ISSN:2320-0790

An International Journal of Advanced Computer Technology

Literature Survey of Association Rule Based Techniques for Preserving Privacy

[1]Vijay Kumar Patidar, [2]Abhishek Raghuvanshi, [3]Vivek Shrivastava

[1]ITM, Bhilwara, [2]M.Tech (MIT, Ujjain), [3]M.Tech (ITM, Bhilwara)

Abstract: The paper gives an overview of privacy preserving in association rule mining techniques. In this paper, all the present privacy preserving using association rule hiding techniques are discussed. This paper also proposes a classification hierarchy that sets the basis for analyzing the work which has been performed in this context. A detailed review of the work accomplished in this area is also given, along with the coordinates of each work to the classification hierarchy.

1. Introduction

Data mining and knowledge discovery in databases are two new research areas that investigate the automatic extraction of previously unknown patterns from large amounts of data. The problem of privacy preserving data mining has become more important in recent years because of the increasing ability to store personal data about users and the increasing sophistication of data mining algorithm to leverage this information. A number of techniques such as classification, k-anonymity, association rule mining, clustering have been suggested in recent years in order to perform privacy preserving data mining. Furthermore, the problem has been discussed in multiple communities such as the database community, the statistical disclosure control community and the cryptography community. Data mining techniques have been developed successfully to extracts knowledge in order to support a variety of domains marketing, weather forecasting, medical diagnosis, and national security. But it is still a challenge to mine certain kinds of data without violating the data owners 'privacy .For example, how to mine patients 'private data is an ongoing problem in health care applications .As data mining become more pervasive, privacy concerns are increasing. Commercial concerns are also concerned with the privacy issue. Most organizations collect information about individuals for their own specific needs. Very frequently, however, different units within an organization themselves may find it necessary to share information. In such cases, each organization or unit must be sure that the privacy of the individual is not violated or that sensitive business information is not revealed .Consider, for example, a government, or more appropriately, one of its security branches interested in developing a system for determining, from passengers whose baggage has been checked, those who must be subjected to additional security measures. The data indicating the necessity for further examination derives from a wide variety of sources such as police records; airports; banks; general government statistics; and passenger information records that generally include personal information (such as name and passport number); demographic data (such as age and gender); flight information (such as departure, destination, and duration); and expenditure data (such as transfers, purchasing and bank transactions). In most countries, this information is regarded as private and to avoid intentionally or unintentionally exposing confidential information about an individual, it is against the law to make such information freely available.

While various means of preserving individual information have been developed, there are ways for circumventing these methods. In our example, in order to preserve privacy, passenger information records can be de-identified before the records are shared with anyone who is not permitted directly to access the relevant data. This can be accomplished by deleting from the dataset unique identity fields, such as name and passport number. However, even if this information is deleted, there are still other kinds of information, personal or behavioral (e.g. date of birth, zip code, gender, number of children, number of calls, number of accounts) that, when linked with other available datasets, could potentially identify subjects. To avoid these types of violations, we need various data mining algorithm for privacy preserving.

2. Classification of Techniques for Protecting Sensitive Data

There are many approaches which have been adopted for privacy preserving data mining. We can classify them based on the following dimensions:

- data distribution
- data modification
- data mining algorithm
- data or rule hiding
- privacy preservation

The first dimension refers to the distribution of data. Some of the approaches have been developed for centralized data, while others refer to a distributed data scenario. Distributed data scenarios can also be classified as horizontal data distribution and vertical data distribution. Horizontal distribution refers to these cases where different database records reside in different places, while vertical data distribution, refers to the cases where all the values for different attributes reside in different places.
The second dimension refers to the data modification scheme. In general, data modification is used in order to modify the original values of a database that needs to be released to the public and in this way to ensure high privacy protection. It is important that a data modification technique should be in concert with the privacy policy adopted by an organization. Methods of 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),
- Blocking, which is the replacement of an existing attribute value with a "?",
- Aggregation or merging which is the combination of several values into a coarser category,
- Swapping that refers to interchanging values of individual records, and
- Sampling, this refers to releasing data for only a sample of a population.

