Research on Privacy Preserving Data Mining

Pingshui Wang[1, *], Tao Chen[1, 2] and Zecheng Wang[1]

Abstract: In recent years, with the explosive development in Internet, data storage and data processing technologies, privacy preservation has been one of the greater concerns in data mining. A number of methods and techniques have been developed for privacy preserving data mining. This paper provided a wide survey of different privacy preserving data mining algorithms and analyzed the representative techniques for privacy preservation. The existing problems and directions for future research are also discussed.

Keywords: Privacy preserving data mining, randomization, anonymization, secure multi-party computation.

1 Introduction

Data mining is the extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data [Han and Kamber (2012)]. Privacy preserving data mining is a novel research direction in data mining. In recent years, with the rapid development in Internet, data storage and data processing technologies, privacy preserving data mining has been drawn increasing attention. In order to make a publicly available system secure, we must ensure not only that private sensitive data have been trimmed out, but also to make sure that certain inference channels have been blocked as well. Under privacy constraints, the privacy preserving data mining problem was extensive researched. A number of effective methods for privacy preserving data mining have been proposed. But most of those methods might result in information loss and side-effects in some extent, such as data utility-reduced, data mining efficiency-downgraded, etc. That is, an essential problem under the context is trade-off between the data utility and the disclosure risk. This paper provided a survey of different privacy preserving data mining algorithms.

The rest of this paper is organized as follows. In Section 1, we will introduce the classification of privacy preserving techniques. In Section 2, we will analyze the method of randomization for privacy preserving on the original data. Anonymization method will be discussed in Section 3. Issues in distributed privacy preserving data mining will be discussed in Section 4. Section 5 contains the conclusions and discussions.

2 Classification of privacy preserving techniques

The topic of privacy preserving data mining has been studied extensively by the data mining community in recent years. A number of techniques for privacy preserving data

[1] Anhui University of Finance and Economics, Bengbu, 233030, China.

[2] University of Kansas, Lawrence, Kansas, 66045, USA.

[*] Corresponding Author: Pingshui Wang. Email: 120081049@aufe.edu.cn.

JIHPP. doi:10.32604/jihpp.2019.05943

mining have been proposed. Most techniques use some form of transformation on the original data in order to perform the privacy preservation [Verykios, Bertino, Fovino et al. (2004)]. We usually classify them into the following three categories:

2.1 The randomization method

Randomization method is a popular method in current privacy preserving data mining studies, which adds noise to the data in order to mask the values of the records. The noise added is sufficiently large so that the individual values of the records can no longer be recovered. However, the probability distribution of the aggregate data can be recovered and subsequently used for privacy-preservation purposes. In general, randomization method aims at finding an appropriate balance between privacy preservation and knowledge discovery. Representative randomization methods include random-noise-based perturbation and Randomized Response scheme.

2.2 The anonymization method

Anonymization method aims at making the individual record be indistinguishable among a group records by using techniques of generalization and suppression. Its representative approach is k-anonymity. The motivating factor behind the k-anonymity approach is that many attributes in the data can often be considered quasi-identifiers which can be used in conjunction with public records in order to uniquely identify the records. Many methods have been proposed, for example, k-anonymity, p-sensitive k-anonymity, (a, k)-anonymity, l-diversity, t-closeness, M-invariance etc.

2.3 The distributed privacy preserving data mining method

Distributed privacy preserving data mining method mainly resolve the problems that people jointly conduct mining tasks based on the private inputs they provide. These mining tasks could occur between mutual un-trusted parties, or even between competitors. So protect privacy becomes a primary concern in distributed data mining setting. There are two different distributed privacy preserving data mining approaches such as the method on horizontally partitioned data and that on vertically partitioned data.

3 Method of randomization for privacy preserving on original data

The randomization method provides an effective yet simple way of preventing the user from learning sensitive data, which can be easily implemented at data collection phase for privacy preserving data mining, because the noise added to a given record is independent of the behavior of other data records.

When the randomization method is carried out, the data collection process consists of two steps [Zhang (2006)]. The first step is for the data providers to randomize their data and transmit the randomized data to the data receiver. In the second step, the data receiver estimates the original distribution of the data by employing a distribution reconstruction algorithm.

Representative randomization methods include random-noise-based perturbation and randomized response scheme. Agrawal et al. proposed a scheme for privacy preserving data mining using random perturbation and discussed how the reconstructed distributions may be used for data mining [Agrawal and Srikant (2000)]. In their randomization scheme, a random number is added to the value of a sensitive attribute. For example, if x_i is the value of a sensitive attribute, $x_i + r_i$, rather than x_i, will appear in the database, where r_i is a random noise drawn from some distribution. It is shown that given the distribution of random noises, reconstructing the distribution of the original data is possible. Subsequently, Evmievski et al. proposed an approach to conduct privacy preserving association rule mining [Evfimievski, Srikant, Agrawal et al. (2004)]. Kargupta et al. [Kargupta, Datta, Wang et al. (2003)] proposed a random matrix-based spectral filtering technique to recover the original data from the perturbed data. Huang et al. [Huang, Du and Chen (2005)] further proposed two other data reconstruction methods: PCA-DR and MLE-DR. In addition, several distribution reconstruction algorithms have been proposed in correspondence to different randomization operators [Agrawal and Aggarwal (2001); Evfimievski, Srikant, Agrawal et al. (2002); Rizvi and Haritsa (2012); Wang, Wang, Guo et al. (2018); Wang, Xiong, Pei et al. (2018)]. The basic idea of most algorithms is to use Bayesian analysis to estimate the original data distribution based on the randomization operator and the randomized data. For example, the expectation maximization (EM) algorithm [Agrawal and Aggarwal (2001)] generates a reconstructed distribution that converges to the maximum likelihood estimate of the original distribution.

