Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Environmental Sciences 12 (2012) 1341 – 1347

Procedia Environmental Sciences

2011 International Conference on Environmental Science and Engineering (ICESE 2011)

An Overview of Privacy Preserving Data Mining

Xinjun Qi , Mingkui Zong

School of Technology,Harbin University,Harbin, 150086, China

sjqxj@hrbu.edu.cn,hljzmk@126.com

Abstract

In recent years, wide available personal data has made privacy preserving data mining issue an important one. An overview of new and rapidly emerging research field of privacy preserving data mining and some exist problems provided in this paper. We also make a classification for the privacy preserving data mining, and analyze some works in this field. 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 protection degree, and data mining difficulty. Finally, the development of privacy preserving data mining for further directions is prospected.

Keywords:privacy preserving; data mining; randomization; k-anonymity; secure multipart computation

1. Introduction

With the advance of data storage capabilities of computer, a variety of new data mining algorithms have been proposed. More and more information can be obtained from all social organization. The traditional privacy protection methods can not do this well, facing urgent 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 mainly considers two aspects. First, how to guarantee that the information such as ID card number, name, address etc does not revealed in the data application process. Sensitive original information data is whether revised or cut from original database. Purpose of doing this is to prevent individual privacy against adverse data receive. The next is how to make more advantageous data application. The service data excavation algorithm the sensitive knowledge which unearths 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 destroy some data

 doi:10.1016/j.proenv.2012.01.432

privacy, so sensitive rules must be eliminated. The essential purpose of privacy protection data mining is revises original data by some way, and develops corresponding data mining algorithm. At present, privacy preserving technology in database application mainly concentrates on data mining and on data anonymity two domains. Current privacy protection mainly research direction shown in Table 1.

Privacy protection research issue is decided by practical application of different privacy protection requirement. General privacy preservation methods are committed to data protection at a lower privacy level, which achieve 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 [1], data encryption, and data released and so on, such as privacy protection classification mining algorithm, privacy protection association rules mining, distributed privacy preserving collaborative recommendation, data release and so on. Many algorithms were developed based on encryption methods, such as association rules mined in horizontally partitioned and vertically partitioned data, clustering mining, classification 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 [2]. This paper reviews privacy protection algorithms and challenge arising in privacy protection mining issues.

The rest of this paper organized as follows. Research methods of privacy preserving data mining algorithms summarized in Section 2. Privacy protection technologies summarized in Section 3. In Section 4 we make conclusion.

2. Privacy Preserving Data Mining Algorithms main research methods

There are many methods of data mining for privacy protection, our privacy preserving classification methods based on the following aspects, such as data distribution, data distortion, data mining algorithms, data or rules hiding, and privacy protection. We do a brief description of each.

Data distribution: Currently, 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.

Data distortion: This method is to modify original data-base record before release, so as to achieve privacy protection purpose. Data distortion methods include perturbation, blocking, aggregation or merging, swapping and sampling. All this methods 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

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

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

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. (1) 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. (2) Encryption technologies, such as secure multiparty computation. If each site know only their input and input but nothing about others, the calculations are safe. (3) Data reconstruction method can reconstruct original data distribution from random data.

3. Privacy Protection Technologies

3.1. 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 contained in the original data; (2) data between hidden and original have the same characteristics (3) get the same data accuracy as original data set. Privacy protection data mining algorithms, such as classification, association rule discovery, clustering, need choose data to modify or purify, and the choice of purified data is a NP hard problem. To deal with this complex problem, the methods of distortion, such as random perturbation, blocking, and condensation, are used.

Association Rules Mining based on Perturbation

Statistical significance 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 purified.

That non-sensitive rules can be dug out in the original data set can also be dug on the clean data set in the same support and confidence.

That sensitive rules can not be dug out in the original data set can not be dug out in the purification data set at the same support and confidence.

