Volume 5, Issue 12, December 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Survey on Association Rule Hiding Approaches

Mohammad Azam Chhipa
Student, Department of Computer Science,
IET DAVV, Indore, M.P., India

Prof. Lalit Gehlod
Professor, Department of Computer Science,
IET DAVV, Indore, M.P., India

Abstract*— There is exceptional growth in the research of data mining and its application. Data Mining is a procedure to extract fruitful knowledge which is hidden in large amounts of data. In recent years, due to increase in Data Mining techniques and its application the disclosure risk of sensitive information of released data is also increased because of this disclosure a research topic in data mining known as privacy preserving data mining (PPDM) is gaining popularity. The goal of PPDM is to provide security to sensitive information by modifying raw data in such a way that the data mining algorithm cannot extract that information. In this survey paper we will discuss a topic of data mining known as association rule mining and techniques to hide sensitive rules which are mined through data mining.

Keywords*— Data mining, PPDM, Association Rule mining, Sensitive Rule Hiding.

I. INTRODUCTION

Data Mining [1] refers to extracting or mining knowledge from large amounts of data. Data Mining is the process of discovering interesting patterns and knowledge from large amounts of data. This data on which mining is done can be stored either in databases, data warehouses, or other information repositories. By applying data mining techniques to this data repositories fruitful and interesting knowledge, regularities or other hidden information or patterns can be extracted and viewed and browsed from different angles. The discovered knowledge can be use in different areas like decision making, process control, information management and query processing. Data Mining has been successfully applied to many domains, such as business intelligence, web search, scientific discovery, digital libraries, etc. Data Mining, with its promise to efficiently discover valuable, non-obvious information from large databases, is particularly vulnerable to misuse. So there might be a conflict between data mining and privacy.

Because of the numerous application of data mining, the risk of disclosure of sensitive information of an organization is increased when the data is released to other parties. For example, consider Indian superstores like Food Bazaar and Reliance Fresh. Suppose shopkeeper of Reliance Fresh mines the association rules related to Food Bazaar, where he found that most of the customers who buy bread also buy milk. Seeing this, shopkeeper of Reliance Fresh exploits this information and puts some discount on the cost of bread. This is how customers of Food Bazaar will now move to Reliance. This scenario leads to the research of sensitive knowledge (or rule) hiding in database. Therefore, before releasing the dataset to the other party, each supermarket is willing to hide sensitive association rules of its own sensitive products. So, the sensitive information (or knowledge) will be protected. The problem of association rule hiding in the area of privacy preserving data mining was first proposed in 1999 by Atallah *et al.* [2].

Rest of this paper is organized as follows: - In Section 2, discusses the association rule mining strategy. The concept of association rule hiding is given in section3. Section 4 presents the existing association rule hiding approaches by identifying open challenges. Section 5 summarizes the recent evolutions in sensitive association rule hiding. The metrics used for evaluating sensitive rule hiding approaches are given in section 6. Section 7 conclude our study by identifying future work with references at the end.

II. ASSOCIATION RULE

Association rule mining is one of the most important data mining tasks, which goal is to find interesting association and correlation relationship among large sets of data items [3]. A typical example of association rule mining is market basket analysis [1], which analyses customer buying habits by finding associations between different items that customer place in their shopping baskets. The association can help retailers develop better marketing strategies. It was first introduced by R. Agarwal [3] in 1993. It works as follows:

Given a set of items $I = \{i1, i2, i3, \dots , im\}$, and a set of transaction $T = \{t1, t2, t3,\dots, tn\}$, where each transaction consists of several items from I. An association rule is an implication of the form: $A \rightarrow B$, where $A \subset I$, $B \subset I$, $A \neq \Phi$, $B \neq \Phi$, and $A \cap B \neq \Phi$. In the rule $A \rightarrow B$, A is called antecedent (Left-hand-side) and B is called consequent (Right-hand-side).The rule $A \rightarrow B$ holds in the transaction set T with support s, where s denotes the percentage of transactions in T that contain $A \cup B$.

$$\text{Support } (A \rightarrow B) = |A \cap B| \,/\, |D|.$$

The rule $A \rightarrow B$ has confidence c in the transaction set T, where c is the percentage of transactions in T containing A that also contains B.

Confidence ($A \rightarrow B$) = $|A \cap B| / |A|$.

In other words, support describes how often the rule would appear in the database, while confidence measures the strength of the rule. A rule $A \rightarrow B$ is strong if support ($A \rightarrow B$) ≥ minimum support and confidence ($A \rightarrow B$) ≥ minimum confidence.

