ISSN:0975-9646

Supriya S. Borhade et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 3042-3045

# Privacy Preserving Data Mining- An Overview

Supriya S. Borhade[1], Dr. S. V. Gumaste[2]

[1]*M.E. Student, Computer Engineering Department,*
*SPCOE , Dumberwadi, Pune*
*Maharashtra, India*
[2] *Professor & Head, Computer Engineering Department,*
*SPCOE , Dumberwadi, Pune*
*Maharashtra, India*

***Abstract-*** **In recent years, ample amount of available personal data has made privacy preserving data mining issue an important one. An overview of new and quickly rising research field of privacy preserving data mining and a few exist problems provided in this paper. Also made a classification for the privacy preserving data mining and analyze some works in this field. But sometimes these patterns can reveal sensitive information about the data holder or individuals whose information are the subject of the patterns. The notion of privacy-preserving data mining is to identify and disallow such revelations as evident in the kinds of patterns learned using traditional data mining techniques Data distortion method for achieving privacy protection association rule mining and privacy protection data release were focused on discussion. Detailed evaluation criteria of privacy preserving algorithm was illustrated, which include algorithm performance, data utility, privacy security degree and data mining difficulty. Finally, the development of privacy preserving data mining for further directions is prospected.**

***Keywords-*****PPDM, randomization , k-anonymity, secure multipart computation**

## I. INTRODUCTION

With the advance of data storage capabilities of computer, a range of new data mining algorithms have been proposed. Growingly information can be obtained from most corporate. The conventional privacy protection methods can not do this well, facing vital need of privacy protection in data mining, since when they protect sensitive information, the knowledge in data is prevented against accessing to. Privacy protection data mining mostly considers two aspects. First, how to guarantee that the information such as PAN card number, mothers name, address etc does not revealed in the data application process. Sensitive original information data is whether altered or cut from original database. Purpose of doing this is to prevent individual privacy against undesirable data receive. The next is how to make more beneficial data application. The service data mine algorithm the sensitive knowledge which uncovers from the database possibly destroys the data privacy, therefore should remove the sensitive rule. Mining useful sensitive information using data mining technology from database may obliterate some data privacy, so sensitive rules must be eliminated. The important function of privacy protection data mining is revises original data by some way and develops equivalent data mining algorithm. At present, privacy preserving technology in database application mainly focuses on data mining and on data anonymity two domains. Current privacy protection mainly research direction shown in Table 1. Privacy protection study issue is decided by practical application of different privacy protection requirement. Common privacy preservation techniques are committed to data protection at a lower privacy level, which accomplish privacy preserving through introduction of statistical models and probability models. Privacy preserving in data mining is mainly applied to achieve privacy protection by different data characteristics in high-level data. Data release based privacy protection is to provide a common privacy protection method in many applications, thus making designed privacy algorithm is also versatile. The research of privacy protection methods are focused on data distortion [3], data released and data encryption and so on, such as privacy protection classification mining algorithm, distributed privacy preserving collaborative recommendation, privacy protection association rules mining, data release and so on. Many algorithms were developed based on encryption methods, such as association rules mined in vertically partitioned and horizontally partitioned data, classification mining, clustering mining and decision tree mining etc. Paper about data streams privacy protection is few. Aggarwal et al. concerned about data streams release of k-anonymity privacy protection [4]. This paper will give reviews privacy protection algorithm and challenges come up in privacy protection mining issues. The rest of this paper organized as follows. Research methods of privacy preserving data mining algorithms summarized in Section II. Privacy protection technologies summarized in Section III. Finally in Section IV ended with conclusion.

## II. PRIVACY PRESERVING DATA MINING ALGORITHMS MAIN RESEARCH METHODS

There are many techniques of data mining for privacy protection. In this paper privacy preserving classification techniques based on the following features, such as, data distribution data distortion, data mining algorithms, data or rules hiding and privacy protection.

A brief description of each is representing.

A. *Data distribution:* At present, some algorithms execute privacy protection data mining on a centralized data and some on distributed data. Distributed data consist of and vertical partitioned data. Different database records in different sites in horizontal partitioned data and in vertically partitioned data each database record attribute values in different sites.

B. *Data distortion:* This technique is to alter original data-base record before release, so as to achieve privacy protection purpose. Data distortion methods include perturbation, blocking, merging or aggregation, swapping and sampling. All these techniques are accomplished by the alteration of an attribute value or granularity transformation of an attribute value.

Table 1. Privacy Protection Research Direction

| Research Direction | Demonstration |
|---|---|
| General Privacy Preservation Technology | Perturbation, Randomization Swapping, Encryption |
| Data Mining Privacy Preservation Technology | Association Rule Mining Classification, Clustering |
| Privacy Protection Data Publishing Principle | K-Anonymity l-Diversity m-Invariance l-Closeness |

C. *Data mining algorithms:* Privacy preserving data mining algorithm include classification mining, clustering, association rule mining and Bayesian networks etc.

