Development of an Association Rule Hiding Algorithm for Privacy Preserving in Market Basket Databases

Gaurav Kumar Ameta, Divya Bhatnagar

Abstract*: Association Rule Hiding is achieved by applying privacy preserving data mining techniques on a database. It becomes necessary before revealing the database to the third party. Major limitations of popular association rule hiding algorithms are failing into hiding all sensitive association rules, losses in terms of large number of non sensitive association rules, generation of ghost rules and false rules during the process of association rule hiding. These drawbacks have a greater impact the factors like privacy, correctness and usefulness of the sanitized database. Trustworthiness of inferences, conclusions and results extracted from sanitized database is affected. In this paper, an efficient algorithm named as selective flip bit was proposed as a solution for association rule hiding that hides all sensitive association rules by generating very less number of lost rules and zero generation of ghost rules and false rules. Developed algorithm was tested on both artifactual and real life databases. A software tool was developed to implement the selective flip bit algorithm on real life database. Another tool was also developed for the performance evaluation of developed rule hiding algorithm. Results indicate that the proposed selective flip bit algorithm is highly efficient in terms of hiding sensitive association rules along with retaining maximum non sensitive association rules as compared to the algorithms in the same field.

***Index Terms*: Association Rule Hiding, Privacy Preserving Data Mining, Sanitized Database, Selective Flip Bit Algorithm.**

I. INTRODUCTION

Databases belonging to the sectors like banking, healthcare, genetics, defence, education, travel and transport, retail and telecommunication etc. generally contains two types of information, sensitive and non sensitive. Sensitive information is that information which needs to be kept secret by the data publisher before exchanging the database to other parties. Non Sensitive information is intended for the receiving party, means no alteration is needed in non sensitive information and it should be easily accessible and inferable by database receiver.

Not even information; association rules in the databases may also be categorized as sensitive and non sensitive. Sensitive association rules are those rules which if applied by miner then it may leak sensitive information which can lead to leakage of secrecy of sensitive associations, secrecy of individual attributes and details about those attributes. These rules are categorized as sensitive association rules.

Non Sensitive Association rules are those rules which need not to be hidden from miner and their leakage is not risky from publisher's opinion. In other words, Non Sensitive rules are intended for the miner so that these rules can be used for analysis purpose by him without any risk.

Lost Rules are the association rules which lost after modifying the database. False Rules are the sensitive association rules which are not hidden by hiding algorithm and can be mined by applying mining algorithm on modified database. Ghost Rules are the rules which are not present in original database but generated after applying hiding algorithm.

Problem of protecting privacy in association rule mining can be stated as follows: If D is the source database of transactions and R is a set of relevant association rules that could be mined from D. The goal is to transform D into a database D' so that the most non sensitive association rules in R can still be mined from D' while others, representing sensitive knowledge, are hidden. D' is referred as the Transformed Database in this scenario. Refer Figure1 to visualize it.

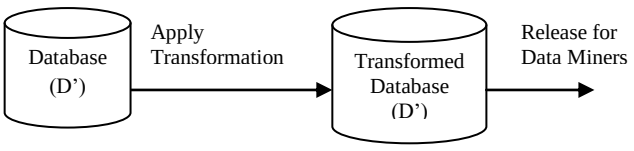

Fig. 1: Association Rule Hiding: Basic Concept

Main challenge is that in the process of hiding sensitive association rules, several non sensitive association rules are also gets compromised. Miner then tries on transformed database to extract fruitful information; he might not get accurate results. This context is associated directly to the NP-Hard Problem. So, in an ideal privacy preserving process all sensitive association rules are hidden along with the condition that all non sensitive rules remain open for analysis. But practically it is not possible; especially

Revised Manuscript Received on December 22, 2018.

Gaurav Kumar Ameta, Department of Computer Science & Engineering, Sir Padampat Singhania University, Udaipur, India.
Divya Bhatnagar , Department of Computer Science & Engineering, Sir Padampat Singhania University, Udaipur, India..

in case if some common items exist in both sensitive and non sensitive association rules in a single database. For ex. Bread ==>Beer comes under category of Sensitive Association rule but Bread ==>Butter is non sensitive association rule.

