International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

A Learning of PPDM Technique in Association Rule Mining

Priyanka Choubey[1], Savita Rathoad[2]

[1]M. Tech. (Final Year), Truba College of Engineering & Technology, Indore, Madhya Pradesh, India

[2]Professor, Truba College of Engineering & Technology, Indore, Madhya Pradesh, India

Abstract: ***This Paper is presents a novel classification that is based on "classification by the where". We consider our categorization is general, widespread and gives better understanding to the field of PPDM in terms of placing each problem in the right category. The new categorization is as follows: PPDM can be attempted at three levels. The first level is raw data or databases where transactions reside. The second level is data mining algorithms and techniques that ensure privacy. The third stage is the output of different data mining algorithms and techniques.***

Keywords: Privacy Preserving Data mining, Data Ware housing, data modification, data or rule hiding, Cryptography-Based Techniques

1. Introduction

Data mining and knowledge discovery in databases are two novel study areas that investigate the automatic mining of previously unknown patterns from large amounts of data. Recent advances in data gathering, data distribution and related technologies have inaugurate a new era of research where existing data mining algorithms should be reconsidered from a different point of observation, this of privacy preservation. It is well documented that this new without limits explosion of new information through the Internet and other media, has reached to a point where pressure against the privacy are very common on a daily basis and they deserve serious thinking.

Privacy preserving data mining [9, 18], is a new research direction in data mining and statistical databases [1], where data mining algorithms are analyzed for the side-effects they acquire in data privacy. The main reflection in privacy preserving data mining is dual. First, sensitive raw data like identifiers, names, addresses and the like, should be modified or trimmed out from the original database, in order for the recipient of the data not to be able to compromise another person's privacy. Next, sensitive knowledge which can be mined from a database by using data mining algorithms should also be barred, because such knowledge can equally well compromise data privacy, as we will indicate. The main aim in privacy preserving data mining is to develop algorithms for modify the original data in some way, so that the private data and private knowledge remain private even after the mining process. The difficulty that arises when confidential information can be derived from released data by not permitted users is also commonly called the "database inference" problem. In this report, I offer a categorization and an extended description of the various techniques and methodologies that have been developed in the area of privacy preserving data mining.

2. Classification Of Privacy Pre – Serving Techniques

There are many approaches which have been adopted for privacy preserving data mining. We can categorize 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. several of the approaches have been developed for centralized data, while others pass on 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, 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 central that a data modification technique should be in concert with the privacy policy adopted by an organization. Methods of alteration include:

a) 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),
b) Blocking, which is the replacement of an existing attribute value with a "?", The data mining algorithm, for which the data modification is taking place. This is actually a little that is not known ahead of time, but it facilitates the analysis and design of the data hiding algorithm. For the time being, different data mining algorithms have been measured in segregation of each other. Among them, the most significant ideas have been

developed for classification data mining algorithms, similar to decision tree inducers, association rule mining algorithms, clustering algorithms, forceful sets and Bayesian networks.

Whether raw data or aggregate data should be hidden. The complication 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 total 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 necessary in order to achieve higher utility for the modified data given that the privacy is not jeopardized. The techniques that have applied for this reason are:

- Heuristic-based techniques similar to 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 calculation, no party knows anything except its own input and the outcome, 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 show. In organize to quantify the degradation of the data, we principally use two metrics. The first one, actions the private data protection, while the second events the beating of functionality. From D, and let Rh be a set of rules in R. How can we change database D into a database D_, the on the loose database, so that all rules. 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. Review Of Privacy Reserving Algorithms

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 basis that selective

3.1.1 Centralized Data Perturbation-Based Association Rule Confusion

A subsequent work described in [10] extends the sanitization of sensitive large itemsets to the sanitization of sensitive rules. The approaches adopted in work was either to prevent the sensitive rules from being generated by hiding the frequent itemsets from which they are resulting, or to decrease the confidence of the sensitive rules by bringing it below a user-specified threshold. These two approaches led to the creation of three strategies for hiding sensitive rules. The important thing to mention regarding these three strategies was 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 change 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 secret and non-frequent rules that became frequent (ghost rules) count towards the reduced utility of the released database. For this rationale, the heuristics used for this later work, must be more sensitive to the utility issues, given that the security is not compromised. A total work which was based on this idea can be found in [24]. 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 usefulness 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 in R can still be mined from D_, apart from for the rules in Rh. The heuristic proposed for the modification of the data was based on data perturbation, and in particular the method was to change a selected.

