A SURVEY ON PRIVACY PRESERVING ASSOCIATION RULE MINING

K. Sathiyapriya[1] and Dr. G. Sudha Sadasivam[2]

Department of Computer Science and Engineering ,
PSG College of Technology, Coimbatore, India
[1]sathya_jambai@yahoo.com
[2]sudhasadhasivam@yahoo.com

ABSTRACT:
Businesses share data, outsourcing for specific business problems. Large companies stake a large part of their business on analysis of private data. Consulting firms often handle sensitive third party data as part of client projects. Organizations face great risks while sharing their data. Most of this sharing takes place with little secrecy. It also increases the legal responsibility of the parties involved in the process. So, it is crucial to reliably protect their data due to legal and customer concerns. In this paper, a review of the state-of-the-art methods for privacy preservation is presented. It also analyzes the techniques for privacy preserving association rule mining and points out their merits and demerits. Finally the challenges and directions for future research are discussed.

KEYWORDS :
Privacy preservation, Association rule mining, Rule hiding, Data blocking, Data perturbation.

1. INTRODUCTION

Recent advances in data mining and knowledge discovery have generated controversial impact in both scientific and technological arenas. On the one hand, data mining is capable of analyzing vast amount of information within a minimum amount of time. On the other hand, the excessive processing power of intelligent algorithms puts the sensitive and confidential information that resides in large and distributed data stores at risk.

Providing solutions to database security problems combines several techniques and mechanisms. An organization may have data at different sensitivity levels. This data is made available only to those with appropriate rights. Simply restricting access to sensitive data does not ensure complete sensitive data protection. Based on the knowledge of semantics of the application, the user may infer sensitive data items from non-sensitive data. Such a problem is known as 'Inference Problem'.

Association rule mining is a technique in data mining that identifies the regularities found in large volume of data. Such a technique may identify and reveal hidden information that is private for an individual or organization.

Consider the case of an airport government security service that may be interested in developing a system for identifying passengers whose baggage must be subjected to additional security measures. The data indicating the necessity for further examination is derived from a wide variety of sources such as police records, airports, banks, general government statistics and passenger information records. These records include personal information such as name and passport number, demographic data such as age and gender; flight information such as departure,

DOI : 10.5121/ijdkp.2013.3208

destination, and duration; and expenditure data such as transfers, purchasing and bank transactions. In most countries, this information is regarded as private and exposing intentionally or unintentionally confidential information about an individual is against the law.

In order to preserve privacy, passenger information records can be de-identified before the records are shared. This can be accomplished by deleting unique identity fields, such as name and passport number from the dataset. However, even if this information is deleted, there are still other kinds of information like personal or behavioral information. It includes date of birth, zip code, gender, number of children, number of calls, number of accounts. When this information is linked with other available datasets, it could potentially identify subjects. To avoid such exposure of sensitive information, algorithm for privacy preservation in association rule mining becomes a must.

Large number of research papers are available in this field, each tackling the problem in different angle using different techniques. Most of the methods result in information loss and side-effects. The side effect includes falsely hiding non sensitive rules and falsely generated spurious rules. so it is important to organize these papers into categories in such a way to identify the merits and demerits of different rule hiding techniques.

The rest of this paper is organized as follows: Section 2 introduces association rule mining strategies; Section 3 the goal of association rule hiding methodologies. Section 4 surveys privacy preserving association rule mining techniques. Section 5 provides the evolution and recent scenarios. Section 6 analyses the major rule hiding algorithms in terms of side effects and quality of database. Section 7 concludes and provides directions for future research.

2. ASSOCIATION RULE MINING STRATEGY

Association rules are an important class of regularities within data which have been extensively studied by the data mining community. The problem of mining association rules can be stated as follows: Given $I = \{i1 , i2 , ... , im \}$ is a set of items, $T = \{t1, t2 , ... , tn\}$ is a set of transactions, each of which contains items of the itemset I . Each transaction t_i is a set of items such that $t_i \subseteq I$. An association rule is an implication of the form: $X \rightarrow Y$, where $X \subset I$, $Y \subset I$ and $X \cap Y = \emptyset$. X (or Y) is a set of items, called itemset. In the rule $X \rightarrow Y$, X is called the antecedent, Y is the consequent. It is obvious that the value of the antecedent implies the value of the consequent. The antecedent, also called the "left handside" of a rule, can consist either of a single item or of a whole set of items. This applies for the consequent, also called the "right hand side", as well. Often, a compromise has to be made between discovering all itemsets and computation time. Generally, only those item sets that fulfill a certain support requirement are taken into consideration. Support and confidence are the two most important quality measures for evaluating the interestingness of a rule.