The third dimension refers to the data mining algorithm, for which the data modification is taking place. This is actually something that is not known beforehand, but it facilitates the analysis and design of the data hiding algorithm. For the time being, various data mining algorithms have been considered in isolation of each other. Among them, the most important ideas have been developed for classification data mining algorithms, like decision tree inducers, association rule mining algorithms, clustering algorithms, rough sets and Bayesian networks.
The fourth dimension refers to whether raw data or aggregated data should be hidden. The complexity for hiding aggregated data in the form of rules is of course higher, and for this reason, mostly heuristics have been developed. The lessening of the amount of public information causes the data miner to produce weaker inference rules that will not allow the inference of confidential values. This process is also known as "rule confusion".
The last dimension which is the most important refers to the privacy preservation technique used for the selective modification of the data. Selective modification is required in order to achieve higher utility for the modified data given that the privacy is not jeopardized. The techniques that have been applied for this reason are:

- Heuristic-based techniques like adaptive modification that modifies only selected values that minimize the utility loss rather than all available values
- Cryptography-based techniques like secure multiparty computation where a computation is secure if at the end of the computation, no party knows anything except its own input and the results, and
- Reconstruction-based techniques where the original distribution of the data is reconstructed from the randomized data.

It is important to realize that data modification results in degradation of the database performance. In order to quantify the degradation of the data, we mainly use two metrics. The first one, measures the confidential data protection, while the second measures the loss of functionality.

3. Review of Techniques for Protecting Sensitive Data

3.1 Heuristic-Based Techniques

A number of techniques have been developed for a number of data mining techniques like classification, association rule discovery and clustering, based on the premise that selective data modification or sanitization is an NP-Hard problem, and for this reason, heuristics can be used to address the complexity issues.

3.1.1 Centralized Data Perturbation-Based Association Rule Confusion

A formal proof that the optimal sanitization is an NP-Hard problem for the hiding of sensitive large item sets in the context of association rules discovery, have been given in [4]. The specific problem which was addressed in this work is the following one. Let D be the source database, R be a set of significant association rules that can be mined from D, and let R_h be a set of rules in R. How can we 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 R_h. The heuristic proposed for the modification of the data was based on data perturbation, and in particular the procedure was to change a selected set of 1-values to 0-values, so that the support of sensitive rules is lowered in such a way that the utility of the released database is kept to some maximum value. 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 subsequent work described in [10] extends the sanitization of sensitive large itemsets to the sanitization of sensitive rules. The approaches adopted in this work was either to prevent the sensitive rules from being generated by hiding the frequent itemsets from which they are derived, or to reduce the confidence of the sensitive rules by bringing it below a user-specified threshold. These two approaches led to the generation of three strategies for hiding sensitive rules. The important things to mention regarding these three strategies were the possibility for both a 1-value in the binary database to turn into a 0-value and a 0-value to turn into a 1-value. This flexibility in data modification had the side-effect that apart from non-sensitive association rules that were becoming hidden; a non-frequent rule could become a frequent one. We refer to these rules as "ghost rules". Given that sensitive rules are hidden, both non-sensitive rules which were hidden and non-frequent rules that became frequent (ghost rules) count towards the reduced utility of the released database. For this reason, the heuristics used for this later work, must be more sensitive to the utility issues, given that the security is not compromised. A complete work which was based on this idea, can be found in [24].

The work in [19] builds on top of the work previously presented, and aims at balancing between privacy and disclosure of information by trying to minimize the impact on sanitized transactions or else to minimize the accidentally hidden and ghost rules.