The Randomized Response (RR) was firstly proposed by Warner [Warner (1965)]. The RR scheme is a technique originally developed in the statistics community to collect sensitive information from individuals in such a way that survey interviewers and those who process the data do not know which of two alternative questions the respondent has answered. In data mining community, Rizvi et al. presented a MASK scheme to mine association rules with secrecy constraints [Rizvi and Haritsa (2012)]. Du et al. proposed an approach to conduct privacy preserving decision tree building [Du and Zhan (2013)]. Guo et al. [Guo, Guo and Wu (2017)] addressed the issue of providing accuracy in terms of various reconstructed measures in privacy preserving market basket data analysis. Hou et al. [Hou, Wei, Wang et al. (2018)] proposed a Privacy Preserving Medical Recommendation (PPMR) algorithm, which can protect patients' treatment information and demographic information during online recommendation process without compromising recommendation accuracy and efficiency in the context of neighborhood-based Collaborative filtering (CF) methods.

The randomization method is a simple technique which can be easily implemented at data collection time. It has been shown to be a useful technique for hiding individual data in privacy preserving data mining.

4 Method of anonymization

With the rapid growth in database, networking, and computing technologies, a large amount of personal data can be integrated and analyzed digitally, leading to an increased use of data mining tools to infer trends and patterns. This has raised universal concerns about protecting the privacy of individuals.

The data records are often made available by simply removing key identifiers such as the name and social-security numbers from individual records. However, the combinations of other record attributes (named as quasi-identifier) can be used to exactly identify individual records. For example, attributes such as age, zip-code and sex are available in public records such as census rolls. When these attributes are also available in a given data set, they can be used to infer the identity of the corresponding individual.

In order to protect privacy, Sweeney [Sweeney (2002)] proposed the k-anonymity model which achieves k-anonymity using generalization and suppression, so that, any individual is indistinguishable from at least k-1 other ones along the quasi-identifier attributes in the anonymized data set. Generalization involves replacing (or recoding) a value with a less specific but semantically consistent value. For example, the date of birth could be generalized to a range such as year of birth, so as to reduce the risk of identification. Suppression involves not releasing a value at all. It is clear that such methods reduce the risk of identification with the use of public records, while reducing the accuracy of applications on the transformed data.

In recent years, numerous algorithms have been proposed for implementing k-anonymity via generalization and suppression. Bayardo et al. [Bayardo and Agrawal (2005)] presented an optimal algorithm that starts from a fully generalized table and specialized the dataset in a minimal k-anonymous table. LeFevre et al. [Lefevre, Dewittd and Ramakrishnan (2005)] described an algorithm that uses a bottom-up technique and a priori computation. Fung et al. [Fung, Wang and Yu (2005)] presented a top-down heuristic to make a table to be released k-anonymous. As to the theoretical results, Sweeney [Sweeney (2002)] proved the optimal k-anonymity is NP-hard and describe approximation algorithms for optimal k-anonymity. However, Machanavajjhala et al. [Machanavajjhala, Gehrke and Kifer (2007)] pointed out that the user may guess the sensitive values with high confidence when the sensitive data is lack of diversity, and introduced the l-diversity method. Subsequently, several models such as p-sensitive k-anonymity [Truta and Vinay (2006)], (a, k)-anonymity [Wong, Li and Fu (2006)], t-closeness [Li and Li (2007)] and M-invariance [24] etc. were proposed in the literature in order to deal with the problem of k anonymity.

The k-anonymity methods mainly focus on a universal approach that exerts the same amount of preservation for all individuals, without catering for their concrete needs. The consequence may be offering insufficient protection to a subset of people, while applying excessive privacy control to another subset. Motivated by this, Xiao et al. [Xiao and Tao (2006)] presented a new generalization framework based on the concept of personalized anonymity. Their technique performs the minimum generalization for satisfying everybody's requirements, and thus, retains the largest amount of information from the original data.

K-anonymity data mining is however a recent research area and many issues are still to be investigated, such as, the combination of k-anonymity with other possible data mining techniques; the investigation of new approaches for detecting and blocking k-anonymity violations.

5 Method for distributed privacy preserving data mining

The growth of Internet has triggered tremendous opportunities for distributed data mining, where people jointly conducting mining tasks based on the private inputs they supply.