The support of the rule $X \rightarrow Y$ is the percentage of transactions in T that contain $X \cap Y$. It determines how frequent the rule is applicable to the transaction set T . The support of a rule is represented by the formula

$$\text{supp}(X \rightarrow Y) = \frac{|X \cap Y|}{n}$$

where $|X \cap Y|$ is the number of transactions that contain all the items of the rule and n is the total number of transactions.

The confidence of a rule describes the percentage of transactions containing X which also contain Y . It is given by

$$\text{conf}\,(X \rightarrow Y) = \frac{|X \cap Y|}{X}$$

Confidence is a very important measure to determine whether a rule is interesting or not. The process of mining association rules consists of two main steps. The first step is, identifying all the itemsets contained in the data that are adequate for mining association rules. These combinations have to show at least a certain frequency and are thus called frequent itemsets. The second step generates rules out of the discovered frequent itemsets. All rules that has confidence greater than minimum confidence are regarded as interesting.

3. GOAL OF ASSOCIATION RULE HIDING METHODOLOGIES

Association rule hiding methodologies aim at sanitizing the original database in order to achieve the following goals[39]:

1. No rule that is considered as sensitive from the owner's perspective and can be mined from the original database at pre-specified thresholds of confidence and support, can be also revealed from the sanitized database, when this database is mined at the same or at higher thresholds
2. All the non sensitive rules that appear when mining the original database at prespecified thresholds of confidence and support can be successfully mined from the sanitized database at the same thresholds or higher.
3. No rule that was not derived from the original database when the database was mined at pre-specified thresholds of confidence and support, can be derived from its sanitized counterpart when it is mined at the same or at higher thresholds.

The first goal requires that all the sensitive rules disappear from the sanitized database, when the database is mined under the same or higher levels of support and confidence as the original database.

The second goal states that there should be no lost rules in the sanitized database. 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.

The third goal 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 false (ghost) rule is an association rule that was not among the rules mined from the original database.

A solution that addresses all these three goals is called exact. Exact hiding solutions that cause the least possible modification to the original database are called ideal or optimal. Non-exact but feasible solutions are called approximate.

The privacy preserving association rule mining algorithms should 1. prevent the discovery of sensitive information. 2. not compromise the access and the use of non sensitive data. 3. be usable on large amounts of data. 4. not have an exponential computational complexity.

Association rule hiding has been widely researched along two principal directions. The first variant includes approaches that aim at hiding specific association rules among those mined from the original database. The second variant includes approaches that hides specific frequent itemsets from those frequent itemset found by mining original database. By ensuring that the itemsets that lead to the generation of a sensitive rule become insignificant in the disclosed database, the data owner can be certain that his or her sensitive knowledge is adequately protected from untrusted third parties.

The common approaches used in association rule hiding algorithms are 1) Heuristic approaches, 2) Border-based approaches and 3) Exact approaches. The Heuristic approaches are used to modify the selected transactions from the database for hiding the sensitive data. The Border-based approaches is the sensitive rule hiding can be done through the modification of the original borders in the lattice of the frequent and the infrequent patterns in the data set. The Exact approaches are non-heuristic algorithms which envisage the hiding process as a constraint satisfaction problem that may be solved using integer programming or linear programming.

Hiding the sensitive association rules by hiding their generating itemsets is a common strategy adopted by the majority of researchers.

4. SURVEY ON PRIVACY PRESERVING ASSOCIATION RULE MINING (PPAM)

Let D be the source database, R be a set of significant association rules that can be mined from D, and let Rh be a set of rules in R. Privacy preserving association rule algorithms transform database D into a database D', the released database, so that all rules in R can still be mined from D', except for the rules in Rh. Based on the privacy protection technologies used, privacy preserving association rule mining algorithms can be commonly divided into three categories.

a) Heuristic-Based Techniques :Heuristic-based techniques resolves how to select the appropriate data sets for data modification. Since the optimal selective data modification or sanitization is an NP-Hard problem, heuristics is used to address the complexity issues. The methods of Heuristic-based modification 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.

