Association Rule Hiding Methodologies: A Survey

Suma B.
Assistant Professor
Department of Computer Science & Engineering
R V College of Engineering, Bangalore, India

Abstract

Data mining provides the opportunity to extract useful information from enormous amount of data. Recent advances in data mining and machine learning algorithms have increased the disclosure risks that one may encounter when releasing data to third party.The objective of the association rule hiding algorithms is to hide sensitive information so that they cannot be discovered through association rule mining algorithm, but at the same time not losing the great benefit of association rule mining. Based on the execution time, the degree of optimality, the level of tolerance of side effects and guaranteed to get solution, different association rule hiding approaches are exist. This paper provides a brief survey of different of association rule hiding approaches and then discusses merits and short comings of these approaches.

1. Introduction.

Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. Data mining is being put into use and studied for databases, including relational databases, object-relational databases and object oriented databases, data warehouses, transactional databases, unstructured and semi- structured repositories such as the World Wide Web, advanced databases such as spatial databases, multimedia databases, time-series databases and textual databases, and even flat files. Two critical factors for success with data mining are: a large, well-integrated data warehouse and a well-defined understanding of the business process within which data mining is to be applied. Various data mining techniques such as, decision trees, association rules, and neural networks are already proposed and become the point of attention for several years.

Providing security to sensitive data against unauthorized access has been a long term goal for the database security research community and for the government statistical agencies. Data mining also poses a threat to privacy and information protection if not done or used properly. Some of the potential security risks associated with knowledge discovery have been first investigated in [1]. Clifton & Marks [2] were the first to propose possible remedies to the protection of sensitive data and sensitive knowledge from the use of data mining.

Privacy-preserving data mining algorithms, the first of which were introduced by Agarwal and Srikant [3] and Lindell and Pinkas [4], allow parties to cooperate in the extraction of knowledge, without any party having to reveal individual data items. The association rule hiding problem is to sanitize database in a way that through association rule mining one will not be able to disclosing the sensitive rules and will be able to mine all the non-sensitive rules. Finding an optimal solution to this problem is NP-hard, proved in [5].

The rest of this paper is organized as follows. In section 2, basics of association rule mining is discussed. Section 3, presents a brief introduction of privacy preserving association rule hiding problem. In section 4 various association rule hiding techniques are discussed and analysis of the hiding techniques are given is given in section 5.Finally Section 6 concludes the paper.

2. Association rule mining.

Association rule mining technique is the most effective data mining technique to discover hidden pattern among the large amount of data. It is responsible to find correlation relationships among different data attributes in a large set of items in a database. Association Rules Mining introduced by R. Agarwal [6] is an important research topic among the various data mining problems.

In this section the basic concepts of association rule mining is illustrated. The problem of mining association rules can be explained as follows:There is the item set I= $\{i_1, i_2 \ldots ..,i_n\}$ where I is a set of 'n' discrete items, and consider $D=\{t_1,t_2,t_3,\ldots,t_m\}$ as a set of transactions, each transaction $t_i \in D$ is an itemset such that $t_i \subseteq I$. A unique identifier, TID, is associated with each transaction. A transaction t supports X, a set of items in I, if $X \subseteq t_i$. It is assumed that the items in a transaction are sorted in lexicographic order. A sample database of transactions is shown in Table 1.

Table 1: Sample transaction data

TID	Transaction Items
T1	A,B,C
T2	A,B,C
T3	A,B,C
T4	A,B
T5	A
T6	A,C

An association rule is an inference of the form $X \Rightarrow Y$, where $X, Y \subset I$ and $X \cap Y = \varphi$. The set of items X is called antecedent and Y the consequent. Two properties support and confidence are generally considered in association rule mining. Support S for a rule $X \Rightarrow Y$, denoted by $S(X \Rightarrow Y)$,

is the ratio of the number of transactions in D that contain all the items in X U Y to the total number of transactions in D defined as :