3.1.2 Centralized Data Blocking-Based Association Rule Confusion

One of the data modification approaches which have been used for association rule confusion is data blocking [6]. the approach of blocking is implemented by replacing certain attributes of some data items with a question mark. It is sometimes more desirable for specific applications (i.e., medical applications) to replace a real value by an unknown value instead of placing a false value. An approach which applies blocking to the association rule confusion has been presented in [22]. The introduction of this new special value in the dataset, imposes some changes on the definition of the support and confidence of an association rule. 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 below the middle in these two ranges of values, then we expect that the confidentiality of data is not violated. Notice that for an algorithm used for rule confusion in such a case, both 1-values and 0-values should be mapped to question marks in an interleaved fashion, otherwise, the origin of the question marks, will be obvious. An extension of this work with a detailed discussion on how effective is this approach on reconstructing the confused rules, can be found in [21].

3.2 Cryptography-Based Techniques

A number of cryptography-based approaches have been developed in the context of privacy preserving data mining algorithms, to solve problems of the following nature. Two or more parties want to conduct a computation based on their private inputs, but neither party is willing to disclose its own output to anybody else. The issue here is how to conduct such a computation while preserving the privacy of the inputs. This problem is referred to as the Secure Multiparty Computation (SMC) problem. In particular, an SMS problem deals with computing a probabilistic function on any input, in a distributed network where each participant holds one of the inputs, ensuring independence of the inputs, correctness of the computation, and that no more information is revealed to a participant in the computation than that's participant's input and output.

Two of the papers falling into this area, are rather general in nature and we describe them first. The first one [11] proposes a transformation framework that allows to systematically transform normal computations to secure multiparty computations. Among other information items, a discussion on transformation of various data mining problems to a secure multiparty computation is demonstrated. The data mining applications which are described in this domain include data classification, data clustering, association rule mining, data generalization, data summarization and data characterization. The second paper [8] presents four secure multiparty computation based methods that can support privacy preserving data mining. The methods described include, the secure sum, the secure set union, the secure size of set intersection, and the scalar product. Secure sum, is often given as a simple example of secure multiparty computation, and we present it here as well, as a representative for the techniques used. Below we present the approaches which have been developed by using the solution framework of secure multiparty computation. It should be made clear, that because of the nature of this

solution methodology, the data in all of the cases that this solution is adopted is distributed among two or more sites.

3.2.1 Vertically Partitioned Distributed Data Secure Association Rule Mining

Mining private association rules from vertically partitioned data, where the items are distributed and each itemset is split between sites, can be done by finding the support count of an itemset. If the support count of such an itemset can be securely computed, then we can check if the support is greater than the threshold, and decide whether the itemset is frequent. The key element for computing the support count of an itemset is to compute the scalar product of the vectors representing the sub-itemsets in the parties. Thus, if the scalar product can be securely computed, the support count can also be computed. The algorithm that computes the scalar product, as an algebraic solution that hides true values by placing them in equations masked with random values, is described in [23]. The security of the scalar product protocol is based on the inability of either side to solve *k* equations in more than *k* unknowns. Some of the unknowns are randomly chosen, and can safely be assumed as private. A similar approach has been proposed in [14]. Another way for computing the support count is by using the secure size of set intersection method described in [8].

3.2.2 Horizontally Partitioned Distributed Data Secure Association Rule Mining

In a horizontally distributed database, the transactions are distributed among *n* sites. The global support count of an itemset is the sum of all the local support counts. An itemset *X* is globally supported if the global support count of X is bigger than *s*% of the total transaction database size. A *k*-itemset is called a globally large *k*-itemset if it is globally supported. The work in [15] modifies the implementation of an algorithm proposed for distributed association rule mining [7] by using the secure union and the secure sum privacy preserving SMC operations.

3.2.3 Vertically Partitioned Distributed Data Secure Decision Tree Induction

The work described in [12] studies the building process of a decision tree classifier for a database that is vertically distributed. The protocol presented in this work, is built upon a secure scalar product protocol by using a third-party server.

