A Study on Association Rule Hiding Approaches

Komal Shah, Asmit Thakkar, Amit Ganatra

Abstract*: In recent years, data mining is a popular analysis tool to extract knowledge from collection of large amount of data. One of the great challenges of data mining is finding hidden patterns without revealing sensitive information. Privacy preservation data mining (PPDM) is answer to such challenges. It is a major research area for protecting sensitive data or knowledge while data mining techniques can still be applied efficiently. Association rule hiding is one of the techniques of PPDM to protect the association rules generated by association rule mining. In this paper, we provide a survey of association rule hiding methods for privacy preservation. Various algorithms have been designed for it in recent years. In this paper, we summarize them and survey current existing techniques for association rule hiding.

Index Terms*: Association Rule Hiding, Data Mining, Privacy Preservation Data Mining.

I. INTRODUCTION

Data mining technology aims to find useful patterns from large amount of data. These patterns represent knowledge and are expressed in decision trees, clusters or association rules. The knowledge discovered by various data mining techniques may contain private information about individual or business. Revelation of any private information may cause threat to security. For example, in medical database, it is useful to share information about diseases but at the same time it is required to preserve patient's identity. Here individual privacy must be maintained. Another example is market basket database which is used to analyze customer's purchasing behavior represented in terms of association rules. In market basket database, instead of data related to individuals, the sensitive information or knowledge derived from data is required to be protected.

The sharing of data and/or knowledge may come at a cost to privacy, primarily due to two main reasons: (1) if the data refers to individuals then its disclosure can violate the privacy of the individuals who are recorded in the data if their identity is revealed to untrusted third parties or if sensitive knowledge about them can be mined from the data, and (2) if the data regards business information, then disclosure of this data or any knowledge extracted from the data may potentially reveal sensitive trade secrets, whose knowledge can provide a significant advantage to business competitors and thus can cause the data owner to lose business over his or her peers [1].

Privacy preservation data mining (PPDM) considers problem of maintaining privacy of data and knowledge in data mining. It allows extracting relevant knowledge and prevents sensitive data or information from disclosure. PPDM algorithms are developed for modifying the original data in such that the no sensitive information is revealed even after mining process. Association rule hiding is one of the privacy preservation techniques to hide sensitive association rules. All association rule hiding algorithm aims to minimally modify the original database such that no sensitive association rule is derived from it.

In next section we described the association rule mining strategy. In section 3, we have explained approaches to association rule hiding and related work in detail. Section 4 discusses about limitations of association rule hiding approaches. Section 5 contains analysis of existing algorithms on association rule hiding based on heuristic approach.

II. ASSOCIATION RULE MINING STRATEGY

Let I = {i1,…., in} be a set of items. Let D be a set of transactions or database. Each transaction t Є D is an item set such that t is a proper subset of I. A transaction t supports X, a set of items in I, if X is a proper subset of t. Assume that the items in a transaction or an item set are sorted in lexicographic order.

An association rule is an implication of the form X→Y, where X and Y are subsets of I and X∩Y= Ø. The support of rule X→Y can be computed by the following equation: Support(X→Y) = |X→Y| / |D|, where |X→Y| denotes the number of transactions in the database that contains the itemset XY, and |D| denotes the number of the transactions in the database D. The confidence of rule is calculated by following equation: Confidence(X→Y) = |X→Y|/|X|, where |X| is number of transactions in database D that contains itemset X. A rule X→Y is strong if support(X→Y) ≥ min_support and confidence(X→Y) ≥ min_confidence, where min_support and min_confidence are two given minimum thresholds.

Association rule mining algorithms scan the database of transactions and calculate the support and confidence of the rules and retrieve only those rules having support and confidence higher than the user specified minimum support and confidence threshold. Association rule hiding algorithms prevents the sensitive rules from being disclosed. The problem can be stated as follows: "Given a transactional database D,

Revised Manuscript Received on February 05, 2012.