D. *Data or rules hidden:* This technique refers to hide original data or rules of original data. Due to rules hidden of original data it is very complex to reform again, some person proposed heuristic method to solve this issue.

E. *Privacy protection:* In order to protect privacy there need to modify data carefully for achieving a high data utility. Do this for some reasons as. [3] Modify data based on adaptive heuristics methods and only modify selected values of, but not all values, which make information loss of data is minimum. [4] encryption technologies, such as secure multiparty computation. If each site know only their input and input but nothing about others, the calculations are safe. [1] Data reconstruction method can reconstruct original data distribution from random data.

## III. PRIVACY PROTECTION TECHNOLOGIES

### A. *Data Distortion Techniques:*

In order to protect privacy in released database, people proposed a lot of effective data mining technology to hide sensitive information. The purpose of privacy protection is as follow.

(1) Hide sensitive information enclosed in the original data;
(2) Data between hidden and original might have the same characteristics.
(3) Obtain the same data accuracy as original data set.

Privacy protection data mining algorithms, such as classification, clustering, association rule discovery, require desired data to modify or sanitize and the choice of sanitize data is a NP hard problem. To deal with this complex problem, the techniques of distortion, such as random blocking, perturbation and condensation can be used.

#### *a) Association Rules Mining based on Perturbation:*

Statistical implication is used to judge rules emergence in data set and support and confidence as a metric. All association rules are greater than or equal to user defined support and confidence, but from point of view of user that some rules are sensitive, some are not. Association rules hiding technique is to use the following method to pure the original data set. All sensitive rules can only appear on original data mining, at the same time (or greater than) the confidence and support is not allowed to appear when the data set is sanitized. That non-sensitive rules can be mine out in the original data set can also be mine on the clean data set in the same support and confidence. Those sensitive rules cannot be mine out in the original data set cannot be mine out in the sanitization data set at the same support and confidence. The optimal sanitization is NP hard [8] for association rules mining to hide large item sets. Reference [7] proposed a major development to clean-sensitive set 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 approaches for hiding sensitive rules. The important things to state 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 alteration had the side-effect that apart from non-sensitive association rules that were becoming hidden and a non-frequent rule could become a frequent one.

#### *b) Mining Association Rules Using block:*

An alternative perturbation for association rules of data alteration method is the data block [5]. Blocking method replace a assets value of data items with mark of question, That using anonymous value instead of actual values rather than using false value instead of actual values is very popular in medicine. Reference [8] proposed a method of association rules mining using blocking, which appropriate changes the definition on the minimum support, replace with minimum support interval and minimum confidence and replace with confidence interval. Privacy is wont be violated as long as support of sensitive rules below the middle of support interval, or confidence of sensitive rules below the middle of confidence interval. Whether 1-value or 0-value should be mapped to a question mark, otherwise original value of question mark will be exposed. Reference [7] is a detailed explanation of the effectiveness of blocking method, this method of reconstruction of the text using the rules of disturbance.

Classification Rule Mining Based on block Reference [13] provides a new framework combining classification rule analysis and frugal decrease, that in this framework, data administrator has as a goal to block values for class label. By doing this, the information receiver, will be unable to build instructive models for the data that is not demoted. Frugal decrease is a framework for formalizing the phenomenon of trimming out information from a data set for downgrading information. In Frugal decrease a cost measure is assigned to the potential downgraded information that it is not sent to low. The main goal to be accomplished in this work is to find out whether the loss of functionality associated with not downgrading the data, is worth the extra confidentiality.

*B. Distributed Privacy Preserving Mining:*

In the privacy preserving data mining environment, a lot of encryption based approach introduced to solve the problem with the following features. Two or more parties mine their data on the basis cooperation, but none of them willing to reveal their data. This is a secure multiparty computation(SMC) problems under distributed environment, which focuses on how to convert various data mining methods to secure multiparty computation issues, such as data classification, association rules mining, data clustering, data generalization and data aggregation. Secure multiparty computation(SMC) methods described include, the secure set union, the secure sum, the secure size of set intersection and the scalar product.

Let us discuss the distributed association rule mining:

a) *Vertically partitioned data association rules mining:* Vertically partitioned data set different attributes for each item in different sites. Mining private association rules from vertically partitioned data by finding the support count of an itemset. If the support count of such an itemset can be securely computed, then one can check if the support is greater than the threshold and decide whether the itemset is frequent. Each party involved in the calculation by the sub-item set composed of a vector and calculate the number of an item set support is the key to computing vector dot product. Therefore, if the dot product can be secure computing, supports can also be calculated in security.

b) *Horizontally partitioned data association rules mining:* The transactions are distributed among n sites in a horizontally partitioned database. The total 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.

