Association Rule Hiding: A Survey

Geeta S. Navale [#1], Dr. Suresh N. Mali*[2]

[#]*Sinhgad Institute of Technology and Science, Pune*

Research Scholar, SKNCOE, Pune, India

[1]geetanavale@gmail.com

[*] *Sinhgad Institute of Technology and Science (SITS)*

Principal, SITS, Pune, India

[2]snmali@rediffmail.com

Abstract— **In today's scenario of internet world, cloud computing and data centres, a lot of importance has to be given to the security of information while sharing it in different organisation. Today's advance data mining techniques gives knowledge about various relationships of itemsets which will be useful to group of people while enhancing their business. However, in many cases the organisation having the source of information may not be interested in sharing such relationships (association rules) with any other organisation or wants to get rights of information otherwise. Therefore, there is a need of hiding the sensitive information in the database prior to sharing the same among in the organization. The focus of this paper is to understand the methodologies of association rules hiding in database while keeping the disturbance in the original database as minimum as possible and categorize these methods in suitable classes. At the time robustness of hidden information has to be kept as high as possible for the intended recipients.**

Keywords— **Association rule hiding, Knowledge hiding, Privacy preserving data mining**

I. INTRODUCTION

An association rule is a rule of type X → Y, where X and Y are itemsets. If transaction contains itemset X, it (probably) also contains itemset Y.

Agrawal, R. et al. were the first to introduce Association rule mining[1]. Based on the concept of strong rules, Agrawal et al. introduced the problem of mining association rules from transaction data as follows: Let I = {i1,i2,...,in} be a set of n binary attributes called items. Let D = {t1, t2,..., tm} be a set of transactions called the database. In a database D, each transaction has an unique transaction ID and contains a subset of the items in I. A rule is defined of the form X ⇒ Y where X, Y ⊆ I and X ∩ Y = ∅. The sets of items (for short itemsets) X and Y is called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule. Association rules give the relationships between a set of items in large datasets to be discovered. When we discover such relationships between items, we get a large number of association rules. From these large number of association rules, it is very difficult to conclude which are important one? Then there should be some way or technique or measure which will decide this. By applying this technique we will get the desired or interested association rules. Various interesting measures have been proposed by Geng L., Hamilton et al. [2] and McGarry, K [3]. G. Dong et al. [4], Lallich, S. et al.[5], A.A. Freitas [6], Michael Steinbach [7] also put a focus on interesting measures. Pang-Ning Tan et al. [8] focuses on how to select correct objective measure to find the association rules. Searching for interesting patterns will optimize the problem. To exemplify the concepts, we use a small example from the superstore domain. The set of items is I = {wheat, pulses, rice, spices} and a small database containing the items is shown in Table 1.

TABLE I

Transaction ID	Items
T1	wheat
T2	pulses, rice, spices
T3	rice, spices
T4	rice, spices
T5	pulses, rice

An example of association rule for the superstore could be {pulses, rice} ⇒ {spices} means if pulses and rice is bought, customers also buy spices. To select interesting rules from all possible rules, constraints on various measures of importance and interest can be used. The best-known constraints are thresholds on support and confidence. The support supp(X) of an itemset X can be defined as the proportion of transactions in the data set which contain the itemset. The support of rule X → Y in data D is given in (1)

$$\text{supp}(X \rightarrow Y) = \text{supp}(X \cup Y) \quad (1)$$

The support is the probability that a transaction contains X ∪ Y. In the example database in Figure 1, the rule {pulses, rice}

⇒ {spices} has a support of 1/5 = 0.2 since it occurs in 20% of all transactions (1 out of 5 transactions). The confidence of rule X → Y in data D is given in (2)

$$conf(X \rightarrow Y) = supp(X \cup Y) / supp(X) \quad (2)$$

The confidence is the conditional probability that transaction contains *Y* given that it contains *X*. For example, the rule {pulses, rice} ⇒ {spices} has a confidence of 1/2 = 0.5 in the database in Table 1, which means that for 50% of the transactions containing pulses and rice the rule is correct.

The association rule mining has many applications. To mention few- shelf management in superstore, sending SMSs to customers when superstores are executing any scheme.