3.2.4 Horizontally Partitioned Distributed Data Secure Decision Tree Induction

The work in [16] proposes a solution to the privacy preserving classification problem using a secure multiparty computation approach, the so-called oblivious transfer protocol for horizontally partitioned data. Given that a generic SMC solution is of no practical value, the authors focus on the problem of decision tree induction, and in particular the induction of ID3, a popular and widely-used algorithm for decision tree induction. The ID3 algorithm chooses the "best" predicting attribute by comparing entropies given as real numbers. Whenever the values for entropies of different attributes are close to each other, it is expected that the trees resulting from choosing either one of these attributes, have almost the same predicting capability. Formally stated, a pair of attributes has x-equivalent information gains if the difference in the information gains is smaller than the value x. This definition gives rise to an approximation of ID3. By denoting as *ID*3, the set of all possible trees which are generated by running the ID3 algorithm and choosing either attribute in the case that they are x-equivalent; the work in [16] proposes a protocol for secure computation of a specific ID3x algorithm. The
Protocol for privately computing ID3x is composed of many invocations of smaller private computations. The most difficult computations among these reduces to the oblivious evaluation of $x \ln x$ function.

3.3 Reconstruction-Based Techniques

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 mining. Below, we list and classify some of these techniques.

3.3.1 Reconstruction-Based Techniques for Numerical Data

The work presented in [3] addresses the problem of building a decision tree classifier from training data in which the values of individual records have been perturbed. While it is not possible to accurately estimate original values in individual data records, the authors propose a reconstruction procedure to accurately estimate the distribution of original data values. By using the reconstructed distributions, they are able to build classifiers whose accuracy is comparable to the accuracy of classifiers built with the original data. For the distortion of values, the authors have considered a discretization approach and a value distortion approach. For reconstructing the original distribution, they have considered a Bayesian approach and they proposed three algorithms for building accurate decision trees that rely on reconstructed distributions.

The work presented in [2] proposes an improvement over the Bayesian-based reconstruction procedure by using an Expectation Maximization (EM) algorithm
for distribution reconstruction. More specifically, the authors prove that the EM algorithm converges to the maximum likelihood estimate of the original distribution based on the perturbed data. They also show that when a

large amount of data is available, the EM algorithm provides robust estimates of the original distribution. It is also shown, that the privacy estimates of [3] had to be lowered when the additional knowledge that the miner obtains from the reconstructed aggregate distribution was included in the problem formulation.

3.3.2 Reconstruction-Based Techniques for Binary and Categorical Data

The work presented in [20] and [13] deal with binary and categorical data in the context of association rule mining. Both papers consider randomization techniques that offer privacy while they maintain high utility for the data set.

4. Evaluation of Techniques for Protecting Sensitive Data

An important aspect in the development and assessment of algorithms and tools, for privacy preserving data mining is the identification of suitable evaluation criteria and the development of related benchmarks. It is often the case that no privacy preserving algorithm exists that outperforms all the others on all possible criteria. Rather, an algorithm may perform better that another one on specific criteria, such as performance and/or data utility. It is thus important to provide users with a set of metrics which will enable them to select the most appropriate privacy preserving technique for the data at hand; with respect to some specific parameters they are interested in optimizing.

A preliminary list of evaluation parameters to be used for assessing the quality of privacy preserving data mining algorithms, is given below:

• The *performance* of the proposed algorithms in terms of time requirements, that is the time needed by each algorithm to hide a specified set of sensitive information;

• The *data utility* after the application of the privacy preserving technique, which is equivalent with the minimization of the information loss or else the loss in the functionality of the data;

• The *level of uncertainty* with which the sensitive information that have been hidden can still be predicted;

Below refer to each one of these evaluation parameters and analyze them.

5. Conclusions

We have presented a classification and an extended description and clustering of various privacy preserving data mining algorithms. The work presented in here, indicates the ever increasing interest of researchers in the area of securing sensitive data and knowledge from malicious users. The conclusions that we have reached from reviewing this area, manifest that privacy issues can be effectively considered only within the limits of certain data mining algorithms. The inability to generalize the results for classes of categories of data mining algorithms might be a tentative threat for disclosing information.

