide

Dataset	Number of Rule Hide	
	Existing	Proposed
768	15	20
668	17	29
568	28	32
468	28	29

Table 7. Execution Time Require

Dataset	Execution Time (Second)	
	Existing	Proposed
768	28	13
668	25	20
568	23	16
468	21	10

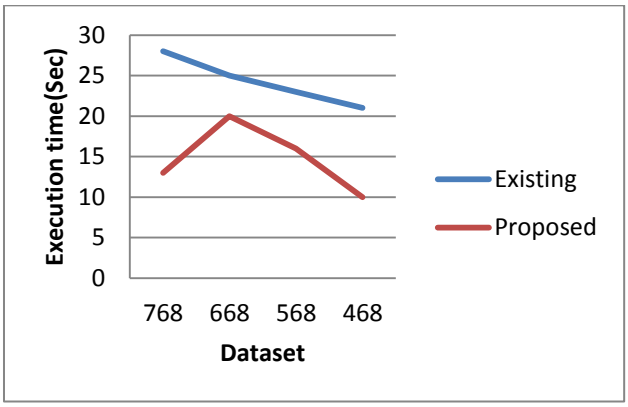

Fig 7: Execution Time

From above comparison we can see that, proposed algorithm takes a less time but more rule hide than existing algorithm. We use a novel approach for ARM using pattern generation that reduces the number of database scanning than the traditional apriori algorithm.

5. CONCLUSION

Traditional apriori algorithm take multiple times scanning of database. A novel approach for association rule mining using pattern generation is efficient with reduction in multiple times scanning of database and less memory space. Hybrid algorithm use ISL and DSR approach that increase and decrease of rule from left hand side and right hand side which gives better result than alone ISL and DSR. RHID algorithm is applicable for incremental environment. It can be conclude that Combine approach of hybrid and RHID algorithm is hide the sensitive rule from LHS and RHS side in an incremental environment. This algorithm can be converted into distributed algorithm so it can be used in the distributed environment.

6. REFERENCES

[1] Khyati B. Jadav,Jignesh Vania, Dhiren R. Patel, "A Survey on Association Rule Hiding Methods", International Journal of Computer Applications, ISSN: 09758887, Volume 82 No 13, November 2013.

[2] M. Atallah, E. Bertino, A. Elmagamind, M. Ibrahim, and V. S. Verykios \Disclo-sure limitation of sensitive rules",In Proc. of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX 1999), pp. 45-52.

[3] Sunil Kumar ,Mahaveer Singh, Nidhi Porwal, "An Algorithm for Hiding Association Rules on Data Mining", National Conference on Communication Technologies & its impact on Next Generation Computing CTNGC 2012 Proceedings publishedby International Journal of Computer Applications (IJCA)

[4] Vikash Shrivastava, Vivek Shrivastava, Vijay Patidar, A Generalized Association Rule Based Method for Privacy Preserving in Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 3, Issue 9, September 2013

[5] Kirtirajsinh Zala, Comparison of ISL, DSR, and New Variable Hiding Counter Algorithm of Association Rule Hiding", International Journal of Scienti_c & Engineering Research, ISSN 2229-5518, Volume 3, Issue 5, May-2012

[6] Chirag N. Modi,Udai Pratap Rao, Dhiren R. Patel, "Maintaining Privacy and Data Quality in Privacy Preserving Association Rule Mining", 2010 Second International conference on Computing, Communication and Networking Technologies

[7] V K S K Sai Vadapalli & G Loshma, "Secure Strategy for Privacy Preserving Association Rule Mining", International Conference on Computer Science and Engineering, April 28th, 2012, Vizag, ISBN: 978-93-81693-57-5

[8] Komal Shah, Amit Thakkar, Amit Ganatra, Association Rule Hiding by Heuristic Approach to Reduce Side E_ects & Hide Multiple R.H.S. Items", International Journal of Computer Applications, ISSN:0975 8887, Volume 45 No.1, May 2012.

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

[10] Pratiksha Sapkal, Minakshi Panchal, Manisha Pol, Madhumita Mane, "Hiding Sensitive items using MDSRRC to Maintain Privacy in Database", Int. Journal of Engineering Research and Applications ISSN : 2248-9622, Vol. 4, Issue 2(Version 1), February 2014, pp.623-627

[11] Vikram Garg, Anju Singh, Divakar Singh, "A Hybrid Algorithm for Association Rule Hiding using Representative Rule", International Journal of Computer Applications, ISSN: 0975 8887, Volume 97 No.9, July 2014

[12] Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.