If owner of one of the superstores get the database of the other store, he can mine the database and gets the various business secrets of that owner and may result in loss of business. Thus there is a need to hide association rules. The hiding process should ensure that it will produce minimum side effects. The meaning of side effects is that after applying the hiding process to original database, it shouldn't drastically increase/decrease the size of database, it should give correct mining results, must hide all the required rules.

II. SURVEY

In this section we are going to present the techniques used by the researchers to hide the association rules. Dasseni et al.[9] were the first to propose a heuristic solution to hide association rule. They proposed three algorithms based on the measures of support and confidence. By using, the confidence measure, an algorithm could impose a rule to decrease confidence and become insignificant. They searched for the transactions that partially support both the left- and the right-hand side of a sensitive rule. The confidence measure of these transactions is minimized below the confidence threshold.

Oliveira et al.[10] introduced a framework for hiding sensitive association rules. This framework scans the database either once or twice to hide the rules. They select the victim item by computing the frequencies of items in the restrictive association rules. They compute the number of sensitive transactions to hide. Then they are sorting the sensitive transactions by size, and finally hide a sensitive transaction.

Verykios et al.[11] progressed on existing techniques by establishing a distortion and a blocking algorithms. The distortion algorithm works like this- a weight is assigned to each rule in the database based on the minimum confidence threshold is its confidence. For each transaction, based on these weights a priority value is calculated. All the transactions are sorted in ascending order of priority.

Wu et al.[12] proposed an approach that classifies the valid modifications and represents each class by three attributes indicating the modification scheme, the set of items to be contained in the transactions to be modified and the item to be modified.

Sun and Yu [13] [14] proposed a border based hiding process. To have separation between sensitive and nonsensitive items, they suggested a border based approach. Their main aim was data quality. They went for minimization of side effects at the cost of degrading efficiency. An algorithm was proposed to recognize the optimal candidate item to be removed, which has a minimal impact on the border. They used the concept of the weight of a border itemset to count the impact of hiding candidates on the border.

Moustakides and Verykios [15][16] proposed approach efficient than the approach proposed by Sun and Yu. Also hiding quality was better. The proposed algorithms sorts the sensitive itemsets based on their support in an increasing order. The hiding process starts with the sensitive itemset having the minimum support. The procedure is repeated until the hiding of every sensitive itemset is done.

Menon et al. [17] were the first to introduce a Constraint Satisfaction Problem (CSP) to hide association rules. The CSP is formulated as an integer program. The authors proposed two approaches blanket approach and intelligent approach. In blanket approach one item is remaining from the original transaction after hiding. In intelligent approach some of items from the transaction were removed which will result in reducing the support of every sensitive itemset. The integer program was used to maximize the accuracy of the shared database by minimizing the number of transactions that had to be changed to hide the frequent itemsets corresponding to the sensitive patterns.

Menon and Sarkar [18] improved the first solution methodology. They minimized the number of transactions that need to be sanitized. The authors also tried to keep the minimum number of nonsensitive itemsets lost during the hiding of the sensitive ones. Because of its high complexity, they reported that the problem could not be solved even for very small number of cases. Because of this reason, they proposed a solution that depends on two smaller solutions. First one is sanitization and the modified frequent itemset-hiding. The sanitization solution identifies how the support of sensitive itemsets can be purged from a specific transaction by removing very few numbers of items from it. Because of this fewer nonsensitive itemsets will be lost during the hiding process of sensitive itemsets. The modified frequent itemset-hiding solution will concentrate on minimizing the number of nonsensitive itemsets lost, while hiding sensitive ones.

Gkoulalas-Divanis and Verykios.[19] proposed an exact approach for association rule hiding problem. For effective hiding, the authors revised the border introduced by Sun and Yu [13][14] to find out a minimum set of itemsets without any side effects. They used the revised border to formulate a

minimal integer program, which is changed into a binary integer program.