In such cases maintaining privacy of sensitive association rules and disclosing all non sensitive association rules become very complex for publisher. To handle such situation publisher adopt some techniques for minimizing leakage of sensitive association rules and maximizing releasing of non sensitive association rules. There are several approaches which can be used by the publisher like Heuristic Approach, Border Approach, Exact Approach, Reconstruction Approach, Cryptographic Technique Approach and Hybrid Technique Approach.

As per Figure 2, there is an Original Database 'D'. Publisher generates several kinds of association rules pertaining to the criteria of minimum support and minimum confidence. After that sensitive association rules are determined. After marking of sensitive association rules, Association rule hiding algorithm is applied. Finally, a reconstructed database is received which is ready for use by other miners.

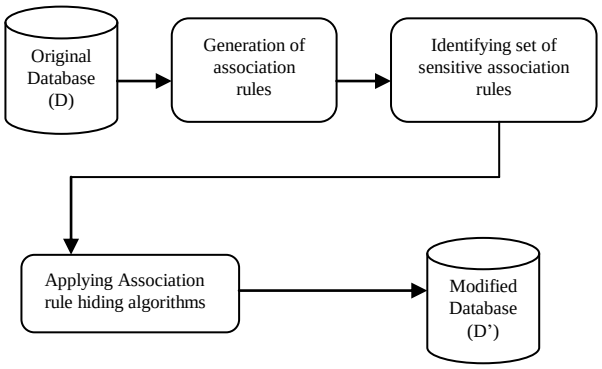

Fig. 2: Framework for Hiding Sensitive Association Rules

Association rule hiding method is dependent on the two parameters support or confidence. Majorly two ways are available to hide any sensitive rule, either increase or decrease the support up to specific level or decrease confidence up to certain threshold. Modifications performed on the database may lead to some side effects that may cause side effects in terms of lost rules, ghost rules and false rules.

Data categorization was discussed in which data is divided into sensitive and non sensitive objects. Access techniques like normal mode and sensitive mode are discussed [1]. Various areas of privacy preserving data mining and algorithms were investigated. Methods are discussed for distributed privacy preserving mining and for horizontally and vertically partitioned data. Issue of degrading the effectiveness of the techniques related to data mining is also discussed [2]. Introductory techniques by which privacy can be preserved are discussed like suppression, randomization, summarization and cryptography [3]. Two algorithms Increase Support Left (ISL) and Decrease Support Right (DSR) for association rule hiding has been discussed. Improved algorithm was discovered and proven better in terms of number of rules hided, processing time and the number of entries updated in the database [4]. A new approach Remove and Reinsert L.H.S. of Rule (RRLR) was discovered for sensitive association rule hiding. It was produced as an improved as compared to approaches like Decreased Support of Right Hand Side item of Rule Cluster (DSRRC) and Advanced Decreased Support of Right Hand Side item of Rule Cluster (ADSRRC). Both improved algorithms are able to hide all sensitive association rules. In RRLR algorithm quantity of lost rules are 22.73% as compared to DSRRC and ADSRR algorithms in which 36.36% non sensitive association rules are lost [5]. Privacy preserving algorithm using suppression technique was elaborated. In this database used by publisher or researchers must be sanitized before exchanging with third party [6]. A new approach in which more focus is given on identifying rare data correlations. Such correlations are considered to be more interesting as compare to mining frequent itemsets. [7]. A novel approach was developed for hiding sensitive association rules in binary transactional databases. Algorithm NDSRRC is developed by authors which emphasizes on rule sensitivity. Developed algorithm was compared with already existing algorithms like DSSRC and MDSRRC [8]. During the process of privacy preserving and hiding sensitive association rules, side effects are occurred in terms of hiding failures, misses cost, artifactual pattern and support loss in the database during modification process. Database modification techniques were discussed for association rule hiding are heuristic based, border based, exact, reconstruction based and cryptography based [9]. Five techniques are summarized named as Heuristic Approach, Border Based Approach, Cryptographic Approach, Exact Approach and Reconstruction Approach used for association rule hiding. More emphasize was to show a need of developing some hybrid algorithms and to decrease modifications in the database during the process of association rule hiding [10]. Privacy preserving techniques were applied on medical databases. Medical dataset generally contains personal information about the patients and their related disease. Revealing of such information may harm to identity of an individual. Techniques discussed in this paper are generalization, bucketization, slicing [11]. Issue of privacy breach through social media was pointed out. People are using social media without being aware of privacy factor. Anybody can extract personal information of anyone through social media platforms. Privacy breaches are divided into three categories as sensitive link disclosure, sensitive attribute disclosure and identity disclosure [12]. A hybrid method was discovered for privacy preserving before publishing the data. Medical database is taken here for reference, in which attributes are divided into three categories quasi attributes, key attributes and sensitive attributes. [13]. Individual privacy and collective privacy were addressed.

