Association Rule Hiding based on Heuristic Approach by Deleting Item at R.H.S. Side of Sensitive Rule

Divya C. Kalariya
Research Scholar
G.H. Patel College of Engineering and Technology.
Vallabh Vidyanagar, India

Vinita Shah
Asst. Prof., IT Dept.
G.H. Patel College of Engineering and Technology.
Vallabh Vidyanagar, India

Jay Vala
Asst. Prof., IT Dept.
G.H. Patel College of Engineering and Technology.
Vallabh Vidyanagar, India

ABSTRACT
Privacy preservation data mining is novel research area where data mining algorithms are analyzed for their side-effects they done on data privacy. Privacy preservation data mining (PPDM) deals with the problem of hiding the sensitive information while analyzing data. Many techniques are available for PPDM like data distortion, data hiding, rule hiding, data modification etc. Association rule hiding is one of the technique of PPDM. It hides sensitive rules which are generated by association rule generation algorithm before releasing database. This paper discusses different approaches of association rule hiding technique. In this paper, we propose a heuristic algorithm which provides privacy for sensitive rules while ensuring data quality. Proposed algorithm hides as many as possible rules at a time by modifying fewer transactions.

General Terms
Association rule generation, Privacy Preservation.

Keywords
Data mining, privacy preservation data mining (PPDM), Support, Confidence, Association rule hiding

1. INTRODUCTION
Data mining aims to extract hidden information from data warehouses. In data mining, different type of algorithms are used to extract different useful information from large amount of data. Algorithms are analyzed for their side effect which incur the data privacy. For example, using data mining algorithm on database, anyone can extract sensitive information like frequent pattern, association rules, unclassified data etc. It means data mining poses a threat to information privacy. To solve that problem, privacy preservation data mining concept is used in data mining and database security field.

Different techniques are used to solve PPDM. Association rule hiding is one technique of PPDM to hide sensitive rules which is generated by rule generation algorithm. Association rule generation algorithm is based on frequent items occurring in database. Frequent items mean the set of item which occurring together in a transaction. Finding that frequent items using different algorithm like apriori, FP growth tree. Generated rules are input in rule hiding algorithm, when applying rule hiding. Result of rule hiding algorithm is sanitized database which is not containing sensitive rules.
To understand the requirement of association rule hiding in PPDM, Here take one example which includes one cancer researcher and two hospitals. When researcher want to do survey on database of cancer hospital so researcher requests for the dataset of two hospitals i.e. A and B for review purpose. Both hospital A and B wants to hide the treatment related information based on symptoms from each other and also from researcher. So before giving database to researcher, both hospital use rule hiding technique to hide sensitive rules i.e. frequent symptoms→treatment 1. Output of association rule hiding algorithm is sanitized database which never generates the sensitive rule define by hospital. And this sanitized database is given to researcher for review approach.

2. PROBLEM DEFINITION
Dataset D is our input then AR is Association rule which is generated from input database D. If user want to hide some sensitive association rule (SR) selected by user then SR can be hidden by applying different rule hiding approaches which are discussed in section 4. Using approaches sanitized database D' can be generated. D' contains only the rules which are not present in SR (AR - SR). Rule hiding approach should try to maintain data quality of D' so dissimilarity between D' and D (D - D') should be as possible as lesser.

3. ASSOCIATION RULE MINING
Association rule mining firstly proposed by Agrawal et al in 1993[1]. An association rule is an implication expression of the form X→ Y, where X and Y are disjoint item sets, i.e., X ∩ Y =∅ .[1] Support and confidence are two basic parameter of association rule mining. Definition of support and confidence is defined below [2]:

Support is percentage of transactions in dataset that contain XUY.

$$Support(XY) = \frac{Total\ no\ of(XY)}{Total\ no\ of\ transaction\ in\ D} \quad \text{.....Eq1}$$

Confidence is the percentage of transactions in dataset containing X that also contain Y. Confidence show the conditional probability.

$$Confidance(XY) = \frac{support\ (XY)}{\text{support X}} \quad \text{.....} \quad \text{.Eq2}$$

Based on Minimum Support Threshold (MST) and Minimum Confidence Threshold (MCT) value, frequent item set and association rules are generated using different algorithm like apriori, FP growth.