*C. Reconstructed Technology:*

Much privacy preserving data mining technology proposed recently use data perturbation or reconstruction in data convergence layer. Reference [11] studied to construct a decision tree classifier using the individual records value of perturbation as training data. Since original values of individual records can not estimate accurately, the author considers estimating original distribution accurately. In order to reconstruct the original distribution, Bayesian method is considered. Reference [9] improves the Bayesian reconstruction process by using EM algorithm in the distributed data. More precisely, the author prove that the EM algorithm dictates the maximum estimated fairly asthe original data on the distribution of disruption, but also proved that when large amounts of data can be obtained, EM algorithm can estimate the original distribution robust. Reference [11] also shows that when background was known by data miner through the reconstruction distribution that, the estimation of privacy will decrease.

*D. Anonymous Privacy Protection:*

Anonymous release chose to publish the raw data. In order to achieve privacy protection, sensitive data does not publish or release sensitive data with lower accuracy. The current study focused on data anonymity technical, namely, Make trade-offs between the privacy disclosure risks and data utility, which selective release of sensitive data and information that may be disclosed sensitive data, but to ensure that sensitive data and privacy disclosure risk within the tolerable range. Data anonymity focuses on two aspects: one of the principles is to design better anonymity methods, so that the data released following this principle can not only better protect privacy, but also has great practical utility. The other hand is to design more efficient anonymity algorithms for specific anonymous principle. With the research depth of anonymity, how to achieve practical application of anonymity data becomes the focus of research. Samarati and Sweeney proposed *k*-anonymity principle which requires that each record in the table released cannot distinguish from other *k*-1 records [13]. Call *k* records, cannot be distinguished, an equivalent class. Here cannot be distinguished in terms of non-sensitive attributes. In general, greater *k* values bring about better degree of privacy protection, but the information loss increase. Due to do not make any constraint for sensitive data, that is flaw of *k*-anonymity. An attacker can use protocol against attack and background knowledge attack to identify sensitive data or personal relationships [12] , which leading to privacy leaks. (α, *k*)-anonymity [14] make a improvement on this basis, which not only ensure that *k*-anonymity publishing is satisfied but also ensure that each records related any attribute value in each released equivalence class is not higher than the percentage of α. Generally, data publishing methods, such as *k*-anonymity, *l*-diversity, *t*-closeness [14] and other anonymous release, use generalization techniques, which reduce accuracy and data utility largely. In terms of data collection, if is closure risks of all sensitive data, in data set D released by data owners, are less than the threshold α, α € (0,1), called the disclosure risk of data set as α. Such as static data release *l* diversity [12] ensures that disclosure risk of published data sets is less than 1/*l* and dynamic data publishing principles m-invariance [17] ensure that the disclosure risk of published data sets is less than 1/*m*.

E. *Evaluation of Privacy Protection Algorithms:*

An important aspect on privacy preserving data mining algorithms and tools for developing and evaluating is to select the appropriate evaluation criteria, but the reality is not a privacy protection data mining algorithms under a variety of indicators to be better than other algorithms, in general, an algorithm may be practical in terms of performance or a little better than others. It is very important to provide users with a set of metrics to make possible them to choose the best appropriate algorithms for data privacy preserving. Next, make simple introduce for algorithm performance, data utility, privacy protection degree and the difficulty of data mining.

*a) Algorithm Performance:*

The algorithm with *O*(*n*2) complexity polynomial time is more efficiency than those with *O*(*en*) index of complexity. An additional approach would be to evaluate the time requirements in terms of the average number of operations, needed to reduce the frequency of specific sensitive information appearance below a particular threshold. This values, perhaps, does not provide an absolute measure, but it can be considered in order to perform a rapid comparison among different algorithms.

b) *Data Utility:*
It is a very important issue for utility of data privacy protection. In order to hide sensitive information, false information should insert the database, or block data values. Even though sample Techniques do not modify the information stored in the database, but that, while their information is incomplete, still reduces data utility. More changes to the database, less data utility of the database. So estimated parameters of data utility is data information loss applied privacy protection. Of course, the estimate of information loss related with the specific data mining algorithms.

c) *Degree of Privacy Protection:*
Privacy protection policy is to protect the information demote to a certain threshold, but hidden information can be derived out by some uncertainty. The uncertainty reconstructed by hidden information can estimate sanitization algorithm. A solution can set a maximum on perturbation information from execution point of view and then consider achieve the degree of uncertainty by constraints of different sanitization method. Hope that the algorithm can achieve the greatest uncertainty and better than all the other algorithms.

d) *Difficulty of Different Data Mining:*
In order to provide the full estimation on sanitization method, one need to measure difficulty of data mining algorithms which is different with sanitization method and this called parameter horizontal difficulty. This estimation of parameter need consider the classification of data mining which is very important on the test. Alternatively, one may need to develop a formal framework that upon testing of a sanitization algorithm against pre-selected data sets, one can transitively prove privacy guarantee for the whole class of sanitization algorithms.