The approach proposed by Gkoulalas-Divanis and Verykios [20] proposed a two-phase iterative algorithm that developed the functionality of the 'inline' algorithm. The two-phase iterative algorithm comprises of two phases that iterate until either (1) an exact solution of the given problem instance is identified or (2) a prespecified number of subsequent iterations of the algorithm has been done. The first phase of the algorithm utilizes the 'inline' algorithm to hide the sensitive knowledge. If it succeeds, the process is terminated and the modified database is returned. If the first phase is unable to identify an exact solution, the algorithm proceeds to the second phase, in which a certain number of selected inequalities from the CSP of the first phase are removed in an 'one-to-one' fashion, until the underlying CSP becomes feasible. The second phase results in the expansion of the positive border, whereas the iteration through the two phases aims to approximate the optimal borderline between the sensitive and public itemsets in the sanitized database.

Gkoulalas-Divanis and Verykios[21] introduced the first exact methodology to hide sensitive frequent itemsets by extending the original database. By utilizing the revised borders as well as the cover relationships among the item sets, authors were able to minimize the set of item sets participating in the CSP, which provides a solution to the sensitive knowledge hiding. The hiding process maximizes the data utility of the sanitized database, with minimum side effects while guaranteeing the satisfaction of all the privacy constraints. The released database (the original and the extended database), can guarantee the protection of the sensitive knowledge, when mined at the same or at a higher support threshold than that used for the mining of the original database. Extending the original database for sensitive itemset hiding has confirmed to offer better solutions for an extended set of hiding problems, compared with the previously presented approaches, as well as to lead to hiding solutions of better quality.

All the discussed methods are shown in Fig. 1.

III. GOALS OF ASSOCIATIONRULE HIDING METHODS

Let us assume that D is the original database which we want to mine. The main aim of the hiding process is to modify database D into sanitized database D^S in such a way that sensitive association rules should not be mined. Following are the goals of the association rule hiding -

A. $D \rightarrow D^S$

Nonsensitive rules can be derived from D & D^S.

B. $D \square D^S$

Difference between D & D^S should be minimum.

C. No false rules should be generated from D^S when we mine it.

All these goals must be satisfied by the hiding process to achieve optimality.

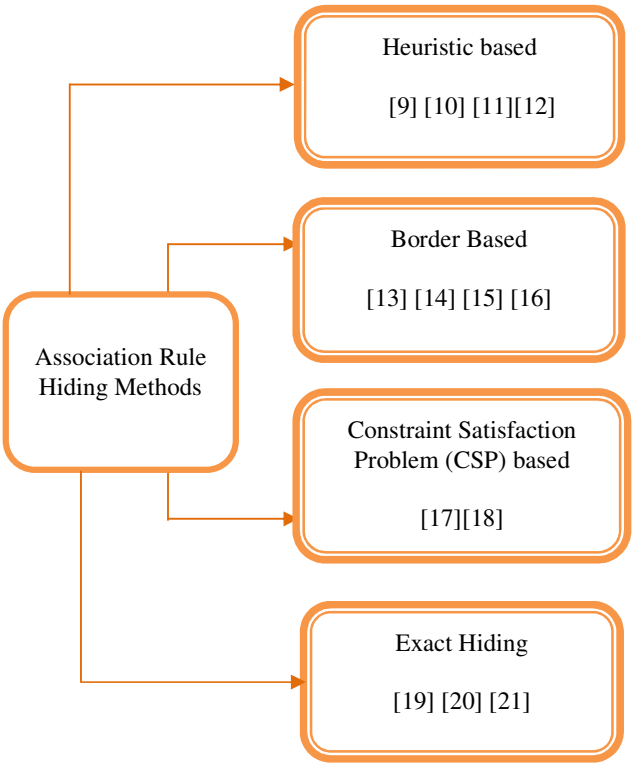

Fig. 1. Different Approaches for Association Rule Hiding

IV. CONCLUSIONS

There are various methodologies of hiding the association rules in the database and optimizing support and confidence. Each methodology has its advantages and disadvantages. One can classify these methodologies in one of the four broad categories presented in this paper. Certain approaches may use combination of the approaches or extend the idea with foundation of basic approaches mentioned in this paper.

ACKNOWLEDGMENT

I am thankful to Sinhgad Institute of Technology and Science, Pune as well as to Smt. Kashibai Navale College of Engineering, Pune for technical support.