If user want to hide the rules then he should try to decrease the confidence value of that rule compare to MCT. User can

do this by decreasing the value of confidence by increasing the value of denominator or by decreasing the value of numerator. And the value of denominator and numerator can be changed by altering the value of support count of Item sets. Altering the values of support count are based on different approach which are explained in next section.

4. ASSOCIATION RULE HIDING APPROACHES AND DIFFERENT ALGORITHMS

The concept of privacy preservation data mining has been recently proposed in response to the concerns of preserving privacy information from data mining algorithms. [3] Basically there are two type of privacy related to data mining which are output privacy and input privacy. Output privacy, means the data is minimally altered so that the mining result will not disclose certain privacy. [4] For output privacy, many technique are developed i.e. heuristic approach based technique like perturbation, blocking, swapping etc. Input privacy, is that the data is manipulated so that the mining result is not affected or minimally affected. [4] For input privacy, many techniques are developed i.e. cryptographic approach based technique like secure multiparty computation etc.

There are main five types of different approaches of association rule hiding which are related to input/output privacies are discussed in following table.

Table1: Comparison of Different Approaches

Approaches	Summary
Heuristic	High Efficiency, scalability and quick response. Totally takes best decision. Produce side effects like Lost of Rules, Artifactual Pattern in modified database
Crypto-graphic	Secure mining of association rule over partitioned Database. Do not protect the output of a computation.
Border Based	Maintains data quality by greedily selecting the modification with minimal side effects. Unable to identify optimal hiding solution
Re-construction	Lesser side effects in database than heuristic approach. limits the number of transactions in the sanitized Database.
Exact	Guarantees of quality. Requires very high time complexity due to integer programming

This summary conclude that heuristic approach is reliable approach then other.

5. HEURISTIC BASED APPROACH

This approach involves efficient, fast and scalable algorithms that selectively sanitize a set of transactions from the original database to hide the sensitive association rules [5].It is divided further in to two types that are Distortion techniques and blocking technique.

Distortion technique delete items by replacing 1-values to 0-values for reducing the confidence of rules or this technique add items by replacing 0-values to 1- values for reducing the support of rules. Sensitive rules are being hidden based on modification in database due to deleting or adding the items. Different algorithm are available for this approach. In [6], authors have presented three algorithms 1.a, 1.b and 2.a for hiding sensitive association rules. Algorithm 1.a inserts the items in transaction therefore increases the support value of L.H.S. side items in rules so confidence of that rule will be decreased automatically. Side effect of insertion new items in database is generation of new association rules. Algorithm 1.b and 2.a deletes the R.H.S. items of rules so confidence will decreased. Sometimes algorithm 1.b & 2.a affect the non-sensitive rules also. Two algorithms Increase Support of L.H.S (ISL) and Decrease Support of R.H.S (DSR) are proposed in [5]. In [7] Algorithm DCIS (Decrease Confidence by Increase Support) and DCDS (Decrease Confidence by Decrease Support) are proposed. ISL and DCIS based on item adding approach while DSR and DCDS is based on Item deleting approach. DSRRC (Decrease Support of R.H.S. item of Rule Clusters) is given in [8], which provides privacy for sensitive rules at certain level while ensuring data quality. DSRRC algorithm clusters the sensitive association rules based on R.H.S. of rules and hides as many as possible rules at a time by modifying fewer transactions. Algorithm DSRRC cannot hide rules having multiple RHS items. To solve the disadvantages of DSRRC algorithm, we proposed new heuristic based algorithm.

Table2: List of Algorithm

Approach	Algorithm	Conclusion
Insertion Based Algorithm(L.H.S.)	Algorithm 1.a	Large number of new rule generation and less number of rules are lost.
	ISL	
	DCIS	
Deletion Based Algorithms(R.H.S.)	Algorithm 2.a	Large number of rules are lost and less number of new rule generation.
	Algorithm 2.b	
	DSR	
	DCDS	
	DSRRC	

Blocking technique replaces an existing value to "?". This technique inserts unknown values in the data to fuzzify the rules. Sometimes providing wrong information to other is not acceptable. Adversary can easily find out the unknown value in sanitized dataset.

