Approaches for Privacy Preserving Data Mining by Various Associations Rule Hiding Algorithms – A Survey

Umesh Kumar Sahu
Student
Computer Science & Engineering
Barkatullah University
Institute of Technology, Bhopal

Anju Singh
Asst. Prof.
Department of IT
Barkatullah University
Institute of Technology, Bhopal

ABSTRACT

Yesteryears, data mining has emerged as a very popular tool for extracting hidden knowledge from collection of huge amount of data. Major challenges of data mining are to find the hidden knowledge in the data while the sensitive information is not revealed. Many Industry ,Defence ,Public Sector and Organisation facing risk or having security issue while sharing their data so it is very crucial concern how to protect their sensitive information due to legal and customer concern. Many strategies have been proposed to hide the information containing sensitive data. Privacy preserving data mining is an answer to such problems. Association rule hiding is one of the PPDM techniques to protect the sensitive association rule .In this paper, all the approaches for privacy preserving data mining have been compared theoretically and points out their pros and cons.

Keywords

Data Mining, Privacy Preserving, sensitive information Association Rule Hiding.

1. INTRODUCTION

Privacy preserving data mining (PPDM) is a fruitful research area in Data Mining (DM), where DM algorithms are analyzed and compared the impacts which occur in data privacy. The goal of PPDM is to transform the existing dataset in some way that the confidentiality of the data and knowledge remains intact even after the mining process. In DM, the users are given the data and they are free to use their own tools. So, the manipulation for privacy has to be applied on the data itself before the mining process. For this reason, there is a need to develop processes that can take us to new privacy preserving control systems to convert a given dataset into a new one in such a way to preserve the general rules mined from the original database. The goal of the proposed Association rule hiding algorithm is to hide certain information from the dataset so that it cannot be discovered through association rule mining algorithm. For example, government wants to launch some new schemes for the development of rural areas. The rural department maintains database of farmers and labours. They wants to analyse the data with help of third party without revealing the personal detail of the farmer and labours. Another example, where shopping malls are trying to understand the purchasing behaviour of the customer. In this case the data items related to individuals is not important, but the knowledge derived from the database is required to be protected.

Data mining is a technique to extract useful information from large data sets by analyzing it. In the current social scenario, sharing and publishing the information has been a common practice for their wealth of opportunities. However, the process of data collection and data sharing may lead to disclosure of their privacy. The privacy preserving data mining (PPDM) has received a tremendous amount of attention in the research literature in the recent past. A lot of techniques have been proposed to achieve the expected goal of privacy preservation. The paper will discuss, different privacy preservation techniques and their advantages and disadvantages. T also discuss some of the popular data mining algorithms like association rule mining.

Data mining have capability of analyzing huge amount of information and knowledge within a short time and intelligent algorithms puts the sensitive and confidential information that resides in large and distributed data stores at risk.The knowledge discovered by various data mining techniques may contain some sensitive information about an individual or organization. Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from the given database. The problem is usually decomposed into two sub problems, one is to find those item set whose occurrence exceeds a predefined threshold in the dataset; those item set are called frequent and large item set. Second sub problem is to generate association rules from those large item set with the constraints of minimal confidence. Association rule hiding refers to the process of modifying the original database in such a way that certain sensitive association rules disappear without seriously affecting the data and the non-sensitive rules.

The process of transformation of the given dataset into a dataset such that it hides some sensitive item sets or rules is called the sanitization process. To make this transformation, a small number of transactions have to be transformed by deleting one or more item sets or even adding noise to the data by turning some items from false to true in some transactions. The released database is called the sanitized database. On one hand, this approach slightly modifies some data, but this is perfectly acceptable in some real applications

The next section explains the approaches of association rule hiding. The Section 3 explains the aim of association rule hiding. Next section is about literature survey on privacy preserving association rule mining. section 5 will be followed by comparative analysis of various rule hiding algorithm.

2. TECHNIQUES OF ASSOCIATION RULE HIDING ALGORITHM

Association rule hiding algorithms prevents the sensitive rules from being disclosed. The problem of association rule hiding can be stated as follows: “Given a transactional database X with minimum confidence, minimum support and a set r of rules which have been mined from database X. A subset R_H of R is denoted as set of sensitive association rules which have to be preventing from being disclosed. The objective of association rule hiding is to transform X into a database X’ in such a way that nobody will be able to mine association rule which belongs to r_H and all non-sensitive rules in r should remain unaffected[7].