Four types of perturbation are elaborated named as additive perturbation, multiplicative perturbation, rotation perturbation and geometric perturbation. Possible attacks on published databases are also introduced as background knowledge attack, minimality attack, unsorted matching attack, temporal attack and homogeneity attacks [14].

II. METHODOLOGY

An improved association rule hiding algorithm named as Selective Flip Bit was developed for boolean market basket database. In this method database is collected, preprocessed and converted into boolean form. '1' and '0' represents to the presence and absence of an item in a database respectively.

A. Selective flip bit method

In the proposed Selective Flip Bit method only specific bits of items containing sensitive association rules are flipped from '1' to '0' in order to hide sensitive association rules. Bits to be flipped are selected after sensitivity calculation among sensitive association rules.

To preserve privacy of sensitive association rules, algorithm is developed which flip bits from presence '1' to absence '0' and not from absence '0' to presence '1', so that quantity of false association rules becomes zero. Flip bit method algorithm works in such a manner that only selected '1' values are perturbed to '0'. This method leads to very less number of lost rules, false rules and ghost rules. Maximum number of non sensitive association rules could be retained along with hiding all sensitive association rules. Algorithm will work as follows:

(i) In this method first a bit is flipped from (1→0) for most sensitive attribute in the most sensitive transaction.

(ii) Check, if all sensitive association rules in transactions are hidden at this point due to propagation effect, the algorithm is stopped.

OR

Sensitive association is prevented from coming into the extractable rules at predefined threshold.

(iii) Otherwise, remaining most sensitive unhidden association rule is identified and bit is flipped for the most sensitive attribute in that particular association in the transaction.

(iv) This process is continued until all sensitive association rules in all the transactions are completely saved.

Large number of perturbation of bits from '1' to '0' leads to higher lost rules and large number of '0' to '1' conversion of bits generates large number of false associations and wrong entries in the modified boolean database. Both things degrade database accuracy and quality.

B. Algorithm implementation for test case

Following associations of items purchased in five transactions are marked as sensitive associations (SA):

Besan, Chana Dal, Poha	SA 1
Ankur Groundnut Oil, Patanjali Dant Kanti	SA 2
Chana Dal, Moong Mogar	SA 3
Ankur Groundnut Oil, Chana Dal, Parle G, Parle Hide & Seek	SA 4
Besan, Poha	SA 5

In above test case items are intentionally selected in such a manner that they will generate same and higher sensitivity values for two most sensitive associations. This case was used as a base test case to implement and test the developed algorithms.

Table I was used to represent the boolean transactional database of Items participating in sensitive associations SA 1 to SA 5. Items are represented as I1→Ankur Groundnut Oil, I2→Besan, I3→Chana Dal, I4→Moong Mogar, I5→Parle G, I6→Parle Hide & Seek, I7→Patanjali Dant Kanti and I8→Poha. Transaction varies from T1 to T5.

Table I: Boolean Database

	I1	I2	I3	I4	I5	I6	I7	I8
T1	0	1	1	0	0	0	0	1
T2	1	0	0	0	0	0	1	1
T3	1	1	1	0	0	0	1	1
T4	1	1	1	1	1	1	0	1
T5	0	1	0	0	0	0	0	1

Sensitivity calculations are done for above sensitive associations by the tool as shown in Table II and Table III. Table II elaborates the sensitivity of an individual item in sensitive association rule.