Komal Shah, U and P.U. Patel Department of Computer Engineering, Chandubhai S Patel Institute of Technology, Changa, India, (e-mail: wishing_kom@yahoo.co.in).

Amit Thakkar, Department of Information Technology, Chandubhai S Patel Institute of Technology, Changa, India, (e-mail: amitthakkar.it@ecchanga.ac.in).

Amit Ganatra, U and P.U. Patel Department of Computer Engineering, Chandubhai S Patel Institute of Technology, Changa, India, (e-mail: amitganatra.ce@charusat.ac.in).

minimum confidence, minimum support and a set R of rules mined from database D. A subset RH of R is denoted as set of sensitive association rules which are to be hidden. The objective is to transform D into a database D‟ in such a way that no association rule in RH will be mined and all non sensitive rules in R could still be mined from D‟.

III. ASSOCIATION RULE HIDING APPROACHES & RELATED WORK

A. Heuristic 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 [1]. Various heuristic algorithms are based on mainly two techniques: (1) Data distortion technique (2) Blocking technique.

Data distortion is done by the alteration of an attribute value by a new value. It changes 1's to 0's or vice versa in selected transactions. There are two basic approaches for rule hiding in data distortion based technique: Reduce the confidence of rules and reduce the support of rules.

Consider sample database given in Table I. Selecting minimum support = 20% and minimum confidence = 80% and applying association rule mining algorithm, two association rules AB→C (confidence = 100%) and BC→A (confidence= 100%) are mined. Now suppose rule AB→C is sensitive and needs to be hidden. Decreasing the confidence of a rule AB→C can be done by either increasing the support of AB in transactions not supporting C (as shown in Table II) or by decreasing the support of C in transactions supporting both AB and C (as shown in Table III). Decreasing support of rule AB→C can be done by decreasing the support of the corresponding large itemset ABC (as shown in Table IV).

The problem for finding an optimal sanitization to a database against association rule analysis has been proven to be NP-Hard [2]. In [3], authors presented three algorithms 1.a, 1.b and 2.a for hiding sensitive association rules. Algorithm 1.a hides association rules by increasing the support of the rule's antecedent until the rule confidence decreases below the minimum confidence threshold. Algorithm 1.b hides sensitive rules by decreasing the frequency of the consequent until either the confidence or the support of the rule is below the threshold. Algorithm 2.a decreases the support of the sensitive rules until either their confidence is below the minimum confidence threshold or their support is below the minimum support threshold. In 1.a algorithm large number of new frequent itemsets is introduced and, therefore, an increasing number of new rules are generated. Algorithm 1.b and 2.a affects number of nonsensitive rules in database due to removal of items from transaction [3].

In [4] two algorithms are proposed ISL (Increase Support of LHS) and DSR (Decrease Support of RHS). Predicting items are given as input for both algorithms to automatically hide sensitive association rules without pre-mining and selection of hidden rules. In [5] two algorithms, DCIS (Decrease Confidence by **I**ncrease **S**upport) and DCDS (Decrease Confidence by Decrease **S**upport) are proposed to automatically hide collaborative recommendation association rules without pre-mining and selection of hidden rules. The ISL and DCIS algorithms try to increase the support of left hand side of the rule and algorithms DSR and DCDS try to decrease the support of the right hand side of the rule. It is observed that ISL requires more running time than DSR. Also both algorithm exhibit contrasting side effects. DSR algorithm shows no hiding failure (0%), few new rules (5%) and some lost rules (11%). ISL algorithm shows some hiding failure (12.9%), many new rules (33%) and no lost rule (0%). Algorithm DCIS requires more running time than DCDS. Similar to ISL and DSR, algorithms DCIS and DCDS also exhibit contrasting side effects. DCDS algorithm shows no hiding failure (0%), few new rules (1%) and some lost rules (4%). DCIS algorithm shows no hiding failure (0%), many new rules (75%) and no lost rule (0%).