3.1.2 Centralized Data Blocking-Based Association

a) 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 exact applications (i.e., medical applications) to replace a real value by an unknown value instead of placing a false value. An move toward 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 description of the support and confidence of an association rule. In this observe the minimum support and minimum confidence will be altered into a minimum support interval and a minimum confidence interval equally. As long as the support and/or the confidence of a sensitive rule lie below the middle in these two ranges of values, then we suppose 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 manner; otherwise, the origin of the question marks will be obvious. An extension of this effort with a detailed discussion on how effective is this approach on reconstructing the confused rules, can be found in [21].

b) 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 natural world. Two or more parties want to manner a computation based on their private inputs, but neither party is willing to disclose its own output to anybody else. The matter 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 among computing a probabilistic function on any input, in a distributed network where each participant holds one of the inputs, ensuring independence of the inputs, exactness of the computation, and that no more information is exposed to a participant in the computation than that's participant's input and output.

Two of the papers falling into this area, are rather common 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. Along with other information items, a discussion on transformation of various data mining problems to a secure multiparty computation is confirmed. The data mining applications which are described in this domain include data categorization, data clustering, association rule mining, data simplification, 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 take in, the secure sum, the secure set union, the secure size of set junction, and the scalar product. Secure sum, is often given as a easy example of secure multiparty computation, and we present it here as well, as an representative for the techniques used. under we present the approaches which have been developed by using the solution framework of secure multiparty calculation. It should be made plain, that because of the nature of this solution methodology, the data in all of the cases that this solution is adopted is dispersed among two or more sites.

3.2 Vertically Partitioned Distributed Data Secure Association Rule Mining

Mining private association rules from vertically partitioned data, where the items are dispersed and each itemset is split between sites, can be done by result the support count of an item set. If the support calculation of such an item set can be securely computed, then we can check if the support is greater than the threshold, and choose whether the item set is frequent. The key ingredient for computing the support count of an itemset is to compute the scalar product of the vectors representing the sub-itemsets in the party. Thus, if the scalar product be able to be firmly computed, the support count can also be computed. The algorithm that computes the scalar product, as an algebraic solution that hide 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. a little of the unknowns are randomly chosen, and can safely be assumed as confidential. 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].

3.2.1 Horizontally Partitioned Distributed Data Secure Association Rule Mining

In a horizontally distributed database, the dealings are distributed amid n sites. The global support count of an itemset is the sum of all the local support counts. An itemset X is worldwide 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.

3.2.2 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 creation protocol by using a third-party server.

3.2.3 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 price, 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 unlike attribute are close to each other, it is expected that the trees ensuing from choosing either one of these attributes, have almost the similar predict capability. right confirmed, a pair of attributes has x-equivalent information gains if the inequality in the information gain is smaller than the value x. This description gives rise to an approximation of ID3. By denote as ID3, the set of all potential trees which are generated by running the ID3 algorithm, and choosing either attribute in the case that they are x-equivalent, the occupation in [16] proposes a procedure for secure computation of a specific ID3x algorithm. The protocol for privately computing ID3x is composed of many invocations of lesser private computation. The most difficult computations among these reduces to the oblivious evaluation of x ln x purpose.

3.3 Reconstruction-Based Techniques

A number of recently proposed techniques address the issue of privacy preservation by disturbing the data and reconstructing the distributions at an aggregate level in order to perform the taking out. Below, we list and classify some of this technique.

3.3.1 Reconstruction-Based Techniques for arithmetic Data

The work presented in [3] addresses the problem of building a conclusion tree classifier from training data in which the values of individual records have been perturbed. While it is not potential to accurately estimate original values in individual data records, the author propose a reconstruction procedure to accurately estimate the distribution of creative data values. By using the reconstructed distributions, they are able to erect classifiers whose accuracy is comparable to the correctness of classifiers built with the original data. For the distortion of values, the authors have measured a

discretization approach and a value buckle approach. For reconstructing the original distribution, they have considered a Bayesian approach and they planned three algorithms for building accurate decision trees that rely on reconstruct distributions. The work presented in [2] proposes an improvement over the Bayesian-based reconstruction modus operandi by using an Expectation Maximization (EM) algorithm for distribution reconstruction. More specifically, the author proves that the EM algorithm converges to the maximum likelihood estimate of the original allocation based on the nervous data. They also show that when a large amount of data is accessible, 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 reconstruct aggregate allocation was included in the problem formulation.