Generally, the process of association rule mining contains the following two steps:

- Step 1: Find all frequent item sets. A set of item is referred as an *itemset*. The occurrence frequency of an itemset is the number of transactions that contain the itemset. A frequent itemset is an itemset whose occurrence frequency is larger than a predetermined minimum support count.
- Step 2: Generate strong association rules from the frequent itemsets. Rules that satisfy both a minimum support threshold (minsup) and a minimum confidence threshold (minconf) are called strong association rule.

Different types of algorithm are available to mine association rule like Apriori algorithm, Partition algorithm, Pincher-search algorithm, Dynamic item set counting algorithm, FP-tree growth algorithm, etc [4]. The Apriori algorithm proposed by Agrawal and Srikant [5] has proved to be one of the most versatile successful algorithm ranking next only to more sophisticated algorithm like *éclat, nonordp and lcm.* It has been proved by the scholar that Apriori outperforms some of these algorithms in areas like space and database content.

III. ASSOCIATION RULE HIDING

Given the thresholds of *support* and *confidence*, the data miner can find a set of association rules from the transactional data set. Some of the rules are considered to be sensitive, either from the data provider's perspective or from the data miner perspective. To hiding these rules, the data miner can modify the original data set to generate a *sanitized* data set from which sensitive rules cannot be mined, while those non sensitive ones can still be discovered, at the same thresholds or higher.

The problem can be stated as follows:

*Given a database D, a set R of relevant
rules that are mined from D and a subset
Rh of R, how can we transform D into a
database D' in such a way that the rules in
R can still be mined, except for the rules in
Rh?*

The goal is to transform original data set *D* into a sanitized data set *D'* in such a way that no association rule in *Rh* , which sensitive rule set, will be mined and all non sensitive rules in *R* could still be mined from *D'*. The main purpose of the association rule hiding algorithms is to make the sensitive rules invisible which can be generated by association rule mining algorithms. M. Attallah et al. [2] have proved that finding an optimal solution of sanitization problem is NP-Hard.

There are two main approaches that can be adopted when we try to hide a set *Rh* of rules: we can either prevent the rules in *Rh* from being generated, by hiding the frequent sets from which they are derived, or we can reduce their confidence by bringing it below a user-specified threshold *(min_con/)*[2]. Decreasing the confidence of a rule $A \rightarrow B$ can be done by either increasing the support of *A* in transactions and not of *B* or by decreasing the support of *B* in transactions supporting both *AB*. Decreasing the support of a rule $A \rightarrow B$ can be done by decreasing the support of the corresponding large itemset *AB*.

Association rule hiding must satisfy following conditions:

- No sensitive rule should be generated from Sanitized database.
- Non sensitive rule must be generated from Sanitized database.
- No new rule, present in database should be generated from Sanitized database.

IV. ASSOCIATION RULE HIDING APPROACHES

Sensitive association rule hiding is a subfield of Privacy Preserving Data Mining (PPDM). Privacy preserving data mining has been recently introduced to cope with privacy issues related to the data subjects in the course of mining of the data. Various kinds of approaches have been proposed to perform association rule hiding [6][7][22]. These approaches can be categorized into the following five groups:

1. Heuristic approaches :

This approach is further divided into two techniques:

- Heuristic Distortion Approaches, which solves how to select the appropriate data sets for data modification. The heuristic proposed for the modification of the data was based on data perturbation, and in particular the procedure was to change a selected set of 1-values to 0-values, so that the support of sensitive rules is lowered. In this the binary 1 value for an item set is modified to binary 0 value and vice versa. This flexibility in data modification had the side-effect that apart from non-sensitive association rules that were becoming hidden, a non-frequent rule could become a frequent one.

- Heuristic Blocking Approaches, which reduce the degree of support and confidence of the sensitive association rules by replacing certain attributes of some data items with a specific symbol (e.g. '?'). The introduction of this special unknown value brings uncertainty to the data, making the support and confidence of an association rule become two uncertain intervals respectively. At the beginning, the lower bounds of the intervals equal to the upper bounds. As the number of "?" in the data increases, the lower and upper bounds begin to separate gradually and the uncertainty of the rules grows accordingly. When either of the lower bounds of a rule's support interval and confidence interval gets below the security threshold, the rule is deemed to be concealed.

2. Border based approach

The border based approach hide sensitive association rules by modifying the border in the lattice of frequent and infrequent item sets of the original database. The item set between frequent and infrequent items make the border. The border consist the item sets which separate the frequent item set from infrequent item set. It uses the border of non-sensitive frequent item and computes the positive and negative borders in the itemset. Then minimal affected modification is selected. If modification is done by greedy selection then it leads to minimum side effects. Sun and Yu [8] were first who introduce the concept of border. The quality of database can be well maintained by controlling modifications according to the impact on the result database. A border-based approach was proposed to efficiently select the modification with minimal side effects. Border based approach used to separated data along border have bad result.