$$S(X \Rightarrow Y) = \sigma(X \cup Y) / |D|$$

where the function σ of a set of items X i.e. σ (X), indicates the number of transactions in D, which contains all the items in X. |D| is the total number of transactions in the database D. Confidence C for a rule $X \Rightarrow Y$, denoted by $C(X \Rightarrow Y)$, is the ratio of the support count of X U Y to that of the antecedent X defined as :

$$C(X \Rightarrow Y) = \sigma(X \cup Y) / \sigma(X)$$

The minimum support S_{min} and minimum confidence C_{min} is defined by the user .The task of association rule mining is to mine from a data set D, that have support and confident greater than or equal to the user specified support value. Note that, while the support is a measure of the frequency of a rule, the confidence is a measure of the strength of the relation between sets of items.

Association rule mining is a two-step process:

1. Find all frequent item sets: All the item set that occur at least as frequently as the user specified minimum support count.
2. Generate strong association rules: These rules must satisfy user defined minimum support and minimum confidence.

For example, let assume the Minimum Support for the items at "table 1" is 40% and the minimum confidence is 60%.We need to find the association rule $\{A, B\} \Rightarrow \{C\}$ is valid or not. This rule has support of 50% and confidence 75%. Therefore this rule is valid association rule because it satisfies the minimum support and minimum confidence.

Agrawal and Srikant proposed the Apriori association rule mining algorithm [7]**.** Apriori algorithm discovers meaningful itemsets and constructs association rules within large databases, but the generation of candidate itemsets needs to perform contrasts against the whole database, level by level, in the process of creating association rules. Performance is considerably affected, as the database is repeatedly scanned to contrast each candidate itemset with the database.

Han et. al. [8] proposed a novel frequent-pattern tree (FP-tree) structure, which is an extended prefix-tree structure for storing compressed, crucial information about frequent patterns and developed an efficient, FP-growth method, for mining the complete set of frequent patterns by pattern fragment growth. FP-growth method is efficient and scalable for mining both long and short frequent patterns and is about an order of magnitude faster than the Apriori algorithm. Tsay and Chiang [9] proposed an efficient cluster based association rule mining method (CBAR) for discovering the large itemsets. A graph-based approach to generate various types of association rules from a large database of customer transactions had been proposed in [10]. This approach scans the database once to construct an association graph and then traverses the graph to generate all large itemsets.

3. Privacy preserving association rules.

Privacy preserving association rule mining should achieve the following goals: (1) All the sensitive association rules must be hidden in sanitized database. (2) All the rules that are not specified as sensitive can be mined from sanitized database. (3) No new rule that was not previously found in original database can be mined from sanitized database. First goal considers privacy issue. Second goal is related to the usefulness of sanitized dataset. Third goal is related to the side effect of the sanitization process.

The objective of the association rule hiding problem is to minimally sanitize database in such a way that association rule mining algorithm will not be able to discover sensitive rules and will be able to mine all the non-sensitive rules. The association rule hiding problem can be stated as follows: Given a transactional database D, a set R of relevant rules that are mined from D and a subset R_H of R ,where R_H is the set of sensitive rules, how can we transform D into a database D' in such a way that the every rule in R can still be mined, except for the rules in R_H.

Thus, the association rule hiding algorithm should transform D to D' that maximizes the number of rules in R - R_H, that can still be mined. There are two main association rule hiding can be adopted to hide a set R_H of rules (i) either prevent the rules in R_H from being generated, by hiding the frequent sets from which they are derived, or (ii) reduce the confidence of the sensitive rules, by bringing it below a user-specified threshold. In [11] the authors demonstrate that solving this problem by reducing the support of the large itemsets by removing items from transactions is an NP-hard problem.

4. Association rule hiding approaches.

Many approaches have been proposed to preserve privacy of sensitive association rules in database. They can be classified in to following categories: heuristic based approaches, cryptography based approaches, border based approaches, exact approaches and reconstruction based approaches.

4.1 *Heuristic based approaches*

These approaches can be further divided in to two groups based on data modification techniques: data distortion techniques and data blocking techniques.