Privacy preserving association rule hiding algorithms can be commonly divided into three sections.

2.1 Heuristic-Based Techniques

Heuristic-based techniques resolve how to identify proper data sets for data transformation. The methods of Heuristic based transformation include perturbation, which is accomplished by the alteration of an attribute value by a new value (i.e., changing a 1-value to a 0- value, or adding noise), and blocking, which is the replacement of an existing attribute value with a “?”. Some of the approaches used are as follows.

2.1.1 Distortion Based Methods

The heuristic proposed for the modification of the data is based on data perturbation. It changes a selected set of 1-values to 0-values, so that the support of sensitive rules is reduced in such that the utility of the released database is kept to some maximum value. The key question of this algorithm is how to change X into X' with the use of heuristic thought.

Agrawal and Srikant [1] used data distortion techniques for transformation of the data items so that the approximate original data distribution could be obtained from the transformed version of the data sets. The mined rules also were approximate of the original rules. The expectation based maximization with distortion for reconstructing the original data distribution [2]. This reconstructed distribution is used to construct a classification model

The authors proposed five algorithms. All of these algorithms fall in the category of distortion based technique. Three algorithms were aimed towards hiding association rules. Remaining two algorithms were related to hiding large item sets. Metrics used in all of these five algorithms were efficiency and side effects. These algorithms were first of their kind in hiding association rules. Side effects of these algorithms were also high [3].

The authors in [4] aims at balancing privacy and disclosure of data items by trying to minimize the impact on sanitized transactions and also to minimize the accidentally hidden and ghost rules. The utility in this work is measured as the number of non-sensitive rules that were hidden based on the side-effects of the data modification process.

2.1.1 Blocking-Based Methods

By reducing the degree of support and confidence of the sensitive association rules by transforming certain data items of some data sets with a question mark or a true value, the approach of blocking is implemented. The minimum support and minimum confidence will be altered into a minimum support interval and a minimum confidence interval correspondingly. If the support and/or the confidence of a sensitive rule lie in the middle of these two ranges of values, the confidentiality of data is not violated

Yucel Saygin [8][9] use blocking for the association rule confusion. After the original data is replaced with some data of unknown value, it is difficult to determine the support and confidence of sensitive association rule, which may be a range of arbitrary values. The paper proposed by Yucel Saygin [8] discusses specific examples with the use of an uncertain symbol used in association rule mining, in which case the support and confidence interval are used to replace support and confidence.

Xiao X. [10] presents a new generalization framework on the concept of personalized anonymity in order to perform minimum generalization for satisfying everybody’s requirements. It provides privacy protection of different size for the records of data table. Liu Mingetal[11] proposes a personalized anonymity model on the basis of (α,k)-anonymization model in order to resolve the problem of privacy self-management. They propose corresponding anonymity method by using local recoding and sensitive attribute generalization.

2.2 Reconstruction-Based Association Rule

A number of recently proposed techniques address the issue of privacy preservation by perturbing the data and reconstructing the distributions at an aggregate level in order to perform the association rules mining. That is, these algorithms are implemented by perturbing the data first and then reconstructing the distributions. According to different methods of reconstructing the distributions and data types, the corresponding algorithm is not the same.

Agrawal [12] used Bayesian algorithm for distribution reconstruction in numerical data. Then, Agrawal [8] proposed a uniform randomization algorithm for reconstruction-based association rule to deal with categorical data items. The authors of [7] improved the work over the Bayesian-based reconstruction procedure with the help of an EM (Expectation Maximization) algorithm for distribution reconstruction.

2.2.1 Data reconstruction methods:

Another variation of data reconstruction methods put the original data aside and starts from sanitizing the so-called “knowledge base”. The new released data is then reconstructed from the sanitized knowledge base. Chen [13] first proposed a Constraint-based Inverse Item set Lattice Mining procedure (CIILM) for hiding sensitive frequent itemsets.Their data reconstruction is based on item set lattice. Another emerging privacy preserving data sharing method related with inverse frequent item set mining is inferring original data from the given frequent item sets. This idea was first proposed by Mielikainen[14].