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 lowered in such a way that the utility of the released database is kept to some maximum value. The key question of this algorithm is how to change D into D' with the use of heuristic thought.

Agrawal and Srikant[1] used data distortion techniques to modify the confidential data values so that the approximate original data distribution could be obtained from the modified version of the database. The mined rules also were approximate of the original rules. Agrawal and Aggarwal[2], used expectation maximization with distortion for reconstructing the original data distribution. This reconstructed distribution is used to construct a classification model.

The authors in [3] presented five algorithms namely 1.a, 1.b, 2.a, 2.b, 2.c. All of these algorithms fall in the category of distortion based technique. Algorithms 1.a, 1.b, and 2.a were aimed towards hiding association rules. Algorithms 2.b, 2.c were related to hiding large itemsets. 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.

Stanley R. M. Oliveira[4] aims at balancing privacy and disclosure of information 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. A sanitization technique is presented by the authors to block forward inference attack and backward inference attack to hide sensitive rules.

The work described by Elena Dasseni[5] extends the sanitization of sensitive large itemsets to the sanitization of sensitive rules. Multiple rule hiding approach is first proposed by the authors in [6]. In this work authors propose strategies and a set of algorithms for hiding sensitive knowledge

from data by minimally perturbing their values. These algorithms are efficient and require two scans of the database irrespective of the number of sensitive items to hide. The hiding strategies proposed are based on reducing the support and confidence of rules that specify how significant they are. The constraint on the algorithms proposed is that the changes in the database introduced by the hiding process should be limited, in such a way that the information incurred by the process is minimal. Wang et al.[7] also proposed an approach to avoid Forward-Inference Attacks, in the sanitized database generated by the sanitization process. Techniques like WSDA, PDA [8] and Border- Based [9] improved the initial heuristic algorithms to greedy algorithms which finds local optimal modification. WSDA technique uses priority values assigned to transactions based on weights for choosing the optimal transaction for modification. These approaches tried to greedily select the modifications with minimal side effects on data utility and accuracy.

Different heuristics use different selection strategies to choose transactions and items for sanitization, as these are the two core issues that affect the hiding effects in the algorithms.

Blocking-Based Methods : The approach of blocking is implemented by reducing the degree of support and confidence of the sensitive association rules by replacing certain attributes of some data items with a question mark or a true value. In this regard, the minimum support and minimum confidence will be altered into a minimum support interval and a minimum confidence interval correspondingly. As long as the support and/or the confidence of a sensitive rule lies in the middle of these two ranges of values, the confidentiality of data is not violated. Yucel Saygin et al.[10][11] 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 et al.[10] 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. et al.[12] 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 Ming et al.[13] 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.

b) 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 et al. [14] used Bayesian algorithm for distribution reconstruction in numerical data. Then, Agrawal et al.[1] proposed a uniform randomization approach on reconstruction-based association rule to deal with categorical data. Before sending a transaction to the server, the client takes each item and with probability p replaces it by a new item not originally present in this transaction. This process is called uniform randomization. It generalizes Warner's "randomized response" method. The authors of [2] improved the work over the Bayesian-based reconstruction procedure by using an EM algorithm for distribution reconstruction.

Another variation of Data reconstruction methods put the original data aside and start from sanitizing the so-called "knowledge base". The new released data is then reconstructed from the sanitized knowledge base. Chen et. al. [15] first proposed a Constraint-based Inverse Itemset Lattice Mining procedure (CIILM) for hiding sensitive frequent itemsets. Their data reconstruction is based on itemset lattice. Another emerging privacy preserving data sharing

method related with inverse frequent itemset mining is inferring original data from the given frequent itemsets. This idea was first proposed by Mielikainen[16]. He showed finding a dataset compatible with a given collection of frequent itemsets is NPcomplete.

