National Conference on Emerging Trends in Computer, Electrical & Electronics (ETCEE-2015)
International Journal of Advance Engineering and Research Development (IJAERD)
e-ISSN: 2348 - 4470 , print-ISSN:2348-6406,Impact Factor:3.134

A Survey on Association Rule Hiding Approaches

Bindiya Sagpariya[1] Kruti Khalpada[2]
[1]Computer Engineering, AITS Rajkot, Gujarat India
[2] Computer Engineering, AITS Rajkot, Gujarat India
Address
[1]bindiyasagpariya34@gmail.com
[2]knkhalpada@aits.edu.in

***Abstract*— In recent years, explosive growth of the amount of data gathered by transactional systems, a challenge for finding new techniques to extract useful patterns from such a huge amount of data arose. As the database is growing day by day the organizations which maintain this database are worried about the importance of such huge transaction database. One of the greatest challenging tasks of data mining is finding hidden patterns without revealing sensitive information. Privacy preserving data mining (PPDM) is the recent research area that deals with the problem of hiding the sensitive information while analyzing data. PPDM algorithms are evolved for modifying the original data in such that the no sensitive information is revealed even after mining procedure. Association rule hiding is one of the privacy preservation techniques to hide sensitive association rules. All association rule hiding algorithms focus to minimally modify the original database such that no sensitive association rule is derived from it. This paper contains the comprehensive survey of privacy preserving data mining methods. Advantages and disadvantages of the existing algorithms are discussed in brief.**

***Keywords*- Data Mining, Privacy Preserving, Association Rule Hiding, PPDM, Sensitive Association Rule**

I. Introduction:

Data Mining is the process of extracting useful knowledge from large amounts of data. It is a knowledge discovery process which is useful to find patterns [9]. Following the explosive growth of the amount of data gathered by transactional systems, a challenge for finding new techniques to extract useful patterns from such a huge amount of data arose. Data mining known as knowledge discovery of data base is an efficient way of extracting required knowledge from given or available large amount of dataset. However, the technologies can be threats to data privacy.

Data mining has numerous applications in business, marketing, medical analysis, bioinformatics, etc. This has increased the revelation risks when the data is released to outside parties. Association mining finds its application across many domains. One of the best known applications of association rule mining is in the business field where discovery of purchase behaviors or association between products is very useful for decision making and developing effective marketing strategy. For example, consider superstores like Food Mall and Star Bazar. Suppose shopkeeper of Star Bazar mines the association rules related to Food Mall, where he found that most of the customers who buy noodles also buy tomato souse. Seeing this, shopkeeper of Star Bazar uses this information and puts some discount on the cost of noodles. This is how customers of Food Mall will now move to Star Bazar. This scenario leads to the research of sensitive knowledge (or rule) hiding in database. So, before releasing the dataset to the other, 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. [4].

As the database is growing day by day the organizations which maintain this database are worried about the importance of such huge transaction database. Sometimes organizations are interested in collaborating their datasets which organizations of similar fields to analyse their databases for mutual benefits. For example some banks want to generate some criterion for credit card policy or scheme and don’t want to leak the details of their customer to other banks than they require a technique which can analyse their data while maintaining the data privacy.

Privacy preservation data mining (PPDM) considers challenge of maintaining privacy of data and knowledge in data mining. It allows obtaining relevant knowledge and averts sensitive data or information from disclosure. PPDM algorithms are evolved for modifying the original data in such that the no sensitive information is revealed even after mining procedure. Association rule hiding is one of the privacy preservation techniques to hide sensitive association rules. All association rule hiding algorithms focus to minimally modify the original database such that no sensitive association rule is derived from it.

Rest of this paper is organized as follows: - In Section 2, introduction to association rule mining. The concept of association rule hiding and the existing association rule hiding approaches by identifying open challenges is given in 3. Section 4 summarizes the recent evolutions in sensitive association rule hiding. Section 5 conclude our study by identifying future work with references at the end.