2.2.2 FP tree method:

A FP-tree based method is presented in [14] for inverse frequent set mining which is based on reconstruction technique. It is a three phase process: The phase one uses frequent item set mining algorithm to generate all frequent item sets with their supports and support counts from data set D. The phase two runs sanitization algorithm over frequent item set FS and get the sanitized frequent item sets of FS’. The third phase is to generate released database D’ from FS’ by using inverse frequent set mining algorithm.

2.3 Cryptography-Based Techniques

Different parties wish to exchange their data, without disclosing any sensitive information. So there is a need for secure and cryptographic protocols for exchanging the information over the different parties.

2.3.1 Vertically Partitioned Distributed Data:

The algorithm follows the concept of "secure sum" for the secure counting of inter-site, the sum of support degree of every sub-item sets which are distributed in different sites is counted. The item set is measured as global frequent item set if its support is greater than the threshold.

Various methods for distributed privacy-preserving data mining is discussed in [15]. These methods include the secure sum, the secure set union, the secure size of set intersection and the scalar product. The methods in [16] discuss how to use to scalar dot product computation for frequent item set counting. This uses secure protocol for calculating the dot-product of two vectors by using linear algebraic techniques.it describe superior performance in terms of computational overhead, numerical stability, and security by using analytical as well as experimental results.

2.3.2 Horizontally Partitioned Distributed:

The main concept to measure global frequent item sets, while ensuring non discloser of inter-site information. It only find the secure sum of support degree inter-sites data items. Thus the overall item sets support degree is founded. The item sets with support degree greater than threshold are the global frequent item sets.

Shaofei Wu [17] proposed an algorithm to balance privacy preserving and knowledge discovery in association rule mining. The solution uses a filter after the mining phase to hide the restricted discovered association rules. Before implementation the algorithms, the data structure of database and sensitive association rule mining set have been analyzed to build an effective model.

Chirag N. Modi [18] proposed an algorithm that provides privacy and security against involving parties and other parties (adversaries) who can receive information via unsecured medium.

2.4 Exact approaches

These approaches follows the hiding process as a constraints satisfaction problem which is solved by binary integer programming (BIP). These approaches gives better solution. But they suffer from high time complexity to CSP.

Gkoulalas and Verykios [19] proposed an approach for finding optimal solution for hiding the rule problem which tries to minimize the distance between the original data set and its sanitized data set.

The authors in [20] proposed a border-based approach that provides an optimal solution to hide the sensitive frequent item sets by extending the original data set by a synthetically generated data set. Extending the original data set for sensitive item set 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.

2 AIM OF ASSOCIATION RULE HIDING

Aim of association rule hiding at sanitizing the original database in order to achieve the following objective [21]

a) No rule that is considered as sensitive that can be mined from the original database at pre-specified thresholds of confidence and support. It can be also revealed from the transformed data set, when this database is mined at the same or at higher thresholds. This requires that all the sensitive rules disappear from the transformed data set, when the data set is mined under the same or higher levels of support and confidence as the original data set.

b) The non-sensitive rules which are mined from the original database at given thresholds of confidence and support can be successfully mined from the transformed data set at the same thresholds or higher. The second objective states that there should be no lost rules in the transformed data set. That is, all the non-sensitive rules that were mined from the original database should also be mined from its sanitized counterpart at the same or higher levels of confidence and support

c) The rules which are not derived from the original database when the database was mined at given thresholds of confidence and support, can be derived from its transformed counterpart when it is mined at the same or at higher thresholds. The third objective states that no false rules also known as ghost rules should be produced when the sanitized database is mined at the same or higher levels of confidence and support. A ghost rule is an association rule that was not among the rules mined from the original database.

The privacy preserving association rule mining algorithms should

1. Privacy of sensitive information is maintained.
2. Not compromise the access and the use of non-sensitive data.
3. Not have an exponential computational complexity

Association rule hiding has been widely researched along two principal directions.

1. The first variant includes approaches that aim at hiding specific association rules among those mined from the original database.