A FP-tree based method is presented in [29] for inverse frequent set mining which is based on reconstruction technique. The whole approach is divided into three phases: The first phase uses frequent itemset mining algorithm to generate all frequent itemsets with their supports and support counts from original database D. The second phase runs sanitization algorithm over frequent itemset FS and get the sanitized frequent itemsets of FS'. The third phase is to generate released database D' from FS' by using inverse frequent set mining algorithm. But this algorithm is very complex as it involves generation of modified dataset from frequent set.

c) Cryptography-Based Techniques: In many cases, multiple parties may wish to share aggregate private data, without leaking any sensitive information at their end. This requires secure and cryptographic protocols for sharing the information across the different parties.

A systematic framework is described in [17] to transform normal data mining problems to secure multi-party computation problems. The problems discussed in [17] include those of clustering, classification, association rule mining, data summarization, and generalization. The privacy preserving distributed data mining that uses cryptography based techniques are catagorised as follows:

1) Vertically Partitioned Distributed Data: Using the idea of "secure sum" for the secure calculation of inter-site, the sum of support degree of every sub-itemsets which are distributed in different sites is calculated. The itemset is determined as global frequent itemset if its support is greater than the threshold.

A set of methods for distributed privacy-preserving data mining is discussed in [18]. These methods include the secure sum, the secure set union, the secure size of set intersection and the scalar product. These techniques can be used as data mining primitives for secure multi-party computation over a variety of horizontally and vertically partitioned data sets. The methods in [20] discuss how to use to scalar dot product computation for frequent itemset counting. This work uses secure protocol for computing the dot-product of two vectors by using linear algebraic techniques. The protocol demonstrates superior performance in terms of computational overhead, numerical stability, and security by using analytical as well as experimental results.

The approach of vertically partitioned mining has been extended to a variety of data mining applications such as decision trees [19], SVM Classification [42], Naive Bayes Classifier [41], and k-means clustering [40]. A number of theoretical results on the ability to learn different kinds of functions in vertically partitioned databases with the use of cryptographic approaches are discussed in [47]. Sanil et al. [21] describe different approaches based on quadratic optimization to solve for coefficients using a form of secure matrix multiplication to calculate off-diagonal blocks of the full-data covariance matrix.

If the transactions are vertically partitioned across the sites, this problem can be solved by generating and computing a set of independent linear equations [22]. A log-linear model approach[23] for strictly vertically partitioned databases suggests general secure logistic regression for problems involving partially overlapping data bases with measurement error.

2) Horizontally Partitioned Distributed : The key idea is to find global frequent itemsets, while ensuring no leakage of inter-site information. It only calculates the secure sum of support degree inter-sites. Thus the overall itemsets support degree is acquired. The itemsets with support degree greater than threshold are the global frequent itemsets.

Kantarcioglu and Clifton[24] use a secure multi-party computation to model the horizontal partitioning of transactions across sites, incorporating cryptographic techniques to minimize the shared information without incurring much overhead in the mining process. But the cost of

mining is much higher. The communication and computation cost will dominate the other costs, as it is linear in |DB|. David W. Cheung [25] proposed an efficient distributed algorithm FDM (Fast Distributed Mining of association rules) for mining association rules.

Shaofei Wu et al. [26] proposed a new algorithm to balance privacy preserving and knowledge discovery in association rule mining. The solution is to implement a filter after the mining phase to weed out or hide the restricted discovered association rules. Before implementing the algorithms, the data structure of database and sensitive association rule mining set have been analyzed to build an effective model.

Chirag N. Modi et al.[27] proposed an algorithm that provides privacy and security against involving parties and other parties (adversaries) who can get information through the unsecured channel.

5. RECENT EVOLUTION

In recent years, numerous algorithms have been proposed for implementing privacy preservation. Wang Yan et al.[28] proposed a privacy preserving association rule mining algorithm based on Secondary Random Response Column Replacement(SRRCR). It can achieve significant improvements in terms of privacy and efficiency.

Yongcheng Luo et al. [29] categorized privacy preserving association rule mining into three categories: heuristic-based techniques, reconstruction-based techniques, cryptography-based techniques. Finally, they conclude with further research directions of privacy preserving algorithms of association rule mining by analyzing the existing work.

Dehkordi et.al [30] proposed a novel method for privacy preserving association rule mining based on genetic algorithms. It also makes sure that no normal rules are falsely hidden (lost rules) and no extra fake rules (ghost rules) are mistakenly mined after the rule hiding process using genetic algorithm. The algorithm sanitizes both rule and itemset with minimal side effects by introducing new hiding strategies.