The optimal purification is NP hard [7] for association rules mining to hide large item sets. Reference [8] proposed a major project to clean-sensitive set to the purification 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, and a non-frequent rule could become a frequent one.

Mining Association Rules Using block

Another perturbation for association rules of data modification method is the data block [6]. Blocking method replace a property value of data items with mark of question, That using unknown value instead of actual values rather than using false value instead of actual values is very popular in medicine. Reference [7] 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. We think that privacy is not 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 [8] is a detailed description 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 [12] provides a new framework combining classification rule analysis and parsimonious downgrading, 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 informative models for the data that is not downgraded. Parsimonious downgrading is a framework for formalizing the phenomenon of trimming out information from a data set for downgrading information. In parsimonious downgrading 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.

3.2. Distributed Privacy Preserving Mining

In the privacy preserving data mining environment, the people made a lot of encryption based approach 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, data clustering, association rules mining, data generalization, and data aggregation. Secure multiparty computation methods described include, the secure sum, the secure set union, the secure size of set intersection, and the scalar product. Let us discuss the distributed association rule mining.

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 we 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.

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.

3.3. Reconstructed Technology

Much privacy preserving data mining technology proposed recently use data perturbation or reconstruction in data convergence layer. Reference [10] 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 [11] 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 as the 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 [10] also shows that when background was known by data miner through the reconstruction distribution that, the estimation of privacy will decrease.

3.4. 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 can not distinguish from other *k*-1 records [12]. We call *k* records, can not be distinguished, an equivalent class. Here can not 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 [13] , 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 [15] and other anonymous release, use generalization techniques, which reduce accuracy and data utility largely. In terms of data collection, if disclosure risks of all sensitive data, in data set D released by data owners, are less than the threshold α, $\alpha \in [0,1]$, called the disclosure risk of data set as α. Such as static data release *l*-diversity [13] ensures that disclosure risk of published data sets is less than $1/l$, and dynamic data publishing principles m-invariance [16] ensure that the disclosure risk of published data sets is less than $1/m$.

3.5. 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 enable them to choose the best appropriate algorithms for data privacy preserving. Next, we make simple introduce for algorithm performance, data utility, privacy protection degree and the difficulty of data mining.

Algorithm Performance

We can see that the algorithm with $O(^{n2})$ complexity polynomial time is more efficiency than those with $O(e^n)$ index of complexity. An alternative 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 specified threshold. This values, perhaps, does not provide an absolute measure, but it can be considered in order to perform a fast comparison among different algorithms.

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. Although sample Techniques do not modify the information stored in the database, but that, since 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.

Degree of Privacy Protection

Privacy protection policy is to protect the information downgrade to a certain threshold, but hidden information can be derived out by some uncertainty. The uncertainty reconstructed by hidden information can evaluate sanitation 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 purification method. We hope that an algorithm can achieve the greatest uncertainty, and better than all the other algorithms.

Difficulty of Different Data Mining

In order to provide the full estimation on purification method, we need to measure difficulty of data mining algorithms which is different with purification 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, we may need to develop a formal framework that upon testing of a sanitization algorithm against pre-selected data sets, we can transitively prove privacy assurance for the whole class of sanitization algorithms.

4. Conclusion

Privacy protection technology as a growing academic 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 we introduce the study of privacy protection status and the main research method, and then introduce privacy protection methods such as distortion, encryption, 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: Mobile data mining and data stream mining concerning about privacy in data mining which is a promising direction. With the growth of spatial and geographic data, new applications based on user mobility patterns of behavior will emerge. Another area of concern is the incremental privacy protection data release, and challenge 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] 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.

[2] 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.

[3] 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.

[4] 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.

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

[8] 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.

[9] 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.

[10] 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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] X. Xiao, and Y. Tao, "M-invariance: towards privacy preserving re-publication of dynamic datasets," Proceedings of the 2007 ACM SIGMOD international conference on Management of data, ACM, Year Published, pp. 689-700, doi: 10.1145/1247480.1247556.