2. The second variant includes approaches that hide specific frequent item sets from those frequent item set found by mining original data set. By ensuring that the item sets which are generation of a sensitive rule become insignificant in the disclosed data set, the data owner can be certain that his or her sensitive knowledge is adequately protected from untrusted third parties.

3 LITERATURE SURVEY

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 [22][23]. There are two types of privacy concern that were raised in reference to the data mining. The first type of privacy is called output privacy that the data is minimally changes so that the mining result will maintain privacy. Many algorithms have been proposed for this type of output privacy [22][24].Techniques like blocking, perturbation, aggregation, swapping, and sampling are the example of output privacy. For hiding the association rules, two approaches have been proposed. The first approach that has been proposed hides one rule at a time [25]. 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 transformation is done by either deleting items 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 changed in such a way that the mining result is remains unaffected or minimally affected [5], like cryptography-based techniques in which 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 [10]. 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 [6] authors summarize the advantages and limitations of associations hiding approaches.

In [8] the authors discussed three algorithms for hiding sensitive association rules. First one hides association rules by increasing the support of the rule's antecedent until the rule confidence decreases below the minimum confidence threshold. Second algorithm hides sensitive rules by decreasing the frequency of the consequent until either the confidence or the Support of the rule is below the threshold. Third algorithm 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 first algorithm large number of new frequent item sets is introduced and therefore, an increasing number of new rules are generated. The other two algorithms affects number of no sensitive rules in database due to removal of items from transaction

In [11] the authors discussed about ISL and DSR. Item sets are given as input to both the algorithms to automatically hide sensitive association rules without mining and selection of hidden rules. In [12] authors proposed two algorithms, DCIS and DCDS were introduced which automatically hides association rules without pre-mining and selecting hidden rules. The ISL and DCIS algorithms try to increase the support of left hand side of the association rule and algorithms DSR and DCDS try to decrease the support of the right hand side of the association rule.

It is found that the complexity of ISL is more than DSR. Also both algorithm has different side effects. In [13] an algorithm DSC is highlighted in which pattern-inversion tree is used to store all the information so that the data set is scanned only once.

In [3] authors discussed a heuristic algorithm DSRRC which provides privacy for sensitive rules at certain level while maintaining quality of data sets. DSRRC algorithm clusters the sensitive association rules based on R.H.S. of rules and hides all possible rules by modifying lesser number of transactions which maintains data quality. DSRRC algorithm cannot hide rules having multiple RHS items. In [9] the authors discussed about four heuristic algorithms: Algorithm Naïve, MinFIA, MaxFIA and IGA. The Naive Algorithm removes the entire items with the highest frequency. In MinFIA algorithm the item with the smallest support in the pattern is identified as a sensitive item and it deletes that item from the sensitive transactions. Unlike the MinFIA, algorithm MaxFIA selects the item set with the maximum support in the data

set as a sensitive item and removes it. The IGA algorithm groups does not allow the patterns in groups of patterns sharing the same item sets so that all sensitive patterns in the group will be hidden in single step.

In [4] the authors introduced an efficient algorithm known as FHSAR for hiding of sensitive association rules more rapidly. The algorithm has the capability to hide any given sensitive association rule by scanning the data set single time, which helps significantly in reducing the execution time. In [8] a Hybrid algorithm is proposed that uses the combination of ISL and DSR technique and hides the association rules by modifying the database transactions so that the confidence of the association rules can be reduced. Such approach will provide better result than using either ISR or DSR. In [7] the proposed algorithm doesn't modifying the database transactions so that the support &confidence of the association rules remains unchanged. It scans the database less number of times and prunes more number of hidden rules.

4 COMPARATIVE ANALYSIS OF VARIOS RULE HIDING ALGORITHM

Table-I Association Rule Hiding Approaches

TECHNIQUE	PROS	CONS
Heuristic Based Approaches (Distortion technique)	Efficiency, scalability and quick responses due to which it is getting focus by majority of the researchers.	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 selecting the changing 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 problem is to prevent the number of trans-actions in the new data set.
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.

In this table shows the comparative analysis the various association rule hiding algorithms study.