S. Vijayarani et. al[31] uses tabu search optimization technique to modify the sensitive items for hiding the sensitive association rules. This approach has the advantage of modifying the sensitive rules accurately without affecting the non-sensitive rules and no false rules are generated. The disadvantage is that it needs several iterations for selecting the optimal transaction for modification. By developing new fitness functions and applying other optimization techniques the number of iterations can be minimized.

Mehmet Kaya[32] proposes a novel method to hide critical fuzzy association rules from quantitative data. For this purpose, the support value of LHS of the rule to be hidden is increased. The experimental results shows that the algorithm provides consistent rule hiding.

Duraisamy et al.[33] proposes an algorithm to minimally modify the database such that no sensitive rules containing sensitive items on the right hand side of the rule will be discovered. The time complexity is reduced because of clustering the sensitive rules and updating database only after all the sensitive rules are hidden. It also modifies minimum numbers of transactions and the alteration in the transactions are stopped when the confidence of the sensitive rules are reduced than the minimum confidence. But the algorithm can hide only sensitive rules with single antecedent and consequent and with the sensitive item in the consequent.

Yogendra kumar et al. [34] proposed a new algorithm that increases and decreases the support of the LHS and RHS item of the rule correspondingly in order to hide the rule. The authors claim

that the proposed algorithm is advantageous as it makes minimum modification to the data entries to hide a set of rules with lesser CPU time than the previous work.

Muhammad Naeem et.al.[35] proposes an architecture which hides the restricted association rules with complete removal of the known side effects like generation of unwanted, non genuine association rules while yielding no hiding failure. This architecture uses other standard statistical measures instead of conventional framework of Support and Confidence to generate association rules, Specifically a weighing mechanism based on central tendency is introduced.

The method proposed by Ramesh Chandra Belwel et.al.[36], has the basis of reduction of support and confidence of sensitive rules but does not edit or disturb the given database of transactions directly. Rather the same task is performed indirectly by modifying some newly introduced terms associated with database transactions and association rules. These new terms are Mconfidence (modified confidence), Msupport (modified support) and Hiding counter. since the algorithm uses modified definition of support and confidence it hides any desired sensitive association rule without any side effect. But it can hide only the rules that has single sensitive item on the left hand side.

The paper proposed by Assaf Schuster et al.[37] presents a cryptographic privacy-preserving association rule mining algorithm in which all of the cryptographic primitives involve only pairs of participants. The advantage of this algorithm is its scalability and the disadvantage is that, a rule cannot be found correct before the algorithm gathers information from k resources. Thus, candidate generation occurs more slowly, and hence the delay in the convergence of the recall. The amount of manager consultation messages is also high.

The data set initially contains both frequent and infrequent items. Total transactions could exceed the main memory limit. To deal with this problem, Tirumala prasad et.al.[38] proposed Distributed Count Association Rule Mining Algorithm(DCARM)that fragments the data set into different horizontal partitions, removes infrequent items from each partition and inserts each transaction into the main memory. The major advantage is DCARM exchanges less messages among different sites to generate globally frequent itemsets. DCARM thus reduces the communication cost by 60 to 80 percent. Reliability and performance of co - ordinator site is a bottleneck, which is a disadvantage.

6. EXPERIMENTAL ANALYSIS

In general, a rule hiding algorithm hides a rule by either decreasing the support count of frequently occurring item below minimum support or decreasing the confidence of a rule below minimum confidence. This is achieved by changing the values of the frequently occurring sensitive items in the database such that the items support goes below minimum support. The resultant sanitized database is released for mining. The goal of these algorithms is to minimize the side effects of lost rules and ghost rules while maintaining the data quality. To decrease the confidence of the rule, either the support of the items on the right hand side of the rule is decreased (DSR) or the support of the items on the left hand side of the rule is increased(ISL). So, we worked with some association rule hiding algorithms and examined their performance in order to analyze their impact in the original database. These algorithms hide the quantitative fuzzy association rules. One algorithm uses ISL approach and the other uses the DSR approach. We worked on two different side effects – one was the number of new rules generated during the hiding process and the other one was the number of non-sensitive rules lost during the process. The dataset used is Wisconsin Breast Cancer dataset from UCI Machine Learning Repository. The dataset consists of nine quantitative attributes and one categorical attribute. We used only nine quantitative attributes and ignored categorical attribute. First two experiments are conducted for 699 transactions. Figure 1 shows the total number of rules and hidden rules in both ISL and DSR approaches for varying confidence of 50, 60, 70, 80, 90, 100 and a constant support of 50.