REFERENCES

[1] Agrawal, R., Imielinski, T., Swami, "*A.: Mining Association Rules between Sets of Items in Large Databases,*" ACM SIGMOD

International Conference on Management of Data (SIGMOD'93), Washington D.C., USA , pp. 207–216, May 1993.

[2] Geng L., Hamilton, H.J., "*Interestingness Measures For Data Mining: A Survey,"* ACM Comput. Surv. (CSUR) 38(3), 9, 2006.

[3] McGarry, K, "*A survey of Interestingness Measures for Knowledge Discovery,"* Knowl. Eng. Rev. 20(1), pp. 39–61, 2005.

[4] G. Dong and J. Li, "*Interestingness of Discovered Association Rules in terms of Neighbourhood-Based Unexpectedness*," Second Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'98), pp. 72-86, 1998.

[5] Lallich, S., Teytaud, O., Prudhomme, E., "*Association Rule Interestingness: Measure and Statistical Validation*," Qual. Measur. Data Min. 43, pp. 251–276, 2006.

[6] A.A. Freitas, "*On Rule Interestingness Measures*," Elsevier, Knowledge-Based Systems 12, pp. 309–315,1999.

[7] Michael Steinbach, Pang-Ning Tan, Hui Xiong, and Vipin Kumar, "*Objective Measures for Association Pattern Analysis*," American Mathematical Society, pp. 1-21,2007.

[8] Pang-Ning Tan, Vipin Kumar, Jaideep Srivastava "*Selecting the Right Objective Measure for Association Analysis*," Elsevier Ltd, Information Systems 29, pp. 293–313, 2004.

[9] Dasseni E, Verykios VS, Elmagarmid AK, Bertino E., "*Hiding Association Rules by Using Confidence and Support*," In: Proceedings of the 4th International Workshop on Information Hiding, pp. 369–383, 2001.

[10] Oliveira SRM, Zaiane OR., "*A Unified Framework for Protecting Sensitive Association Rules in Business Collaboration,"* Int J Bus Intell Data Mining, pp. 1:247–287, 2006.

[11] Verykios VS, Pontikakis ED, Theodoridis Y, Chang L., "*Efficient Algorithms for Distortion and Blocking Techniques in Association Rule Hiding*," Distributed Parallel Databases, pp. 22:85–104, 2007.

[12] Wu Y-H, Chiang C-M, Chen ALP., "*Hiding Sensitive Association Rules with Limited Side Effects*," IEEE Trans Knowledge Data Eng, pp. 19:29–42, 2007.

[13] Sun X, Yu PS., "*A Border-Based Approach for Hiding Sensitive Frequent Itemsets*," In: Proceedings of the Fifth IEEE International Conference on Data Mining, pp. 426–433, 2005.

[14] Sun X, Yu PS., "*Hiding Sensitive Frequent Itemsets by A Border-Based Approach*," JCSE, pp. 1:74–94, 2007.

[15] Moustakides GV, Verykios VS., "*A Max-Min Approach for Hiding Frequent Itemsets*," In: Proceedings of the Sixth IEEE International Conference on Data Mining—Workshops, pp. 502–506, 2006.

[16] Moustakides GV, Verykios VS., "*A Maxmin Approach for Hiding Frequent Itemsets,"* Data Knowledge Eng, pp. 65:75–89, 2008.

[17] Menon S, Sarkar S, Mukherjee S., "*Maximizing Accuracy of Shared Databases When Concealing Sensitive Patterns,"* Inf Syst Res, pp. 16:256–270, 2005.

[18] Menon S, Sarkar S., "*Minimizing Information Loss and Preserving Privacy*," *Manage Sci*, pp. 53:101–116, 2007.

[19] Gkoulalas-Divanis A, Verykios VS., "*An Integer Programming Approach For Frequent Itemset Hiding*," In:International Conference on Information and Knowledge Management, Proceedings, pp. 748–757, 2006,.

[20] Gkoulalas-Divanis A, Verykios VS., "*Hiding Sensitive Knowledge Without Side Effects*," Knowledge Inf Syst , pp. 20:263–299, 2009.

[21] Gkoulalas-Divanis A, Verykios VS., "*Exact Knowledge Hiding Through Database Extension*," IEEE Trans Knowledge Data Eng, pp. 21:699–713, 2009.