Table 2 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	RHS	YES
By Deletion of Sensitive item set	DSR	LHS	
	DCDS	RHS	
	DSC	BOTH	
	NAÏVE	BOTH	
	MinFIA	BOTH	
	MixFIA	BOTH	

5 CONCLUSION

Association rule hiding is the approach which highly useful while the analysis of data sets in sharing environment. It protects the privacy of sensitive information in databases against the association rule mining approaches. This paper presents a classification of privacy preserving association rule mining approaches is presented and major algorithms in each class are discussed. The pros and cons of different techniques are also analyzed on the basis of decreasing and increasing the support and confidence of item sets.

It also presents a comprehensive survey on the list of existing association rule hiding techniques to hide sensitive item set without revealing pattern. 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.

6 REFERENCES

[1] Alexandre Evfimievski, Ramakrishnan Srikant, Rakesh Agrawal, Johannes Gehrke. Privacy Preserving Mining of Association Rules. SIGKDD 2002, Edmonton, Alberta Canada.

[2] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms", In Proceedings of the 20th Symposium on Principles of Database Systems, Santa Barbara, California, USA, May, 2001.

[3] Stanley R. M. Oliveira and Osmar R. Zaiane, "Privacy preserving frequent itemset mining, InProceedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining (2002), pp.43–54.

[4] S.R.M. Oliveira, O.R. Zaıane, Y. Saygin, "Secure association rule sharing, advances in knowledge discovery and data mining, in: Proceedings of the 8th Pacific-Asia Conference (PAKDD2004), Sydney, Australia, 2004, pp.74–85.

[5] 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.

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

[7] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms", In Proceedings of the 20th Symposium on Principles of Database Systems, Santa Barbara, California, USA, May, 2001.

[8] Alexandre Evfimievski, Ramakrishnan Srikant, Rakesh Agrawal, Johannes Gehrke. Privacy Preserving Mining of Association Rules. SIGKDD 2002, Edmonton, Alberta Canada.

[9] Yucel Saygin, Vassilios S. Verykios, and Ahmed K. Elmagarmid, "Privacy preserving association rule mining," In Proceedings of the 12th International Workshop on Research Issues in Data Engineering (2002), 151–158.

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

[11] Shyue-Liang Wang, Bhavesh Parikh, Ayat Jafari "Hiding informative association rule sets", ELSEVIER, Expert Systems with Applications 2007.

[12] Rakesh Agrawal and Ramakrishnan Srikant, "Privacy-preserving data mining," In Proceedings of the ACM SIGMOD Conference on Management of Data (2000), pp.439–450.

[13] 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. 47-56.

[14] Yongcheng Luo, Yan Zhao, Jiajin Le, "A Survey on the Privacy Preserving Algorithm of Association Rule Mining", isecs, vol.1, pp.241-245, 2009

[15] Chris Clifton, Murat Kantarcioglou, XiadongLin and Michaed Y.Zhu, "Tools for privacy preserving distributed data mining," SIGKDD Explorations 4, no. 2, 2002

[16] Ioannidis, I.; Grama, A, Atallah, M., "A secure protocol for computing dot-products in clustered and distributed environments," Proceedings of International Conference on Parallel Processing, 18-21 Aug. 2002, pp.379–384.

[17] Shaofei Wu and Hui Wang ,"Research On The PrivacyPreserving Algorithm Of Association Rule Mining InCentralized Database", IEEE International Symposiums on Information Processing, 2008.

[18] Chirag N. Modi, Udai Pratap Rao and Dhiren R. Patel, "An Efficient Approach for Preventing disclosure of Sensitive Association Rules in Databases", International Conference on Advances in Communication, Network, and Computing,IEEE, 2010

[19] 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

[20] 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.

[21] Gkoulalas-Divanis, Aris, Verykios, Vassilios S. “Association Rule Hiding for Data Mining”,Springer Series: Advances in Database Systems, Vol. 41, 1st Edition., 2010, p.13.

[22] C. Clifton, “Protecting against data mining through samples” In Proceedings of the thirteenth annual IFIP WG 11.3 working conference on database security, 1999.

[23] R. Agrawal & R. Srikant, “Privacy preserving data mining” In ACM SIGMOD conference on management of data, Dallas, Texas, May 2000

[24] C. Clifton, “Using sample size to limit exposure todata mining” Journal of Computer Security, 2000.

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