3. Exact approach

This approach is a nonheuristic algorithm which takes the hiding process as a constraints satisfaction problem or an optimization problem which is solved by integer programming. This algorithm provides optimal hiding solution without any side effects. An exact algorithm for association rule hiding is proposed in [8] which tries to minimize the distance between the original database and its sanitized version. In [9] proposed an exact border based approach to achieve optimal solution as compared to previous approaches. This can be achieved by applying good sanitization method which minimally distorts the original database. This approach can be considered as descendant of border based methodology. It works in the following way as follows. First border revision method is applied to small portion of item sets from the original database whose status is determined (that is frequent versus infrequent) and recorded. Then, exact methodologies incorporate unknowns to the original database and generate inequalities that control the status of selected item sets of the border. These inequalities along with an optimization criterion require minimal modification of the original database to facilitate sensitive knowledge hiding, formulate an optimization problem whose solution is guaranteed to lead to optimal hiding.

4. Reconstruction based approach

These approaches are efficient than the Heuristic based approaches and generate less side effects than heuristic based approaches. In this approach first frequent item set is extracted from non frequent item set and privacy protected data is released. The new released data is then reconstructed from the sanitized knowledge base. This approach, first perform data perturbing and then reconstruct the database. Basically this approach reconstructs the database in a manner that all sensitive information has been hidden. This method cannot guarantee to find a consistent one within a polynomial time [11]. Y. Guo [10] proposed a FP tree based algorithm which reconstruct the original database by using non characteristic of database and efficiently generates number of secure databases.

5. Cryptographic based approach

These approaches are used in multiparty computation, where two or more parties want to conduct computation on their own inputs but without leakage of output to other. The data is present at different sites, data is not organized in central manner. These approaches are classified in two categories: 1) Vertically partitioned distributed data 2) Horizontally partitioned distributed data.

These approaches the original database is encrypted not distorted as in previous approaches. Vaidya and Clifton [12] proposed a secure approach for sharing association rules when data are vertically partitioned. In terms of communication cost this approach is very effective, but it is very expensive for large amount of datasets. The authors in [13] addressed the secure mining of association rules over horizontal partitioned data. This approach mines association rules securely with reasonable communication cost and computation cost.

V. LITERATURE REVIEW

Many authors has done research on hiding sensitive rule some of their studies are as follows:

In [14], Jain et al. proposed a distortion based approach for hiding sensitive rules, where to reduce confidence of sensitive rule, position of the sensitive item is altered, modification is done in such a way that the support of the rule is remains same. The size of the database also remains same.

Hybrid partial hiding algorithm is employed by Zhu et al [15] on dataset and reconstruct the support of itemset, then applied Apriori [1] algorithm to generate frequent itemsets based on which only non sensitive rules can be obtained.

Le et al. [16] propose a heuristic algorithm to hide sensitive rules, this algorithm works on the intersection lattice of frequent itemsets. The algorithm first identifies the victim item such that the modification of this item causes less impact on the frequent itemsets. After this, the transaction which are need to be modified are specified. Then the victim item is removed from the specified transactions and the data is sanitized.

Modi et al. [17] propose a heuristic algorithm DSRRC (decrease support of right hand side item of rule clusters) for hiding sensitive association rules. The algorithm based on certain criteria clusters the sensitive rules to hide as much as rules at one time. This algorithm has a drawback that it can hide those rules which have multiple items in antecedent (left hand side) and consequent (right hand side). To remove this drawback Radadiya et al. [18] propose an improved version DSRRC named ADSRRC (Advance DSRRC) in this the item ,witch have highest count, in right hand side of the sensitive rules are iteratively deleted in data sanitization process.

Pathak et al. [19] propose an approach based on impact factor. In this cluster of association rules is build based on impact factor. Impact factor of a transaction is number of the itemsets those are present in the itemset which represents the sensitive rule. It means if the impact factor is higher than sensitivity of the rule is higher.

Mielikainen [20] investigated the problem of IFM (inverse frequent set mining) which is described as follows [21]: given a collection of frequent itemset and their support, find a transactional data set such that the data set precisely agrees with the support of the given frequent itemset collection while the supports of other itemsets would be less then predetermined threshold. Guo et al. [10] proposed a reconstruction based approach by solving an IFM problem to hide sensitive rules. Their approach consist three steps first, all frequent itemset with their support and support counts are generated using frequent itemset mining algorithm. Second, itemsets which are related to sensitive rules are removed. Third generate new transactional data set via inverse frequent set mining using the rest itemsets.