In [6] an algorithm DSC (Decrease Support and Confidence) is proposed in which pattern-inversion tree is used to store related information so that only one scan of database is required. The proposed algorithm can automatically sanitize informative rule sets without pre-mining and selection of a class of rules under one database scan. There are about 4% of new rules generated and about 9% of rules are lost on the average for DSC algorithm and it also shows hiding failure for two predicting items.

TABLE I. SAMPLE DATABASE

TID	Items
1	A, B, C
2	A, B, C
3	A, C
4	A, E
5	C, D

Rule	Confidence
AB→C	100%
BC→A	100%

TABLE II. HIDING AB→C BY INCREASING SUPPORT OF AB

TID	Items
1	A, B, C
2	A, B, C
3	A, C
4	A, **B**, E
5	C, D

Rule	Confidence
AB→C	66%
BC→A	100%

TABLE III. HIDING AB→C BY DECREASING SUPPORT OF C

TID	Items
1	A, B
2	A, B, C
3	A, C
4	A, E
5	C, D

Rule	Confidence
AB→C	50%
BC→A	100%

TABLE IV. HIDING AB→C BY DECREASING SUPPORT OF ABC

TID	Items
1	A, C
2	A, B
3	A, C
4	A, E
5	C, D

Rule	Confidence
AB→C	0%
BC→A	0%

TABLE V. HIDING A→C BY BLOCKING

A	B	C	D
1	1	1	0
1	0	1	0
0	1	0	1
1	1	1	0
1	0	1	1

→

A	B	C	D
1	1	1	0
1	0	?	0
?	1	0	1
1	1	1	0
1	0	1	1

In [7] authors proposed an efficient algorithm, FHSAR (Fast Hiding Sensitive Association Rules), for fast hiding sensitive association rules. The algorithm can completely hide any given sensitive association rule by scanning database only once, which significantly reduces the execution time. In this algorithm correlations between the sensitive association rules and each transaction in the original database are analyzed, which can effectively select the proper item to modify.

In [9] four heuristic algorithms are proposed: Algorithm Naïve, MinFIA (Minimum Frequency Item Algorithm), MaxFIA (Maximum Frequency Item Algorithm) and IGA (Item Grouping algorithm). Each algorithm selects the sensitive transactions to sanitize based on degree of conflict. Naive Algorithm removes all items of selected transaction except for the item with the highest frequency in the database. The MinFIA algorithm selects item with the smallest support in the pattern as a victim item and it removes the victim item from the sensitive transactions. Unlike the MinFIA, algorithm MaxFIA selects the item with the maximum support in the restrictive pattern as a victim item. Algorithm IGA groups restricted patterns in groups of patterns sharing the same itemsets so that all sensitive patterns in the group will be hidden in one step.

In [8] a heuristic algorithm named DSRRC (Decrease Support of R.H.S. item of Rule Clusters) is given, which provides privacy for sensitive rules at certain level while ensuring data quality. Proposed 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. Because of less modification in database it helps maintaining data quality. Algorithm DSRRC cannot hide rules having multiple RHS items.

Blocking is the replacement of an existing value with a "?". This technique inserts unknown values in the data to fuzzify the rules. In some applications where publishing wrong data is not acceptable, then unknown values may be inserted to blur the rules. When unknown values are inserted, support and confidence values would fall into a range instead of a fixed value. Consider the database shown in Table V. For rule A→C, Support (A→C) = 80% and Confidence (A→C) = 100%. After fuzzifying the values, support and confidence becomes marginal. So in new database: $60\% \leq$ Confidence (A → C) $\leq 100\%$ and $60\% \leq$ Support (A→C) $\leq 80\%$.

In [10] two algorithms are built based on blocking for rule hiding. The first one focuses on hiding the rules by reducing the minimum support of the itemsets that generated these rules (i.e., generating itemsets). The second one focuses on reducing the minimum confidence of the rules. In [11] and [12] algorithms based on blocking technique are proposed and analyzed.