4.1.1 ***Data distortion technique.*** Data-Distortion technique is based on data perturbation or data transformation. This technique changes a selected set of 1-values to 0-values (delete items) or 0-values to 1- values (add items), if we consider the transaction database as a two-dimensional matrix. Its aim is to reduce the support or confidence of the sensitive rules below the user predefined threshold.Verykios et al. [12] proposed five assumptions which are used to hide sensitive knowledge in database by reducing support or confidence of sensitive rules.

The authors in [13] presented four algorithms Naïve, MinFIA, MaxFIA and IGA. Each algorithm selects the

sensitive transactions to sanitize based on degree of conflict. Oliveira, S.R.M. and Zaïane, O.R in [14] introduced an efficient Sliding Window Algorithm (SWA) that improves the balance between protection of sensitive knowledge and pattern discovery. However this algorithm doesn't take the effect of non sensitive rules into consideration. In [15] authors proposed a method to reduce the side effects in sanitized database.

4.2.2 ***Data blocking technique.*** This technique adds uncertainty in the database by replacing 0's and 1's by unknowns ("?") in selected transaction instead of inserting or deleting items in a way that the database can still be used by a data miner that receives the database and at the same time an adversary cannot infer the sensitive rules that blocking technique will hide.

Y.Saygin et al. [16][17] were the first to introduce data blocking technique for hiding sensitive rules. The safety margin is also introduced in [16] to show how much below the minimum threshold, the new support and confidence of a sensitive rule should. Wang and Jafari [18] proposed a more efficient approach in [16][17]. The algorithm aims to achieve the following two goals: a) Reduce the minimum confidence of sensitive rules. b) Do not reduce the minimum confidence of non-sensitive rules.

If the adversary finds the maximum confidence of all the rules in the modified database, many new ghost rules will be found that did not exist in the initial database so the adversary cannot assume with certainty which of the rules that have maximum confidence above minimum confidence threshold were the sensitive rules. On the other hand, a data miner who wants to find useful information from the database can find the minimum confidence of all the rules, excluding in that way the sensitive rules from his/her information

4.2 *Cryptography based approaches*

Cryptography based approaches used in multiparty computation. If the database of one organization is distributed among several sites, then secure computation is needed between them. These approaches encrypt original database instead of distorting it for sharing. So they provide input privacy. Vaidya and Clifton [19] proposed a secure approach for sharing association rules when data are vertically partitioned. The authors in [20] addressed the secure mining of association rules over horizontal partitioned data.

4.3 *Border based approaches*

Border based approach uses the theory of borders presented in [21]. These approaches pre-process the sensitive rules so that minimum numbers of rules are given as input to hiding process.The sensitive association rules are hidden by modifying the borders in the lattice of the frequent and the infrequent item set of the original database. The item sets which are at the position of the borderline separating the frequent and infrequent item sets forms the borders. So, they maintain database quality while minimizing side effects.

Sun and Yu [22][23] were the first to introduce the frequent item set hiding methodology that is based on the notion of the border is proposed in. The proposed scheme, first computes positive border and negative border in the lattice of all item sets and focus on preserving the quality of the computed borders during the hiding process by greedily selecting the modifications with minimal side effect. Then in [24][25] more efficient algorithms based on border theory are presented.

4.4 *Exact approaches*

This approach formulates the hiding process as a constraints satisfaction problem (CSP) or an optimization problem which is solved by binary integer programming (BIP). These approaches provide better solution than other approaches. But they suffer from high time complexity to CSP. Gkoulalas and Verykios [26] proposed an approach to find optimal solution for rule hiding problem which tries to minimize the distance between the original database and its sanitized version.

The authors in [27] proposed a novel, exact border-based approach that provides an optimal solution for the hiding of sensitive frequent itemsets by minimally extending the original database by a synthetically generated database part - the database extension. Extending the original database for sensitive itemset hiding is proved to provide optimal solutions to an extended set of hiding problems compared to previous approaches and to provide solutions of higher quality.