References

[1] Nabil Adam and John C. Wortmann, *Security- Control Methods for Statistical Databases: A Comparison Study*, ACM Computing Surveys **21** (1989), no. 4, 515–556.

[2] Dakshi Agrawal and Charu C. Aggarwal, *On the design and quantification of privacy preserving data mining algorithms*, In Proceedings of the 20th ACM Symposium on Principles of Database Systems (2001), 247–255.

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

[4] Mike J. Atallah, Elisa Bertino, Ahmed K. Elmagarmid, Mohamed Ibrahim, and Vassilios S. Verykios, *Disclosure Limitation of Sensitive Rules*, In Proceedings of the IEEE Knolwedge and Data Engineering Workshop (1999), 45–52.

[5] LiWu Chang and Ira S. Moskowitz, *Parsimonious downgrading and decision trees applied to the inference problem*, In Proceedings of the 1998 New Security Paradigms Workshop (1998), 82–89.

[6] LiWu Chang and Ira S. Moskowitz, *An integrated framework for database inference and privacy protection*, Data and Applications Security (2000), 161–172, Kluwer, IFIP WG 11.3, The Netherlands.

[7] 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).

[8] Chris Clifton, Murat Kantarcioglou, Xiadong Lin, and Michael Y. Zhu, *Tools for privacy preserving distributed data mining*, SIGKDDExplorations **4** (2002), no. 2.

[9] Chris Clifton and Donald Marks, *Security and privacy implications of data mining*, In Proceedings of the ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (1996), 15–19.

[10] 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), 369–383.

[11] Wenliang Du and Mikhail J. Attalah, *Secure multi-problem computation problems and their applications: A review and open problems*, Tech.

[12] Wenliang Du and Zhijun Zhan, *Building decision tree classifier on private data*, In Proceedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining (2002).

[13] Alexandre Ev.mievski, Ramakrishnan Srikant, Rakesh Agrawal, and Johannes Gehrke, *Privacy preserving mining of association rules*, In Proceedings of the 8th ACM SIGKDDD International Conference on Knowledge Discovery and Data Mining (2002).

[14] Ioannis Ioannidis, Ananth Grama, and Mikhail Atallah, *A secure protocol for computing dot products in clustered and distributed environments*, In Proceedings of the International Conference on Parallel Processing (2002).

[15] Murat Kantarcioglou and Chris Clifton, *Privacy-preserving distributed mining of association rules on horizontally partitioned data*, In Proceedings of the ACM SIGMOD Workshop on Research Isuues in Data Mining and Knowledge Discovery (2002), 24–31.

[16] Yehuda Lindell and Benny Pinkas, *Privacy preserving data mining*, In Advances in Cryptology - CRYPTO 2000 (2000), 36–54.

[17] Ira S. Moskowitz and LiWu Chang, *A decision theoretical based system for information downgrading*, In Proceedings of the 5th Joint Conference on Information Sciences (2000).

[18] Daniel E. O'Leary, *Knowledge Discovery as a Threat to Database Security*, In Proceedings of the 1st International Conference on Knowledge Discovery and Databases (1991), 107–516.

[19] 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), 43– 54.

[20] Shariq J. Rizvi and Jayant R. Haritsa, *Maintaing data privacy in association rule mining*, In Proceedings of the 28th International Conference on Very Large Databases (2002).

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

[22] 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.

[23] Jaideep Vaidya and Chris Clifton, *Privacy preserving association rule mining in vertically partitioned data*, In the 8th ACMSIGKDD International Conference on Knowledge Discovery and Data Mining (2002), 639–644.

[24] Vassilios S. Verykios, Ahmed K. Elmagarmid, Bertino Elisa, Yucel Saygin, and Dasseni Elena, *Association Rule Hiding*, IEEE Transactions on Knowledge and Data Engineering (2003), Accepted