VI. CONCLUSION

Association rule hiding is an important concept of privacy preserving data mining. In this paper we have studied association rule and algorithms to mine association rule in data set. We have also studied different approaches to hide sensitive rules. These approaches are efficient but they are based on modification and alteration of original data set. Their is a need of research and studies to find an optimal solution to preserve privacy while data mining.

REFERENCES

[1] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. San Mateo, CA, USA: Morgan Kaufmann, 2006.

[2] M. Atallah, E. Bertino, A. Elmagamind, M. Ibrahim, and V. S. Verykios "Disclosure limitation of sensitive rules," .In Proc. of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop(KDEX 1999), pp. 45-52.

[3] R. Agrawal, T. Imieli«ski, and A. Swami, ``Mining association rules between sets of items in large databases," in *Proc. ACM SIGMOD Rec.*,

[4] S. Vijayarani, A. Tamilarasi and R. SeethaLakshmi, "Privacy Preserving Data Mining Based on Association Rule-A Survey". In Proc. of the International Conference on Communication and Computational Intelligence-2010, pp. 99-103.

[5] Rakesh Agrawal and Ramakrishnan Srikant, "Fast algorithms for mining association rules in large databases", In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, Proc. of the 20th International Conference on Very Large Data Bases, VLDB, pages 487-499, Santiago,

[6] V. S. Verykios, ``Association rule hiding methods," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 3, no. 1, pp. 28_36, 2013.

[7] K. Sathiyapriya and G. S. Sadasivam, ``A survey on privacy preserving association rule mining," Int. J. Data Mining Knowl. Manage. Process, vol. 3, no. 2, p. 119, 2013.

[8] X. Sun, and P. Yu, "A Border-Based Approach for Hiding Sensitive Frequent Itemsets," In: Proc. Fifth IEEE Int'l. Conf. Data Mining (ICDM 2005), pp. 426–433, 2005.

[9] A. Gkoulalas-Divanis, V. Verykios, "An Integer Programming Approach for Frequent Itemset Hiding," In: Proc. ACM Conf. Information and Knowledge Management (CIKM 2006), pp. 748–757 2006.

[10] Y. Guo, "Reconstruction-Based Association Rule Hiding," In Proc. Of SIGMOD2007 Ph.D. Workshop on Innovative Database Research 2007(IDAR2007), June 2007.

[11] Khyati B. Jadav, Jignesh Vania, Dhiren R. Patel "A Survey on Association Rule Hiding Methods" International Journal of Computer Applications, November 2013.

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

[13] 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, No. 9, pp. 1026-1037, Sept. 2004.

[14] D. Jain, P. Khatri, R. Soni, and B. K. Chaurasia, ``Hiding sensitive association rules without altering the support of sensitive item(s)," in Proc. 2nd Int. Conf. Adv. Comput. Sci. Inf. Technol. Netw. Commun., 2012,pp.500_509.

[15] J.-M. Zhu, N. Zhang, and Z.-Y. Li, ``A new privacy preserving association rule mining algorithm based on hybrid partial hiding strategy," *Cybern Inf. Technol.*, vol. 13, pp. 41_50, Dec. 2013.

[16] H. Q. Le, S. Arch-Int, H. X. Nguyen, and N. Arch-Int, ``Association rule hiding in risk management for retail supply chain collaboration," *Comput Ind.*, vol. 64, no. 7, pp. 776_784, Sep. 2013.

[17] C. N. Modi, U. P. Rao, and D. R. Patel, ``Maintaining privacy and data quality in privacy preserving association rule mining," in *Proc. Int. Conf* Comput. Commun. Netw. Technol. (ICCCNT), Jul. 2010, pp. 1_6.

[18] N. R. Radadiya, N. B. Prajapati, and K. H. Shah, ``Privacy preserving in association rule mining," *Int. J. Adv. Innovative Res.*, vol. 2, no. 4, pp. 203_213, 2013.

[19] K. Pathak, N. S. Chaudhari, and A. Tiwari, ``Privacy preserving association rule mining by introducing concept of impact factor," in *Proc. 7th* IEEE Conf. Ind. Electron. Appl. (ICIEA), Jul. 2012, pp. 1458_1461.

[20] T. Mielikäinen, ``On inverse frequent set mining," in *Proc. 2nd Workshop* Privacy Preserving Data Mining, 2003, pp. 18_23.

[21] X. Chen and M. Orlowska, ``A further study on inverse frequent set mining," in Proc. 1st Int. Conf. Adv. Data Mining Appl., 2005, pp. 753_760.

[22] C. Jiang, "Information Security in Big Data: Privacy and Data Mining" IEEE, October 2014.