Figure 2 shows the number of lost rules in both the approaches for varying confidence and constant support of 50.

Figure 3 shows the number of new rules generated for different number of transactions. Figure 4 shows the number of entries modified in both the approaches.

The total number of entries modified and the execution time in seconds for varying number of transactions were shown in table1 and 2 respectively

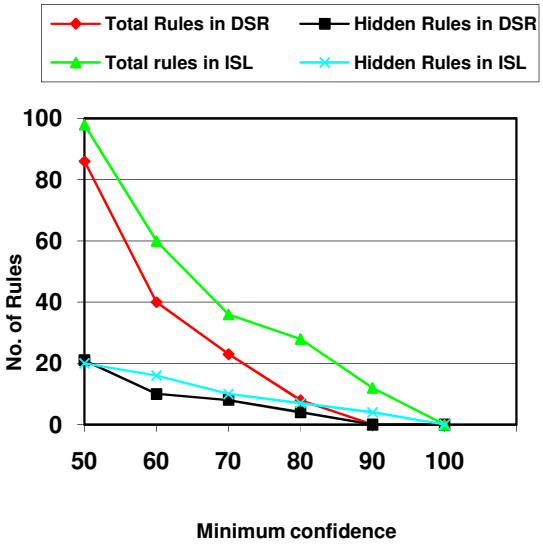

Figure 1. Total Rules Vs. Hidden Rules

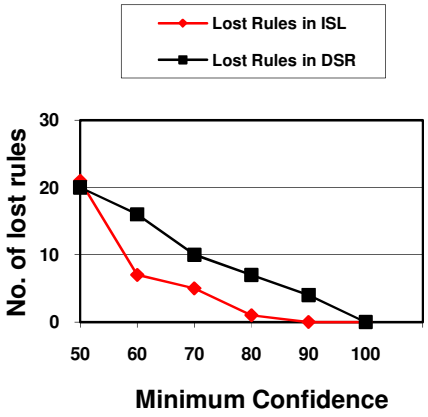

Figure 2. Lost Rules in ISL and DSR

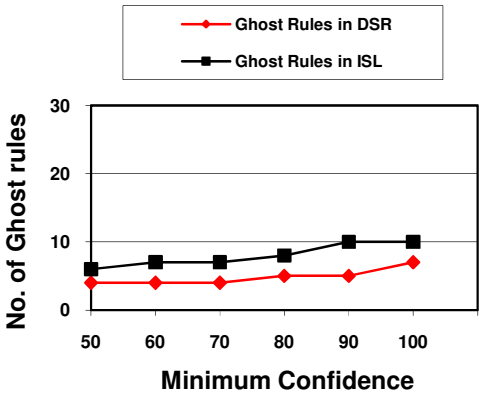

Figure 3. Ghost Rules in ISL and DSR

Table 1. Number of entries modified

Transactions	Total No. of entries	No. of entries modified	
		ISL	DSR
100	900	88	116
200	1800	172	216
300	2700	268	351
400	3600	387	456
500	4500	440	527
600	5400	258	579

Table 2. Execution time in seconds Vs. Transactions

Transactions	Time for Execution in Seconds	
	ISL	DSR
500	7	20
1000	18	42
1500	25	36

As side effects, we considered both the loss and the introduction of information in the database. We lose information whenever some rules, originally mined from the database, cannot be retrieved after the hiding process. We add information whenever some rules that could not be retrieved before the hiding process can be mined from the released database. From the results we can see that the number of ghost rules is more in ISL approach and the number of lost rule is lesser in DSR approach. we concluded that there is no best solution for all problems. The choice of the algorithm to adopt depends on which criteria one considers as the most relevant: the time required or the information loss or the information that is added.