4.5 *Reconstruction based approaches*

Data reconstruction approaches place the original data aside and start from sanitizing the so-called "knowledge base".The authors in [28] proposed a novel framework that can be regarded as "knowledge sanitization" approach, which is inspired by the inverse frequent set mining problem. This framework first performs sanitization on an itemset lattice called a knowledge base. The new released data is then reconstructed from the sanitized knowledge base. Mielikainen [29] was the first analyzed the computational complexity of inverse frequent set mining and showed in many cases the problems are computationally difficult.

Y. Guo [30] proposed a FP tree based inverse frequent set mining algorithm which reconstruct the original database by using non characteristic of database and efficiently generates number of secure databases. The FP-tree reduces the gap between a database and its frequent itemsets, transformation from given frequent itemsets to database can be carried out more smoothly, naturally and easily.

5. Analysis of association rule hiding approaches

Heuristic algorithms may suffer from undesirable side-effects on the non-sensitive rules in the data that lead them to identify approximate hiding solutions; these approaches have been getting focus of attention for majority of the researchers due to their efficiency, scalability and quick responses. Some

of the non-sensitive rules may be lost along with sensitive rules, and new ghost rules may be created because of the distortion or blocking process. 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.

Cryptographic approaches addresses secure mining of association rules over partitioned database but data do not protect the output of a computation. Instead, it prevents privacy leaks in the process of computation. Thus, it falls short of providing a complete answer to the problem of privacy preserving data mining.

The algorithms in exact approaches provide an exact (optimal) hiding solution that satisfies all the constraints with ideally no side effects. However if there is no exact solution exists in database, some of the constraint are relaxed. The time complexity these algorithms are very high due to the time that is taken by the integer programming solver to solve the optimization problem.

In border based approaches, theory of border revision is critical for the understanding. Although border-based approaches provide an improvement over pure heuristic approaches, they are still dependent on heuristics to decide upon the item modifications that they apply on the original database.

Reconstruction based approaches create privacy aware database by extracting sensitive characteristics from the original database. These approaches results in lesser side effects in database than heuristic approaches.

6. Conclusion

Association rule hiding is one of the techniques of privacy preserving data mining to protect the association rules generated by association rule mining algorithms. In this paper, a classification of privacy preserving association rule mining approaches is presented and major algorithms in each class are discussed. The merits and short comings of different techniques are also presented. All the proposed methods provides only approximate solution for the goal of privacy preservation, we need to further perfect those approaches or develop some efficient methods.

References

[1] O'Leary, D.E., "Knowledge Discovery as a Threat to Database Security," in *Knowledge Discovery in Databases*, AAAI Press/MIT Press, 1991, pp. 507-516.

[2] Clifton, C. and Don Marks, "Security and privacy implications of data mining" In: Proc. of the *ACM SIGMOD Workshop Data Mining and Knowledge Discovery*. 1996, pp. 15-19.

[3] R. Agarwal and R. Srikant,"Privacy-preserving data mining", *In ACM SIGMOD*, May 2000, pp. 439-450.

[4] Y. Lindell and B. Pinkas, "Privacy preserving data mining", Journal of Cryptology, 2002,15(3): pp.177-206.

[5] 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)*, 1999, pp. 45-52.

[6] R. Agrawal, T. Imielinski, and A. Swami.. "Mining association rules between sets of items in large databases". In Proceedings of the 1993 *ACM SIGMOD International Conference on Management of Data*, Washington, DC, May 26-28 1993, pp. 207-216.

[7] R. Agarwal, R. Srikant, "Fast algorithm for mining association rules in large databases", Proceedings of *1994 International Conference on VLDB*, 1994, pp. 487-499.

[8] Jiawei Han , Jian Pei , Yiwen Yin , Runying Mao," Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach", *Data Mining and Knowledge Discovery*, 8, 2004, pp. 53-87.

[9] Yuh-Jiuan Tsay, Jiunn-Yann Chiang, "CBAR: an efficient method for mining association rules", *Knowledge-Based Systems 18*, 2005, pp. 99-105.