In blocking technique the maximum confidence of a sensitive rule cannot be reduced. If the blocking algorithm does not add much uncertainty in the database, adversary can infer the hidden values if he applies a smart inference technique. In database both 0's and 1's must be hidden during blocking, because if only 1's were hidden the adversary would simply replace all the ?'s with 1's and would restore easily the initial database and many ?'s must be inserted, if we don't want an adversary to infer hidden data.

B. Border based approach

This approach hides sensitive association rule by modifying the borders in the lattice of the frequent and the infrequent itemsets of the original database. The itemsets which are at the position of the borderline separating the frequent and infrequent itemsets forms the borders. The algorithms in this approach differ in the methodology they follow to enforce the new, revised borders, in the modified database. Border based approach uses the theory of borders presented in [13]. The first frequent itemset hiding methodology that is based on the notion of the border is proposed in [14, 15]. It maintains the quality of database by greedily selecting the modifications with minimal side effect. Then in [16, 17] more efficient algorithms based on border theory are presented.

C. Exact approach

This approach contains nonheuristic algorithms which formulates the hiding process as a constraints satisfaction problem or an optimization problem which is solved by integer programming. These algorithms can provide optimal hiding solution with ideally no side effects. In [18] an exact algorithm for association rule hiding is proposed which tries to minimize the distance between the original database and its sanitized version. In [19] proposed an exact border based approach to achieve optimal solution as compared to previous approaches.

IV. LIMITATIONS OF ASSOCIATION RULE HIDING APPROACHES

Heuristic algorithms may suffer from undesirable side-effects that lead them to identify approximate hiding solutions. This is due to fact that heuristics always aim at taking locally best decisions with respect to the hiding of the sensitive knowledge which, however, are not necessarily also globally best [1]. Heuristic algorithms may cause undesirable side effects to nonsensitive rules, e.g. lost rules and new rules.

In border based approach, theory of border revision is critical for the understanding. Although border-based approaches provide an improvement over pure heuristic approaches, they are still reliant on heuristics to decide upon the item modifications that they apply on the original database. As a result, in many cases these methodologies are unable to identify optimal hiding solutions, although such solutions may exist for the problem at hand [1].

Algorithms in exact approaches have very high time complexity due to the time that is taken by the integer programming solver to solve the optimization problem [1].

TABLE VI. COMPARATIVE ANALYSIS OF HEURISTIC ALGORITHMS

Rule Hiding Strategy	Algorithm	Item Hiding Algorithm		Rule Hiding Algorithm
		LHS	RHS	
Insertion	ISL	Y		
	DCIS		Y	
	1.a			Y
Deletion	DSR	Y		
	DCDS		Y	
	DSC	Y	Y	
	Naïve	Y	Y	
	MinFIA	Y	Y	
	MaxFIA	Y	Y	
	IGA	Y	Y	
	FHSAR			Y
	1.b			Y
	2.a			Y
	DSRRC			Y