5.1 Proposed Heuristic Based Algorithm

In order to hide the sensitive rule like X→Y, we can decrease either confidence or support of the rule below the user specified minimum threshold. To decrease the confidence of the rule, we can choose two methods like (1) increase the support of X or (2) decreasing support of Y. Proposed algorithm hides rules with multiple items in L.H.S and multiple items in R.H.S. Proposed algorithm decrease the support of the R.H.S. and decrease the confidence of the rule below MCT. We replace '1' to '0' in some transaction to decrease the support of selected items.

Step of proposed algorithm
INPUT:
MCT (Minimum Confidence Threshold), Original database D, MST (Minimum support threshold).
OUTPUT:
Database D' with all sensitive rules are hidden.
1. Apply apriori algorithm on given database D. Generate all possible association rules R.
2. If User select Sensitive rules manually (SR€ R) then Go to step 5 otherwise go to step 3.

3. Enter the lift threshold value.
4. Calculate Lift value of all rules from R and consider rules as SR if its lift value is higher than lift threshold.

$$Lift(L \rightarrow R) = \frac{\text{Supp (L U R)}}{\text{Supp (L).Supp (R)}} \text{.....Eq3[16]}$$

5. Create itemset H1={h1,h2,.....,hn};
Where h=R.H.S. items of SR; H1 arranged in descending order of support h.
6. Calculate sensitivity of each item associate with first item h1 from H1.
4. Calculate sensitivity of each Transaction for h1.
5. Delete h1 from higher sensitive transaction.
6. Remove h1 from H1 and update H1(h1=h2,....,hn-1=hn).
7. Check, H1 is empty? IF H1 is not empty then go to step 6 otherwise go to step 8.
8. Updated database D.
9. Apply apriori algorithm on database D'. Generate all possible association rules R'.
10.Check, SR is present in R'? If present then go to step 5. otherwise go to step 11.
11. Take D' as output that is sanitized Database.

Proposed algorithm can provide automatic selection of sensitive rules and also hide the sensitive rules which contain more than one item at R.H.S. side.
To calculate the accuracy of this algorithm, following parameter are considered.

5.2 Performance Parameters

Evolution parameter are used to evaluating the performance of association rule hiding algorithms. Detailed description of parameter are discussed below. [12]

Dissimilarity:
It shows difference between original database and sanitized database.

$$\text{Diss}(D, D') = \frac{\sum|D(i) - D'(i)|}{D(i)} \text{.....Eq4.}$$

Lost Rules cost:
It measures the number of no sensitive association rules found in the original database but not in sanitized database.

$$\text{Lost Rules} = \frac{AR - AR'}{AR} \text{.....Eq5.}$$

Artifactual Patterns:
Artifactual pattern (AP), is measured in terms of the percentage of the discovered patterns that are artifacts. The formula is:

$$AP = \frac{AR' - |AR \cap AR'|}{AR'}$$

Hiding Failure:
The percentage of sensitive information that is still discovered, after the data has been sanitized.

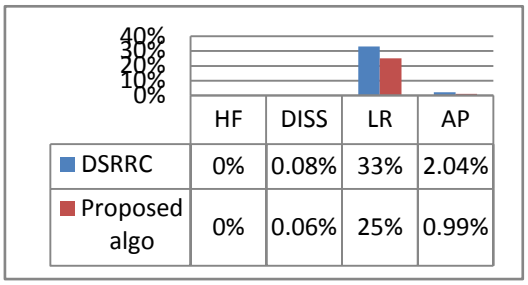
$$HF = \frac{SR'}{SR} \text{.....Eq6}$$

6. COMPARATIVE ANALYSIS

Comparing the result of proposed algorithm with exiting algorithm(DSRRC) based on accuracy parameter which are discussed earlier. In following table comparative analysis for different sensitive rules for supermarket[15] are describe.

Table3: Comparative analysis

Sensitive Rules(SR)	Accuracy Parameter	Existing Algorithm (DSRRC)	Proposed Algorithm
(SR1)backing-needs-> bread-and-cake; biscuits-> bread-and-cake; breakfast-food-> bread-and-cake	HF	0%	0%
	DISS	0.078%	0.061%
	LR	33%	25%
	AP	2.04%	0.99%
(SR2)backing-needs-> bread-and-cake; backing-needs-> vegetables; fruit->bread-and-cake, vegetables	HF	100%	0%
	DISS	-------	0.41%
	LR	-------	34.65%
	AP	------	0.0%



We can conclude that our proposed algorithm gives less dissimilarity ,less lost rule cost and less artifactual pattern compare to DSRRC algorithm. DSRRC algorithm is not applicable to rules which contain more then one R.H.S. item for exampleSR2.