3.3.2 Reconstruction-Based Techniques for Binary and definite Data

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

4. Evaluation of Privacy Preserving Algorithms

An important aspect in the advance and assessment of algorithms and tackle, for privacy preserving data mining is the identification of fitting evaluation criteria and the development of related benchmark. It is often the case that no isolation preserving algorithm exist that outperforms all the others on all possible criterion. Rather, an algorithm may perform better that a different one on specific criteria, such as performance and/or data efficacy. It is thus important to provide users with a set of metrics which will enable them to select the most apposite privacy preserving technique for the data at tender with respect to some detailed parameters they are interested in optimizing.

A preliminary list of evaluation parameter to be used for assessing the eminence of privacy preserving data mining algorithm, is given beneath:

- the performance of the proposed algorithms in terms of time requirements, that is the moment needed by each algorithm to hide a specified set of sensitive in rank;
- the data utility after the application of the retreat preserving technique, which is different 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 in turn that have been hidden can still be predicted; Below I refer to each one of these valuation parameters and I analyze them.

4.1 Performance of the proposed algorithms

A first approach in the assessment of the time requirements of a privacy preserving algorithm is to weigh up the computational cost. In this case, it is undemanding that an algorithm having a O(n2) polynomial complexity is more efficient than an extra one with O(en) exponential complexity.

An alternative draw near would be to evaluate the time requirements in terms of the average integer of operations, needed to reduce the frequency of manifestation of specific sensitive in sequence below a specified threshold. This values, perhaps, does not provide an absolute determine but it can be considered in order to perform a fast evaluation among different algorithms.

The communication cost incur during the exchange of information among a number of collaborating sites, should also be considered. It is imperative that this cost ought to be kept to a minimum for a distributed privacy preserve data mining algorithm.

4.2 Data Utility

The utility of the data, at the end of the isolation preserving process, is an important issue, because in order for sensitive information to be hidden, the database is essentially modified through the insertion of false in turn (swapping of values is a side effect in this case)or through the blocking of data values. We should observe here that some of privacy preserving techniques, like the use of case, do not modify the information stored in the database, but still, the effectiveness of the data falls, since the in turn is not complete in this case. It is noticeable that the more the changes are made to the database, the less the database reflects the sphere of interest. Therefore, an evaluation parameter for the data utility should be the amount of information that is lost after the use of privacy preserving process. Of course, the compute used to evaluate the information loss depends on the specific data taking out practice with respect to which a privacy algorithm is performed.

For illustration, information loss in the context of association rule withdrawal will be measured either in terms of the number of rules that were both lingering and lost in the database behind sanitization, or even in terms on the reduction/increase in the support and assurance of all the rules. For the casing of classification, we can use metrics similar to those used for organization rules. Finally, for clustering, the variance of the distances between the clustered items in the creative database and the sanitized database can be the basis for evaluating in rank loss in this case.

4.3 Uncertainty Level

The privacy preservation strategy operates by fall the information that we want to protect below certain thresholds. The hidden in turn, however, can still be inferred even though with some vagueness level. A sanitization algorithm then can be evaluated on the basis of the ambiguity that it introduces during the modernization of the hidden information. From an equipped point of view, a scenario would be to set a maximum to the perturbation of information, and then consider the degree of uncertainty achieved by each sanitization algorithm under this constraint. We expect that the algorithm that will attain the maximum uncertainty level, will be there the one which will be ideal over all the rest.

5. Conclusions

I have presented a classification and an extended portrayal and clustering of various privacy preserving data mining algorithms. The work on hand in here indicates the ever increasing interest of researchers in the area of secure sensitive data and knowledge from malicious users. The conclusions that we have reach from reviewing this area, manifest that privacy issues can be in effect considered only within the limits of certain data mining algorithms. The inability to generalize the results for classes of categories of data mining algorithms capacity 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).