[10] Show-Jane Yen and Arbee L.P. Chen, "A Graph-Based Approach for Discovering Various Types of Association Rules. *IEEE Transaction on Knowledge and Data Engineering*, Vol. 13, NO. 5, Septeember/October 2001, pp. 839-845.

[11] Atallah**,** M., Bertino, E., Elmagarmid, A., Ibrahim, M., and Verykios, V.S,"Disclosure limitation of sensitive rules", In: Scheuermann P, ed. Proc. of the *IEEE Knowledge and Data Exchange Workshop (KDEX'99)*. IEEE Computer society, 1999. pp. 45-52.

[12] Verykios, V.S., Elmagarmid, A., Bertino, E., Saygin, Y., and Dasseni, E. "Association rule hiding", *IEEE Transactions on Knowledge and Data Engineering*, 2004, 16(4): pp. 434-447.

[13] Oliveira, S.R.M. and Zaïane, O.R. Privacy preserving frequent itemset mining. In: Proc. of the 2 nd *IEEE ICDM Workshop on Privacy, Security and Data Mining*. Australian Computer Society, 2002, pp. 43-54.

[14] Oliveira, S.R.M. and Zaïane, O.R,"Protecting sensitive knowledge by data sanitization", In: Proc. of the *3rd IEEE Int'l Conf. on Data Mining (ICDM'03)*, IEEE Computer Society, USA, 2003, pp. 613-616.

[15] Y. H. Wu, C.M. Chiang and A.L.P. Chen, "Hiding Sensitive Association Rules with Limited Side Effects", *IEEE Transactions on Knowledge and Data Engineering*, vol.19(1), Jan. 2007, pp. 29-42.

[16] Y.Saygin, V. S. Verykios, and C. Clifton, "Using Unknowns to Prevent Discovery of Association Rules", *ACM SIGMOD*, vol.30(4), Dec. 2001, pp. 45-54,.

[17] 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.

[18] S.L.Wang and A. Jafari, "Using unknowns for hiding sensitive predictive association rules", In Proc. *IEEE Int'l Conf. Information Reuse and Integration (IRI 2005)*, Aug. 2005, pp. 223-228.

[19] J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data" ,In proc. Int'l Conf. *Knowledge Discovery and Data Mining*, July 2002, pp. 639-644,.

[20] M. Kantarcioglu and C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data", *IEEE Transactions on Knowledge and Data Engineering*, vol. 16(9), Sept. 2004, pp. 1026-1037.

[21] H. Mannila and H. Toivonen, "Levelwise search and borders of theories in knowledge discovery" *Data Mining and Knowledge Discovery*, vol.1 (3), Sep. 1997, pp. 241-258.

[22] 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)*, 2005, pp. 426-433.

[23] X. Sun and P. S. Yu., "Hiding sensitive frequent itemsets by a border–based approach", *Computing science and engineering*, 1(1):74-94, 2007.
[24] 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)*, 2006, pp. 502-506.
[25] G. V. Moustakides and V. S. Verykios.," A maxmin approach for hiding frequent itemsets", *Data and Knowledge Engineering*, 2008, 65(1):pp. 75-89.
[26] 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.
[27] Gkoulalas-Divanis and V.S. Verykios, "Exact Knowledge Hiding through Database Extension," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21(5), May 2009, pp. 699-713.
[28] Chen, X., Orlowska, M., and Li, X., "A new framework for privacy preserving data sharing", In: Proc. of the 4th *IEEE ICDM Workshop: Privacy and Security Aspects of Data Mining*, IEEE Computer Society, 2004, pp.47-56.
[29] T. Mielikainen, "On inverse frequent set mining",In Proc. of *3rd IEEE ICDM Workshop on Privacy Preserving Data Mining*. IEEE Computer Society, 2003, pp.18-23.
[30] Y. Guo, "Reconstruction-Based Association Rule Hiding" In Proc. of *SIGMOD2007 Ph.D. Workshop on Innovative Database Research 2007(IDAR2007)*, June 2007, pp.51-56.