II. Association rule mining:

Association rule mining technique is the most potential data mining technique to discover hidden pattern among the large dataset. It is accountable to find correlation relationships among different data attributes in a large set of items in a database. Association Rules Mining introduced by R. Agarwal[3] is an important research topic among the various data mining problems. Let I = {i1, i2, …., in} be a set of items from a database. Let D be a set of transactions. Each transaction t Є D is an itemset such that t is a proper subset of I. A transaction t supports A, a set of items in I, if A is a proper subset of t. An association rule of the form A→B, where A and B are subsets of I and A∩B= Ø. The support denoted as σ of rule A→B can be computed by the following equation:

Support (A→B) = |AUB| / |D|, where |AUB| denotes the number of transactions in the database that contains the item set AB, and |D| denotes the number of the transactions in the database D which means that σ% of the transactions in D supports item set AB. The confidence denoted as τ of rule A→B is calculated by following equation:

Confidence (A→B) = |AUB|/|A|, where |A| is number of transactions in database D that contains item set A which means τ% of the transactions in D that supports A also supports B. A rule A→B is strong if support(A→B) ≥ min_support and confidence(A→B) ≥ min_confidence, where min_support and min_confidence are two given minimum thresholds [13].

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.

III. Association rule hiding approaches & related work:

Privacy preserving data mining (PPDM) deals with protecting the privacy of individual data or sensitive knowledge without sacrificing the utility of the data. 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. Recent advances in data mining technologies have increased the disclosure risks of sensitive data [19].

The main approach to hide sensitive association rules is to reduce the support or the confidence of the rules. This is done by modifying transactions or items in the database. However, the modifications will generate side effects, i.e., non-sensitive rule falsely hidden (i.e., lost rules) and spurious rules falsely generated (i.e., new rules). There is a trade-off between sensitive rules hidden and side effects generated.

Association Rule Hiding approaches can be classified into five classes: heuristic based approaches, border based approaches, exact approaches, reconstruction based approaches and cryptography based approaches. Some of the earliest work on the challenges of association rule mining for database security may be found in M. Attallah et al. [4]

A. Heuristic Based Approaches:

This approach is further divided into two techniques: i) Data distortion technique and ii) Data Blocking Technique.

Distortion: In distortion [12], the entry for a given transaction is modified to a different value. Since, we are typically dealing with binary transactional data sets, the entry value is flipped.

Blocking: In blocking [16], the entry is not modified, but is left incomplete. Thus, unknown entry values are used to prevent discovery of association rules. We note that both the distortion and blocking processes have a number of side effects on the non-sensitive rules in the data. 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.

B. Border Based Approaches:

Sun and Yu [17] were the first to propose 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. 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.

C. Exact Approaches:

This approach is better than other approaches but requires high time complexity. In this approach minimally extends the original database by a synthetically generated database called extended database. Gkoulalas-Divanis and Verykios[1] introduced the first exact methodology to perform sensitive frequent itemset hiding based on the notion of a hybrid database generation.

D. Reconstruction Based Approach:

This approach is implemented by perturbing the data first and reconstructing the distributions at an aggregate level in order to perform the association rules mining. Mielikainen [18] was the first analysed the computational complexity of inverse frequent set mining and showed in many cases that the problems are computationally difficult. In this approach it first places the original data aside and start from knowledge base. To sanitize, it conceals the sensitive rules by sanitizing itemset lattice rather than sanitizing original dataset.

E. Cryptography Based Approaches:

These approaches are used for multiparty computation, when database is distributed among several sites. Multiple parties may wish to share their private data, without leaking any sensitive information at their end. This approach is categorised as: vertically partitioned distributed data and horizontally partitioned distributed data. Many existing encryption techniques invented which are used to avoid the information theft. In recent days of wireless communication, the encryption of data plays a major role in securing the data in online transmission focuses mainly on its security across the wireless. Different encryption techniques are used to protect the confidential data from unauthorized use.

Encryption is a very common technique for promoting the information security. The evolution of encryption is moving towards a future of endless possibilities. Everyday new methods of encryption techniques are discovered.

The concept of privacy preserving in data mining came in to existence in response to the concerns that were raised for preserving the private information which are produced as a result of data mining algorithms [14]. There are two types of privacy concern that were raised in reference to the data mining. The first type of privacy concern termed as output privacy is that the data is minimally altered so that the mining result will preserve privacy. Many techniques have been proposed for this type of output privacy. Techniques like blocking; perturbation, aggregation, swapping, and sampling are the example of output privacy. In output privacy for hiding a given specific rules or patterns, there are many proposed techniques available for hiding association rule, classification and clustering rules. For hiding the association rules, two approaches have been proposed. The first approach that has been proposed hides one rule at a time [11]. It first selects transactions that contain the items in a given rule. It then attempts to modify transaction by transaction until the support or confidence of the rule fall below minimum support or minimum confidence. The modification is done by either deleting items from the transaction or adding new items to the transactions.