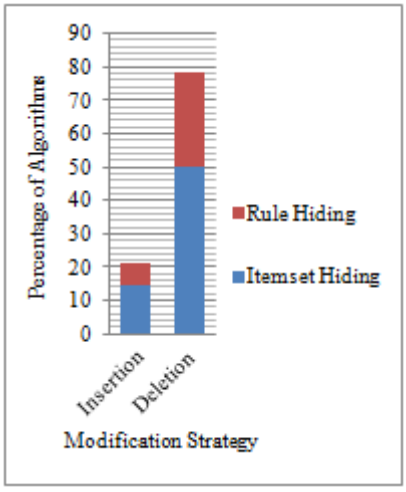

Figure 1. Heuristic algorithms proposed in each strategy

V. ANALYSIS OF ALGORITHMS BASED ON HEURISTIC APPROACH

From work done in heuristic approach, we have analyzed that there are mainly two strategies to hide any association rule. First strategy hides sensitive rule by inserting a new itemset in selected transactions. For rule X→Y, inserting itemset X in transactions not supporting Y will decrease confidence of rule X→Y. Second strategy is to remove itemset from selected transactions. It can be done in two ways. For rule X→Y, we can remove itemset Y to decrease the confidence of rule or we can reduce the support count of large itemset XY by removing items in X or Y from selected transactions. In this survey we have selected heuristic algorithms shown in Table VI, which are proposed for association rule hiding. Among them algorithms 1.a, ISL, and DCIS are using first strategy of inserting items to selected transactions. While other algorithms like 1.b, 2.a, DSR, DCDS, FHSAR, DSRRC, DSC, Naïve, MinFIA, MaxFIA and IGA are using second strategy of removing items.

Another way by which we can categorize these algorithms is by hiding approach. Item hiding algorithms hide all the association rules having particular itemset. Rule hiding algorithms hide all the sensitive rules given. Algorithm 1.a, 1.b, FHSAR and DSRRC are based on rule hiding approach while others are based on item hiding approach. Fig. 1 shows percentage of heuristic algorithms proposed in item insertion and deletion strategies.

It is observed that whenever item inserting strategy is used, it creates more artifactual patterns because it increases the support of some itemsets such that they becomes frequent. It sometimes fails to hide some sensitive association rules due to new patterns created as side effects. On the other hand if item removing strategy is used then some frequent item set becomes infrequent. So it affects the nonsensitive rules which are hidden as side effects.

VI. CONCLUSION

In this survey we have first described association rule hiding approaches and we have surveyed existing algorithm for association rule hiding. Based on this, we have done comparative analysis of heuristic algorithms. From that analysis it is concluded that whenever item insertion approach is used, there are more artifactual patterns created and it also have more hiding failure than deletion approach. So item insertion affects two parameters more, which are hiding failure and artifactual patterns. Sometimes it also shows missing cost. Whenever deletion approach is used at that time it affects misses cost parameter more. All other parameters are affected comparatively less than insertion approach.