Table II: Sensitivity Percentage of each item in sensitive association rule

Name of Item	Frequency in Sensitive Association Rule	Sensitivity Percentage
Besan	2	15.38
Chana Dal	3	23.07
Poha	2	15.38
Ankur Groundnut Oil	2	15.38
Patanjali Dant Kanti	1	7.69
Moong Mogar	1	7.69
Parle G	1	7.69
Parle Hide & Seek	1	7.69
	Total Count=13	

To calculate sensitivity percentage for the sensitive items in sensitive association rules, following formula is used.

Sensitivity Percentage = (Frequency of an Item in all Sensitive Association Rules/ Total Count)*100

As per calculation in Table III, two most sensitive associations as per sensitivity calculations are SA 1 and SA 4. Both are representing same sensitivity values 53.84. It is clear that SA 1 and SA 4 are two most sensitive associations. But only one association out of these two has to be selected first for application of selective flip bit method.

Table III: Sensitivity calculation to determine most sensitive association

Transaction Id	Sensitive Association	Sensitivity	Unaffected Rules Count	Strength of left over rule
1	SA 1	53.84	1	23.07
7	SA 4	53.84	1	30.76
6	SA 3	30.77	0	0
12	SA 5	30.77	0	0
2	SA 2	23.07	0	0

To decide this both sensitive associations are tested individually. Generalization approach is used to test the strength of both the equally sensitive rules. At first all attributes of SA 1 are generalized in order to hide all sensitive association rules. Besan is replaced by Lentil Flour, Chana Dal is replaced by Lentils and Poha is replaced by Flakes. Due to this generalization four sensitive rules SA 1, SA 3, SA 4 and SA 5 were saved out of five. Only one rule SA 2 is remaining. Text in Bold shows the rule generalized SA 1 and the text in Bold and Italic shows that items are generalized due to the propagation effect occurred through the generalization of rule.

Lentil Flour, Lentils, **Flakes** SA 1(Safe)

Ankur Groundnut Oil, Patanjali Dant Kanti SA 2 (Unsafe)

Lentils, Moong Mogar SA 3 (Safe)

Ankur Groundnut Oil, ***Lentils***, Parle G, Parle Hide & Seek SA 4 (Safe)

Lentil Flour, Flakes SA 5(Safe)

This testing is halted here and now the impact is tested by doing generalization for the equal sensitive rule SA 4.

Besan, ***Lentils***, Poha SA 1(Safe)

Groundnut Oil, Patanjali Dant Kanti SA 2(Safe)

Lentils, Moong Mogar SA 3(Safe)

Groundnut Oil, Lentils, Glucose Biscuit, Chocolate Biscuit SA 4(Safe)

Besan, Poha SA 5(Unsafe)

Strength of remaining one unsafe rule is checked. Strength refers to sum of attribute sensitivity values in remaining association rule. It is 23.07 for SA 1 generalization. If same procedure is adopted for SA 4 generalization, strength of remaining rule is 30.76. Remaining rule which having lower strength value is preferred first for flip bit. Lower value implicates that particular generalization is more powerful. So, sensitive association SA 1 is taken first for rule hiding because of its lesser strength value for Selective Flip Bit as per Table III.

Bit of most sensitive item in most sensitive association SA 1 is flipped first. After that effect of this change is checked on remaining sensitive rules. Total five bits are flipped in process to hide all five sensitive association rules. In SA 1, Chana Dal is most sensitive item. Bit of Chana Dal is flipped thrice to save three associations SA 1, SA 4 and SA 3. Then one bit of Poha and Ankur Groundnut is flipped respectively to save remaining two associations SA 5 and SA 2. Table IV represents that total Five highlighted bits are flipped in order to save all sensitive association rules as per the developed selective flip bit algorithm.

Table IV: Selective Flip Bit output for Test Case

	I1	I2	I3	I4	I5	I6	I7	I8
T1	0	1	0	0	0	0	0	1
T2	0	0	0	0	0	0	1	1
T3	1	1	0	0	0	0	1	1
T4	1	1	0	1	1	1	0	1
T5	0	1	0	0	0	0	0	0

C. Algorithm implementation on real data set

A real market basket database was collected from a famous retail store of the city. Database initially consisted of 757 items and 221 transactions. The database collected was in printable textual format. It was transformed to .csv format. After preprocessing a relevant dataset was extracted for implementation of the developed selective flip bit algorithm for privacy preservation. The association rule mining was implemented on this dataset using WEKA 3.8.1 (Waikato Environment for Knowledge Analysis). Apriori algorithm in WEKA was applied to both original database and the dataset obtained after applying selective flip bit algorithm. Algorithm was implemented on varying confidence values of 0.9, 0.8, 0.7 and 0.6. Based on the supplied confidence value, the best N association rules were generated.