7. CONCLUSION AND FUTURE WORK

Proposed Algorithm hides many sensitive association rules at a time while maintaining database quality. We have analyzed experimental results for Proposed Algorithm on different Dataset, which show that performance of the Proposed algorithm is better than Existing(DSRRC) algorithm on the bases of parameter those are Dissimilarity and lost rule cost. Existing algorithm hides only rules that contain single item on R.H.S. of the rule. Proposed algorithm can overcome this problem by hide sensitive rules which contain different more than one number of item at R.H.S. side.

For hiding sensitive association rules, sensitive rules are selected manually in existing algorithm. So, in case of larger dataset it takes more time to select rules manually. The proposed algorithm select the sensitive rules automatically using a LIFT value. It is experimented with sample data set. In future, the work can be done to find some method which is more accurate and do less modification in sanitized dataset.

8. REFERENCES

[1] Vikram Garg, Anju Singh & Divakar Singh "A Survey of Association Rule Hiding Algorithms" Fourth International Conference on Communication Systems and Network Technologies, IEEE, 2014, pp. 404-407.

[2] Komal Shah, Amit Thakkar & Amit Ganatra "A Study on Association Rule Hiding Approaches" International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-1, Issue-3, February 2012, pp. 72-76.

[3] Khyati B. Jadav, Jignesh Vania & Dhiren R. Patel "A Survey on Association Rule Hiding Methods"

International Journal of Computer Applications (0975 – 8887) Volume 82 – No 13, November 2013, pp. 20-25.

[4] R. Natarajan, Dr. R. Sugumar, M .Mahendran, K. Anbazhagan "Design and Implement an Association Rule hiding Algorithm for Privacy Preservation Data Mining" International Journal of Advanced Research in Computer and Communication Engineering(IJARCCE) Vol. 1, Issue 7, September 2012, pp. 486-492.

[5] Shyue-Liang Wang , Bhavesh Parikh, Ayat Jafari "Hiding informative association rule sets" Expert Systems with Applications 33, ELSEVIER, 2007, pp. 316-323.

[6] S. Kasthuri, T. Meyyappan, "Detection of Sensitive Items in Market Basket Database using Association Rule Mining for Privacy Preservation" International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME), IEEE, February 2013, pp. 200-203.

[7] F. Shahzad, s. Asghar, "Hiding Sequential Patterns Using FP Growth Technique" International Conference on Computer Networks and Information Technology (ICCNIT), 2011 IEEE, pp.125-129.

[8] Chirag N. Modi, Udai Pratap Rao, Dhiren R. Patel "Maintaining Privacy and Data Quality in Privacy Preservation Association Rule Mining" Second International conference on Computing, Communication and Networking Technologies, IEEE, 2010, pp.1-6.

[9] Mr. Pravin R. Ponde , Dr. S. M. Jagade (Ph. D) "Maintaining Privacy and Data Quality to Hide Sensitive items from Database", International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 3, Issue 7, July 2014, pp. 253-256.

[10] Shyue-Liang Wang , Bhavesh Parikh, Ayat Jafari "Hiding informative association rule sets" Expert Systems with Applications 33, ELSEVIER, 2007, pp. 316-323.

[11] Divya C. Kalariya, Vinita shah, Jay Vala," A Survey of Association Rule Hiding Approaches for Privacy Preservation Data mining"

[12] Charu C. Aggarwal, Philip S. Yu, Privacy-Preserving Data Mining:Models and Algorithms. Springer Publishing Company Incorporated,2008, pp. 267-286.

[13] Data mining Concepts and Techniques; Jiawei Han and Micheline Kamber; Second Edition, Morgan kaufmann publishers.

[14] Data Mining Techniques; Arun K Pujari; Universities Press.

[15] weka: http://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/supermarket.arff

[16] http://facweb.cs.depaul.edu/mobasher/classes/ect584/weka/associate.html