The second type of privacy concern which is related with the input privacy of the data is that the data is altered in such a way that the mining result is not affected or affected minimally[7], like cryptography-based techniques which allow users access to only a subset of data while global data mining results can still be discovered. The example includes multiparty computation. The second approach deals with groups of restricted patterns or association rules at a time [15]. It first selects the transactions that contain the intersecting patterns of a group of restricted patterns. After that on the basis of disclosure threshold supplied by users, it hides the restricted patterns by sanitizing the percentage of the selected transactions. In [10] authors summarize the advantages and limitations of associations hiding approaches.

Table I Summary of Association Rule Hiding Approaches [10]

Technique	**Advantages**	**Limitation**
Heuristic Based Approaches (Distortion technique)	Efficiency, scalability and quick responses due to which it is getting focus by majority of the researchers. Totally takes best decision	Produce undesirable side effects in new database (i.e. Lost rules and new rules).
Heuristic Based Approaches (Blocking technique)	Maintains truthfulness of the underlying data. Minimizes side effects.	Difficult to reproduce original dataset.
Border Based Approaches	Maintains data quality by greedily selecting the modification with minimal side effects. Improvement over pure heuristic approach.	Unable to identify optimal hiding solution But still dependent on heuristic to decide upon the item modification.
Exact Approaches	Guarantees quality for hiding sensitive information than other approaches.	But requires very high time complexity due to integer programming
Reconstruction Approaches	Create privacy aware database by exacting sensitive characteristic from the original database. Lesser side effects in database than heuristic approach.	The open problem is to restrict the number of transactions in the new database.

Cryptographic Approaches	Secure mining of association rule over partitioned database.	Do not protect the output of a computation. Falls short of providing a complete answer to the problem of privacy preserving data mining. Communication and computation cost should be low.

C N Modi et al. [5] proposed a heuristic algorithm named DSRRC (Decrease Support of R.H.S. item of Rule Clusters) which was able to hide many sensitive association rules at a time. They have analyzed experimental results for DSRRC, which show that performance of the DSRRC algorithm is better than other existing heuristic approaches. They have achieved improvement in misses cost, artifactual patterns, dissimilarity and maintain data quality in comparison to Algorithm 1b of [7]. This approach was able to hide only the rules that contain single item on R.H.S. of the rule. Nikunj et al. [6] introduced a heuristic based algorithm named MDSRRC (Modified Decrease Support of R.H.S. item of Rule Clusters) to hide sensitive association rules with multiple items on L.H.S and R.H.S. This algorithm is the improved version of DSRRC [5]. This algorithm does modification on minimum number of transaction in database in order to hide maximum sensitive rules and also to maintain data quality. They have also showed the performance comparison between DSRRC and MDSRRC.

IV. Comparative analysis of association rule hiding algorithms

In this section the table shows the comparative analysis the various association rule hiding algorithms on the basis of theoretical study.

Table II Comparative Analysis of Algorithm

Method of Rule Hiding	**Name of Algorithm**	**Item Hiding (LHS or RHS)**	**Rule Hiding Algorithm**
By Adding the Sensitive Item Set	ISL	LHS	
	DCIS	RHS	
	Algorithm1.a		YES
By Deletion of Sensitive item set	DSR	LHS	
	DCDS	RHS	
	DSC	BOTH	
	NAÏVE	BOTH	
	MinFIA	BOTH	
	MaxFIA	BOTH	
	Algorithm 1.b		YES
	Algorithm 2.a		YES
	DSRRC		YES
	MDSRRC		YES

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 non-sensitive rules which are hidden as side effects.