First a dataset is supplied to WEKA tool for determination of best 10 association rules. After determination of these sensitive association rules were marked. Same dataset is supplied to developed Association Rule Hiding (ARH) tool. ARH tool was developed to apply association rule hiding algorithms on the database. Developed Selective Flip Bit algorithm was applied through the ARH tool and changes were made in the original database accordingly. This database is known as modified database. Modified database is again supplied to WEKA to check the efficiency of the changes done. It is done to

become assure that all sensitive association rules are eliminated from appearing into best 10, 20, 30 and 40 rules after WEKA analysis.

It is an iterative process in which WEKA and ARH Tool are used alternatively to mine the rules and hide sensitive rules respectively. WEKA is used to check the best 10, 20, 30 and 40 rules. After applying one operation at a time database is tested through WEKA to verify that a particular sensitive association has been circumvented from coming into best 10, 20, 30 and 40 rules or not? This process is continued until all sensitive association rules are prevented from coming into best 10, 20, 30 and 40 rules.

III. RESULTS

Best rules generated by WEKA at minimum confidence values 0.9, 0.8, 0.7 and 0.6 are given in Table V. These rules will be considered as sensitive association rules. To hide these association rules Flip Bit Method was applied by using Association Rule Hiding Tool. Results shows that total 2, 9, 42 and 35 are flipped at minimum confidence values 0.9, 0.8, 0.7 and 0.6 respectively in order to hide all sensitive association rules from coming into best 10, 20, 30 and 40 association rules.

It can be concluded from Table VI that with the newly developed Selective Flip Bit approach, the total association rules retained after privacy preservation ranges from 93.52% to 99.67% which is better as compared to the other popular approaches for privacy preservation and association rule (AR) hiding like generalization, data distortion, blocking, suppression etc. Algorithm available for association rule

Table V: Selective Flip Bit Method Output

Minimum Conf.	Sensitive Rules	Flip bits required to hide all sensitive association rules
0.9	1. Chana Dal, Saras Pure Ghee, Sugar=>Poha 2. Sahakar Sing Dana=> Sugar	2
0.8	1. Parle Hide & Seek, Poha=>Sugar 2. Poha,Tata Salt=>Sugar 3. Parle Hide & Seek, Sugar=>Poha	9
0.7	1. Poha=>Sugar 2. Sugar=>Poha	42
0.6	1. Poha=>Sugar 2. Sugar=>Poha	35

hiding like ISL and DSR are not efficient and complete. Both of these algorithms are not able to hide all sensitive association rules. In both algorithms quantity of ghost rules, lost rules and false rules are higher. A Permutation tool was developed to calculate lost rules, false rules and ghost rules generated during the process of sensitive association rule hiding. Algorithms like DSRRC, ADSRRC and RRLR were developed as an improvement of ISL and DSR. These algorithms are capable to hide all sensitive association rules, but quantity of lost rules are higher 36.36%, 36.36% and 22.73% as compared to selective flip bit algorithm, in which quantity of lost rules ranges from 0.31% to 5.62%.

Table VI: Analysis of Results Generated by Selective Flip Bit Method

Selective Flip Bit Method						
Min. Conf.	Total AR generated from Original File	Total AR generated after Selective Flip Bit	Retained AR (No Change in Support Count)	Lost Rules	Support Count decreased AR	Total AR retained
0.9	43104	42968	96.58%	0.31%	3.09%	99.67%
0.8	43104	40370	75.53%	5.50%	18.95%	94.48%
0.7	43104	40312	72.24%	6.47%	21.28%	93.52%
0.6	43104	40682	77.16%	5.62%	17.22%	94.38%

IV. CONCLUSION

Moving ahead from traditional rule hiding algorithms like ISL, DSR, DSRRC, ADSRRC and its several variants a new approach is developed. The optimized algorithm based on selective flip bit method was proposed. It works in highly efficient manner. It can be clearly observed from Table VI that the range of retained non sensitive association rules is from 93.52% to 99.67% in modified database. It is important to note that maximum numbers of non sensitive rules are prevented from being lost during the process of association rule hiding along with the achievement of hiding 100% of marked sensitive association rules. Quantity of ghost rules and false rules generated after completion of this process is zero. Reason behind being zero of these factors is that bits are flipped from '1' to '0' means presence to absence not from '0' to '1' absence to presence.