REFERENCES

1. Aris Gkoulalas–Divanis;Vassilios S. Verykios "Association Rule Hiding For Data Mining" Springer, DOI 10.1007/978-1-4419-6569-1, Springer Science + Business Media, LLC 2010
2. M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. S. Verykios "Disclosure limitation of sensitive rules,".In Proc. of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99), pp. 45–52, 1999.
3. Vassilios S. Verykios, A.K. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni, "Association Rule Hiding," IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 4, pp. 434-447, 2004.
4. Shyue-Liang Wang; Bhavesh Parikh,; Ayat Jafari, "Hiding informative association rule sets", ELSEVIER, Expert Systems with Applications 33 (2007) 316–323,2006
5. Shyue-LiangWang ;Dipen Patel ;Ayat Jafari ;Tzung-Pei Hong, "Hiding collaborative recommendation association rules", Published online: 30 January 2007, Springer Science+Business Media, LLC 2007
6. Shyue-Liang Wang; Rajeev Maskey; Ayat Jafari; Tzung-Pei Hong " Efficient sanitization of informative association rules" ACM , Expert Systems with Applications: An International Journal, Volume 35, Issue 1-2, July, 2008
7. Chih-Chia Weng; Shan-Tai Chen; Hung-Che Lo, "A Novel Algorithm for Completely Hiding Sensitive Association Rules", IEEE Intelligent Systems Design and Applications, 2008.,vol 3, pp.202-208, 2008
8. Modi, C.N.; Rao, U.P.; Patel, D.R., "Maintaining privacy and data quality in privacy preserving association rule mining", IEEE 2008 Seventh International Conference on Machine Learning and Applications, pp 1-6, 2010
9. Stanley R. M. Oliveira; Osmar R. Za¨_ane, "Privacy Preserving Frequent Itemset Mining", IEEE International Conference on Data Mining Workshop on Privacy, Security, and Data Mining, Maebashi City, Japan. Conferences in Research and Practice in Information Technology, Vol. 14.2002
10. Y.Saygin, V. S. Verykios, and C. Clifton, "Using Unknowns to Prevent Discovery of Association Rules," ACM SIGMOD, vol.30(4), pp. 45–54, Dec. 2001.
11. Y. Saygin, V. S. Verykios, and A. K. Elmagarmid, "Privacy preserving association rule mining," In Proc. Int'l Workshop on Research Issues in Data Engineering (RIDE 2002), 2002,pp. 151–163.
12. E. Pontikakis, Y. Theodoridis, A. Tsitsonis, L. Chang, and V. S. Verykios. A quantitative and qualitative analysis of blocking in association rule hiding. In Proceedings of the 2004 ACM Workshop on Privacy in the Electronic Society (WPES), pages 29–30, 2004.
13. H. Mannila and H. Toivonen, "Levelwise search and borders of theories in knowledge discovery," Data Mining and Knowledge Discovery, vol.1(3), pp. 241–258, Sep. 1997.
14. X. Sun and P. S. Yu. A border–based approach for hiding sensitive frequent itemsets. In Proceedings of the 5th IEEE International Conference on Data Mining (ICDM), pages 426– 433, 2005.
15. X. Sun and P. S. Yu. Hiding sensitive frequent itemsets by a border–based approach. Computing science and engineering, 1(1):74–94, 2007.
16. G. V. Moustakides and V. S. Verykios. A max–min approach for hiding frequent itemsets. In Workshops Proceedings of the 6th IEEE International Conference on Data Mining (ICDM), pages 502–506, 2006.
17. G. V. Moustakides and V. S. Verykios. A maxmin approach for hiding frequent itemsets. Data and Knowledge Engineering, 65(1):75–89, 2008.
18. A. Gkoulalas-Divanis and V.S. Verykios, "An Integer Programming Approach for Frequent Itemset Hiding," In Proc. ACM Conf. Information and Knowledge Management (CIKM '06), Nov. 2006.
19. A. Gkoulalas-Divanis and V.S. Verykios, "Exact Knowledge Hiding through Database Extension," IEEE Transactions on Knowledge and Data Engineering, vol. 21(5), pp. 699–713, May 2009.

AUTHOR PROFILE

Komal Shah is a M.Tech student in U and P.U. Patel Department of Computer Engineering in Chandubhai S Patel Institute of Technology, Changa, India. She graduated from Saurashtra University, Rajkot, India with a B.E. degree in Computer Engineering in 2008. Her current research work focuses on Association Rule Hiding in Privacy Preservation Data Mining. She expects to obtain the M.Tech degree in 2012.

Amit Thakkar has received his B.E degree in Information Technology from Gujarat University, Gujarat, India in 2002 and master Degree from Dharmsinh Desai University, Gujarat, India in 2007. He has joined his Ph.D in the area of Multi relational Classification at KadiSarvavishvidhalayaUniversity, Gandhinagar, India in June 2010.

Since 2002 he has been with faculty of Engineering and Technology, Charotar University of Science and Technology, Changa, Gujarat, Where he is currently working as an Associate Professor in the Department of Information Technology. He has published more than 20 research papers in the field of data mining and web technology. His current research interest includes Multi relational Data Mining, Relational Classification and Associate Classification.

Amit Ganatra has received his B.E degree in Computer Engineering from Gujarat University, Gujarat, India in 2000 and master Degree from Dharmsinh Desai University, Gujarat, India in 2004. He has joined his Ph.D in the area of Multiple Classifier System (Information Fusion) at KadiSarvavishvidhalayaUniversity, Gandhinagar, India in August 2008.

Since 2000 he has been with faculty of Engineering and Technology, Charotar University of Science and Technology, Changa, Gujarat, Where he is currently working as an Associate Professor in the Department of Computer Engineering. He has published more than 50 research papers in the field of data mining and Artificial Intelligence. His current research interest includes Multiple Classifier System, Sequence Pattern Mining.