V. Conclusion

In this paper, we have presented comprehensive survey on the list of existing association rule hiding techniques to hide sensitive item set without revealing pattern. Association rule hiding is an important concept in the area of privacy preserving data mining. It protects the privacy of sensitive information in databases against the association rule mining approaches. The main aim of association rule hiding algorithms is to reduce the modification on original database in order to hide sensitive knowledge, deriving non sensitive knowledge and do not producing some other knowledge. Existing approaches provide only the approximate solution to hide sensitive knowledge. There is need of finding exact solution to the privacy problem in database disclosure. In future, hybrid technique can be found to reduce the side effects and increase the efficiency by reducing the modifications on database, while hiding the association rules.

VI. References

[1] A. Gkoulalas-Divanis, and V. S. Verykios, "*Exact knowledge hiding through database extension*" IEEE Trans Knowledge Data Eng 2009, pp. 699–713.

[2] Aggarwal, Charu C., and S. Yu Philip. "*Privacy-Preserving Data Mining: A Survey*." Handbook of Database Security. Springer US, 2008. 431-460.

[3] Agrawal, Rakesh, and Ramakrishnan Srikant. "*Fast algorithms for mining association rules.*" Proc. 20th int. conf. very large data bases, VLDB. Vol. 1215. 1994.

[4] Atallah, Mike, et al. "*Disclosure limitation of sensitive rules.*" Knowledge and Data Engineering Exchange, 1999.(KDEX'99) Proceedings. 1999 Workshop on. IEEE, 1999.

[5] C. N. Modi, U. P. Rao, and D. R. Patel, "*Maintaining privacy and data quality in privacy preserving association rule mining,*" 2010 Second International conference on Computing, Communication and Networking Technologies, pp. 1–6, Jul. 2010.

[6] Domadiya, Nikunj H., and Udai Pratap Rao. "*Hiding sensitive association rules to maintain privacy and data*

quality in database." Advance Computing Conference (IACC), 2013 IEEE 3rd International. IEEE, 2013.

[7] E. Dasseni, V. Verykios, A. Elmagarmid & E. Bertino, "*Hiding association rules by using confidence and support*" In Proceedings of 4th information hiding workshop, Pittsburgh,2001.

[8] Garg, Vikram, Anju Singh, and Divakar Singh. "*A Survey of Association Rule Hiding Algorithms*." Communication Systems and Network Technologies (CSNT), 2014 Fourth International Conference on. IEEE, 2014.

[9] J. Han, and M. Kamber, *Data Mining: Concepts and Techniques*, pp. 227–245. Morgan Kaufmann Publishers, San Francisco, 2001.

[10] Jadav, Khyati B., Jignesh Vania, and Dhiren R. Patel. "*A Survey on Association Rule Hiding Methods*." International Journal of Computer Applications 82.13 (2013): 20-25.

[11] Komal Shah, Amit Thakkar, Amit Ganatra," *A Study on Association Rule Hiding Approaches*" International Journal of Engineering and Advanced Technology (IJEAT), February 2012.

[12] Oliveira, Stanley RM, Osmar R. Zaiane, and Yücel Saygin. "*Secure association rule sharing*." Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2004. 74-85.

[13] Padam Gulwani "*Association Rule Hiding by Positions Swapping of Support and Confidence*" International journal of Information Technology and Computer Science, 2012

[14] R. Agrawal & R. Srikant, "*Privacy preserving data mining*" In ACM SIGMOD conference on management of data, Dallas, Texas, May 2000.

[15] S. Oliveira & O. Zaiane, "*Algorithms for balancing privacy and knowledge discovery in association rule mining*" In Proceedings of 7 international database engineering and applications symposium (IDEAS03), Hong Kong, July 2003.

[16] Saygin, Yücel, Vassilios S. Verykios, and Chris Clifton. "*Using unknowns to prevent discovery of association rules*." ACM SIGMOD Record 30.4 (2001): 45-54.

[17] Sun, Xingzhi, and Philip S. Yu. "*A border-based approach for hiding sensitive frequent itemsets*." Data Mining, Fifth IEEE International Conference on. IEEE, 2005.

[18] 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.

[19] Verykios, Vassilios S., et al. "*Association rule hiding*." Knowledge and Data Engineering, IEEE Transactions on 16.4 (2004): 434-447.

[20] Vijayarani, S., A. Tamilarasi, and R. SeethaLakshmi. "*Privacy preserving data mining based on association rule-a survey*." Communication and Computational Intelligence (INCOCCI), 2010 International Conference on. IEEE, 2010.