7. CONCLUSION AND FUTURE DIRECTION

In this paper, a classification of privacy preserving techniques is presented and major algorithms in each class is surveyed. The merits and demerits of different techniques were pointed out. The optimal sanitization is proved to be NP- Hard and always there is a tradeoff between privacy and accuracy. All the proposed methods provides only approximate solution for the goal of privacy preservation. To address this, following issues should be studied.

The algorithms for hiding sensitive association rules like privacy preserving rule mining using genetic algorithm, Tabu search based algorithms are limited to binary data, which can be extended to quantitative data.

The rule hiding techniques based on fuzzy methods requires membership function to be specified by an expert. These algorithms either use ISL or DSR approaches to hide sensitive association rules. Hybrid technique can be applied with which side effects of rule hiding can be reduced. Metrics for measuring the side effects can also be developed.

Although the personalized generalization approaches are flexible, the definitions of sensitive attributes are the same as other approaches. Thus, specifying sensitive information dynamically needs be future researched.

Rule interestingness measures in privacy preserving algorithms are generally limited to support and confidence. Depending on the nature of application, different measures can be used to measure the interestingness of quantitative rules. Semantic relation between attributes can be exploited in order to hide sensitive association rules with less side effects.

As each user may have different concern over privacy, user - oriented privacy preserving techniques can be developed.

Parallel algorithms could be developed to prevent revealing of sensitive association between items and to improve the performance of the algorithm for large datasets.

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, In Proceedings 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] Elena Dasseni, Vassilios S. Verykios, Ahmed K.Elmagarmid, and Elisa Bertino, “Hiding Association Rules by using Confidence and Support,” In Proceedings of the 4th Information Hiding Workshop (2001), pp.369– 383.

[6] 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):434-447.

[7] E.T. Wang, G. Lee, “An efficient sanitization algorithm for balancing information privacy and knowledge discovery in association patterns mining,” Data Knowl. Engg. (2008), doi:10.1016/j.datak.2007.12.005.

[8] E.D. Pontikakis, A. Tsitsonis, and V.S. Verykios."An experimental study of distortion-based techniques for association rule hiding". In Proc. of the 18th Annual IFIP WG 11.3 Working Conf. on Data and Applications Security. 2004.

[9] X. Sun, and P.S. Yu, "A border-based approach for hiding sensitive frequent itemsets". In: Proc. of the 5th P IEEE Int’l Conf. on Data Mining (ICDM'05). IEEE Computer Society, 2005. 426-433.

[10] Yucel Saygin, Vassilios Verykios, and Chris Clifton, “Using unknowns to prevent discovery of association rules,” SIGMOD Record 30 (2001),no. 4,pp. 45–54.