## IV. CONCLUSION

Privacy protection technology as a on the rise educational research has a wide range of applications in many fields in recent years. This paper focuses on the review of privacy protection technologies involves in data mining. First introduced the study of privacy protection status and the main research method and then introduce privacy protection methods such as encryption, distortion, privacy and anonymity. For the three protections corresponding literature is illustrated. Because privacy protection technology involves the development of multi-disciplines, there are still many issues to be further study: Data mining and data stream mining concerning about privacy in data mining which is a promising direction. With the development of spatial and geographic data, new applications based on user mobility patterns of behavior will come into view. Another area of concern is the incremental privacy protection data release and face in this area is to redesign data mining algorithms to process data increment. Finally, in addition to the field-driven research, a framework for estimating and comparing a variety of privacy protection data mining algorithms should be design.

## REFERENCES

[1] C C Aggarwal, P S Yu, "On static and dynamic methods for condensation-based privacy-preserving data mining," ACM Trans Database Syst, vol. 33, no. 1, 2008,doi: 10.1145/1331904.1331906.

[2] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim and V. Verykios, "Disclosure Limitation of Sensitive Rules,"Proceedings of the IEEE Knowledge and Data Engineering Workshop, 1999, pp. 45-52.

[3] J Lin, Y Cheng, "Privacy preserving itemset mining through noisy items," Expert Systems with Applications, vol. 36, Mar. 2009, pp. 5711-5717, doi: 10.1016/j.eswa.2008.06.052.

[4] V.S. Verykios, E. Bertino, I.N. Fovino, L.P. Provenza, Y. Saygin and Y. Theodoridis, "State-of-the-art in privacy preserving data mining," ACM SIGMOD Record, vol. 33, no. 1, 2004, pp. 50-57, doi: 10.1145/974121.974131.

[5] Dasseni E, Verykios V S, Elmagarmid A K, et al. Hiding association rules by using confidence and support[J]. Lecture Notes in Computer Science. 2001, 2137: 369-383.

[6] E. Dasseni, V.S. Verykios, A.K. Elmagarmid and E. Bertino, "Hiding association rules by using confidence and support," Lecture Notes In Computer Science, vol. 2137, 2001, pp. 369-383

[7] B.J. Ramaiah, A.R.M. Reddy and M.K. Kumari, "Parallel privacy preserving association rule mining on pc clusters," 2009 IEEE International Advance Computing Conference, Inst. of Elec. 2009, pp. 1538-1542, doi: 10.1109/IADCC.2009.4809247.

[8] L. Chang and I. Moskowitz, "An Integrated Framework for Database Privacy Protection," Data and Application Security, Springer Boston, 2002, pp. 161-172.

[9] R. Agrawal and R. Srikant, "Privacy-preserving data mining," Proceedings of the 2000 ACM SIGMOD international conference on Management of data, ACM, 2000, pp. 439-450.

[10] Y. Saygm, V.S. Verykios and C. Clifton, "Using Unknowns to Prevent Discovery of Association Rules," SIGMOD Record, vol. 30, no. 4, 2001, pp. 45-54.

[11] L. Chang and I.S. Moskowitz, "Parsimonious Downgrading and Decision Trees Applied to the Inference Problem," Proceedings of the 1998 workshop on New security paradigms, ACM, 1998, pp. 82-89

[12] L. Sweeney, "k-anonymity: A model for protecting privacy," International Journal of Uncertainty Fuzziness and Knowledge Based Systems, vol. 10, no. 5, 2002, pp. 557-570, doi: 10.1142/S021848850 2001648.

[13] D. Agrawal and C.C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms, "Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, ACM New York, NY, USA,2001, pp. 247-255.

[14] R. Wong, J. Li, A. Fu and K. Wang, "(α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing," Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM,2006, pp. 754-759.

[15] A.Machanavajjhala, D. Kifer, J. Gehrke and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," ACM Transactions on Knowledge Discovery from Data, vol. 1, no. 1, 2007, doi: 10.1145/1217299.1217302.

[16] X. Xiao and Y. Tao, "M-invariance: towards privacy preserving re-publication of dynamic datasets," Proceedings of the2007 ACM SIGMOD international conference on Management of data, ACM, Year Published, pp. 689-700, doi: 10.1145/1247480.1247556.

[17] L. Ninghui, L. Tiancheng and S. Venkatasubramanian, "t-Closeness: Privacy beyond *k*-anonymity and *l*-diversity," *Proceedings of the 23rd International Conference on Data Engineering*, Inst. of Elec. and Elec. Eng. Computer Society, 2007, pp.106-115, doi: 10.1109/ ICDE.2007.367856.