Lesser values of ghost rules and false rules indicate that modified database will be of good quality in terms of accuracy Data user or miner should be able to extract maximum correct inferences from the modified database which were intended to him. Sensitive inferences and conclusions which were not intended to the data miner were already preserved by the developed algorithm.

ACKNOWLEDGMENT

We thank to all authors and researchers of several technical papers for sharing their work without which it was very difficult to achieve successful completion. We also thank the department of CSE, Sir Padampat Singhania University, Udaipur for providing us with all necessary laboratory facilities required.

REFERENCES

1. Vasudevan L., Sukanya S. and Aarthi N., Privacy Preserving Data Mining using Cryptographic Role Based Access Control Approach, Proc. International MultiConference of Engineers and Computer Scientists., Hong Kong. March 19 – 21, (2008), p. 474.
2. Aggarwal C. & Yu P. (Eds.), General Survey of Privacy Preserving Data Mining Models and Algorithms. Retrieved from DOI 10.1007/978-0-387-70992-5 (2008).
3. Evfimievski A. and Grandison T., Privacy-Preserving Data Mining, IGI Global, P(1), (2009).
4. Jain Y., Yadav V. and Panday G., An Efficient Association Rule Hiding Algorithm for Privacy Preserving Data Mining, International Journal on Computer Science and Engineering, 3(7), 2792 (2011).
5. Shah A. and Gulati R., Privacy Preserving Data Mining: Techniques, Classification and Implications - A Survey, International Journal of Computer Applications, 137(12), 40 (2016).
6. Ebin M. and Brilley C., Privacy Preserving Suppression Algorithm for Anonymous Databases, International Journal of Science & Research, 2(1), 5 (2013).
7. Cagliero L. and Garza P., Infrequent Weighted Itemset Mining Using Frequent Pattern Growth, IEEE Transactions on Knowledge and Data Engineering, 26(4), 903 (2014).
8. Sharad A. and Singh S., Preserving Data Privacy by Susceptible Association Rule Hiding Approach, International Journal of Computer Engineering and Applications, 7(1), 41 (2014).
9. Fouladfar M. and Dehkordi M., A Survey on the Privacy Preserving Algorithm and techniques of Association Rule Mining, Advances in Computer Science: an International Journal, 4(4), 1 (2016).
10. Aleem G., Ellatif L. and Sharf A., Association Rules Hiding for Privacy Preserving Data Mining: A Survey, International Journal of Computer Applications, 150(12), 34 (2016).
11. Rakshatha V. and Salian S., Hiding Personal Detail using Overlapping Slicing, American Journal of Intelligent Systems, 7(3), 50 (2017).
12. Mayil S., Vanitha M., A Review on Privacy Preserving in Social Network, International Journal of Scientific & Engineering Research, 8(1), 1034 (2017).
13. Saranya N., Karpagam M. and Muruganandham N., Hybrid Approach for Data Publishing Using Privacy Preservation Techniques, ARPN Journal of Engineering and Applied Sciences, 12(1), 155 (2017).
14. Chandrakanth P. and Anbarasi S., Additive Data Perturbation Approach for Privacy Preserving Data Mining, International Journal of Current Engineering and Scientific Research, 5(2), 9 (2018).

AUTHORS PROFILE

Gaurav Kumar Ameta is a Research Scholar in Department of Computer Science & Engineering at Sir Padampat Singhania University,Udaipur, India. His area of specialization is Data Mining, Computer Graphics, Cryptography and Privacy Preservation. He is a Lifetime Member of Computer Society of India.

Divya Bhatnagar is working as a Professor and Head of Department., Computer Science & Engineering at Sir Padampat Singhania University, Udaipur. She holds 19 years of teaching experience. Her specialization areas include Data Mining and Neural Networks.