[11] 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.
[12] Xiao X, Tao Y, "Personalized privacy preservation", Proceedings of ACM Conference on management of Data (SIGMOD). ACM Press, New York: 2006, pp.785–790.
[13] Liu Ming, Xiaojun Ye, "Personalized K-anonymity", Computer Engineering and Design, Jan.2008, pp.282–286.
[14] Rakesh Agrawal and Ramakrishnan Srikant, "Privacy-preserving data mining," In Proceedings of the ACM SIGMOD Conference on Management of Data (2000), pp.439–450.
[15] 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.
[16] Mielikainen, T. "On inverse frequent set mining". In: Proc. of the 3rd IEEE ICDM Workshop on Privacy Preserving Data Mining. IEEE Computer Society, 2003. 18-23.
[17] DuW., AtallahM.: SecureMulti-party Computation: A Review and Open Problems.CERIAS Tech. Report 2001-51, Purdue University, 2001.
[18] Chris Clifton, Murat Kantarcioglou, XiadongLin and Michaed Y.Zhu, "Tools for privacy preserving distributed data mining," SIGKDD Explorations 4, no. 2, 2002.
[19] Vaidya J., Clifton C.: Privacy-Preserving Decision Trees over vertically partitioned data. Lecture Notes in Computer Science, Vol. 3654, 2005.
[20] 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.
[21] A. Sanil, A. Karr, X. Lin, and J. Reiter, "Privacy preserving analysis of vertically partitioned data using secure matrix products," Journal of Official Statistics, 2007.
[22] Vaidya, J. & Clifton, C.W., "Privacy preserving association rule mining in vertically partitioned data," In Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining, Edmonton, Canada, July 2002.
[23] ZongBo Shang; Hamerlinck, J.D., "Secure Logistic Regression of Horizontally and Vertically Partitioned Distributed Databases," Data Mining Workshops, ICDM Workshops 2007. Seventh IEEE International Conference on 28-31 Oct. 2007, pp.723–728.
[24] M. Kantarcioglu, C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data," The ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'02). ACM SIGMOD'2002 [C]. Madison, Wisconsin, 2002, pp.24–31.
[25] David W. Cheung, Jiawei Han, Vincent T. Ng,Ada W. Fu, and Yongjian Fu, "A fast distributed algorithm for mining association rules," In Proceedings of the 1996 International Conference on Parallel and Distributed Information Systems (1996).
[26] Shaofei Wu and Hui Wang ,"Research On The PrivacyPreserving Algorithm Of Association Rule Mining InCentralized Database", IEEE International Symposiums on Information Processing, 2008.
[27] 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.
[28] Wang Yan, Le Jiajin and Huang Dongmei, " A Method for Privacy Preserving Mining of Association Rules Based on Web Usage Mining", International Conference on Web Information Systems and Mining,IEEE 2010, pp.33-37.
[29] Yongcheng Luo, Yan Zhao, Jiajin Le, "A Survey on the Privacy Preserving Algorithm of Association Rule Mining", isecs, vol.1, pp.241-245, 2009 .
[30] Mohammad Naderi Dehkordi, Kambiz Badie, Ahmad Khadem Zadeh, " A Novel Method for Privacy Preserving in Association Rule Mining Based on Genetic Algorithms", Journal of software, vol. 4, no. 6, August 2009
[31] S. Vijayarani, A. Tamilarasi, R. SeethaLakshmi, "Tabu Search based Association Rule Hiding", International Journal of Computer Applications 19(1):12-18, April 2011.
[32] T. Berberoglu and M. Kaya, "Hiding Fuzzy Association Rules in Quantitative Data",The 3rd InternationalConference on Grid and Pervasive Computing Workshops, 2008, pp. 387-392.
[33] Dr. Duraiswamy. K, Dr. Manjula. D, and Maheswari. N "A New Approach to Sensitive Rule Hiding", ccsenet journal, vol 1, No. 3, August 2008 , 107-111.

[34] Yogendra Kumar Jain, Vinod Kumar Yadav, Geetika S. Panday," An Efficient Association Rule Hiding Algorithm for Privacy Preserving Data Mining", International Journal on Computer Science and Engineering (IJCSE), Vol. 3 No. 7 July 2011,pp. 2792 - 98.

[35] Muhammad Naeem, Sohail Asghar, Simon Fong, "Hiding Sensitive Association Rules Using Central Tendency",Advanced Information Management and Service(IMS), 2010 6th International conference on, On page(s):478 - 484, Nov.30 2010 -Dec. 2 2010.

[36] Ramesh Chandra Belwal, Jitendra Varshney, Sohel Ahmed Khan, Anand Sharma, Mahua Bhattacharya, "Hiding Sensitive Association Rules Efficiently By Introducing New Variable Hiding counter", IEEE International conference on Service Operations, Logistics and informatics, Vol.1,Oct. 2008, pp 130-134.

[37] Assaf Schuster, Ran Wolff, Bobi Gilburd," Privacy-Preserving Association Rule Mining in Large-Scale Distributed Systems", fourth IEEE symposium on Cluster Computing and Grid, 2004.

[38] Tirumala prasad B, Dr. MHM Krishna Prasad, "Distributed Count Association Rule Mining Algorithm", International Journal of Computer Trends and Technology, July to Aug Issue 2011, pp.280-284.

[39] 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.

[40] Vaidya J., Clifton C. "Privacy-Preserving k-means clustering over vertically partitioned Data". ACM KDD Conference, 2003.

[41] Vaidya J., Clifton C. "Privacy-Preserving Naive Bayes Classifier over vertically partitioned data". SIAM Conference, 2004.

[42] Yu H., Vaidya J., Jiang X. "Privacy-Preserving SVM Classification on Vertically Partitioned Data". PAKDD Conference, 2006.

[43] http://mlearn.ics.uci.edu/databases/breast-cancerwisconsin/breast-cancer-wisconsin.data