These mining tasks could occur between mutual un-trusted parties, or even between competitors. So protect privacy becomes a primary concern in distributed data mining setting. Distributed privacy preserving data mining algorithms require collaboration between parties to compute the results or share no-sensitive mining results, while provably leading to the disclosure of any sensitive information.

In general, distributed data involves two forms: horizontally partitioned and vertically partitioned. Horizontally partitioned data: each site has complete information on a distinct set of entities, and an integrated dataset consists of the union of these datasets. In contrast, vertically partitioned data has different types of information at each site; each has partial information on the same set of entities.

Most privacy preserving distributed data mining algorithms are developed to reveal nothing other than the final result. Kantarcio et al. [Kantarcioglu and Clifton (2004)] have studied the privacy-preserving association rule mining problem over horizontally partitioned data. Their methods incorporate cryptographic techniques to minimize the information shared, while adding little overhead to the mining task. Lindell et al. [Lindell and Pinkas (2000)] have researched how to privately generate ID3 decision trees on horizontally partitioned data. The problem of privately mining association rules on vertically partitioned data was addressed in Ioannidis et al. [Ioannidis, Grama and Atallah (2012); Vaidya and Clifton (2002)]. Vaidya et al. [Vaidya and Clifton (2002)] have first showed how secure association rule mining can be done for vertically partitioned data by extending the Apriori algorithm. Du et al. [Du and Zhan (2012)] have developed a solution for constructing ID3 on vertically partitioned data between two parties. Vaidya et al. [Vaidya and Clifton (2014)] have developed a Naive Bayes classifier for privacy preservation on vertically partitioned data. Vaidya et al. [Vaidya and Clifton (2013)] have proposed the first method for clustering over vertically partitioned data. All these methods are almost based on Secure Multiparty Computation (SMC) technology.

Secure multiparty computation originated with Yao's Millionaires' problem [Yao and Andrew (1986)]. The basic problem is that two millionaires would like to know who is richer, with neither revealing their net worth. Abstractly, the problem is to simply compare two numbers, each held by one party, without either party revealing its number to the other. The SMC literature defines two basic adversarial models:

- *Semi-Honest Model*: Semi-honest adversaries follow the protocol faithfully, but can try to infer the secret information of the other parties from the data they see during the execution of the protocol.
- *Malicious Model*: Malicious adversaries may do anything to infer secret information. They can abort the protocol at any time, send spurious messages, spoof messages, collude with other (malicious) parties, etc.

SMC technologies used in distributed privacy preserving data mining areas mainly consist of a set of secure sub-protocols, such as, secure sum, secure comparison, dot product protocol, secure intersection, and secure set union and so on. Most existing work on very efficient privacy preserving data mining only provides the protocols against semi-honest adversaries. An important area for future research is to develop efficient mining protocols that remain secure and private even if some of the parties involved behave maliciously.

6 Conclusions and future work

In this paper, we carried out a wide survey of the different approaches for privacy preserving data mining, and analyzed the major algorithms available for each method and pointed out the existing drawback. All the purposed methods are only approximate to our goal of privacy preservation. We need to further perfect those approaches or develop some efficient methods. For this, we recognize that the following problems should be concentrated on.

- Privacy and accuracy are a pair of contradiction; improving one usually incurs a cost in the other. How to apply various optimizations to achieve a trade-off should be deeply researched.
- Side-effects are unavoidable in data sanitization process. How to measure and reduce their negative impact on privacy preserving needs to be considered carefully. We also need to define some metrics for measuring.
- In distributed privacy preserving data mining areas, we should try to develop more efficient algorithms and look for a balance between disclosure cost, computation cost and communication cost.
- How to deploy privacy-preserving techniques into practical applications also needs to be further studied.

Acknowledgment: We thank the anonymous reviewers and editors for their very constructive comments. This work was supported by the National Social Science Foundation Project of China under Grant 16BTQ085.

References

Agrawal, D.; Aggarwal, C. C. (2001): On the design and quantification of privacy preserving data mining algorithms. *Proceeding of the 20th ACM SIGMOD-SIGACTSIGART Symposium on Principles of Database Systems*, pp. 247-255.

Agrawal, R.; Srikant, R. (2000): Privacy-preserving data mining. *ACM SIGMOD Record*, vol. 29, pp. 439-450.

Bayardo, R.; Agrawal, R. (2005): Data privacy through optimal k-anonymization. *Proceeding of the 21st International Conference on Data Engineering*, pp. 217-228.

Du, W.; Zhan, Z. (2013): Using randomized response techniques for privacy preserving data mining. *Proceeding of 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 505-510.

Du, W. L.; Zhan, Z. J. (2012): Building decision tree classifier on private data. *Proceeding of the IEEE International Conference on Data Mining Workshop on Privacy, Security, and Data Mining*, pp. 1-8.

Evfimievski, A.; Srikant, R.; Agrawal, R.; Gehrke, J. (2002): Privacy preserving mining of association rules. *Proceeding of the 8th ACM SIGKDD International Conference on Knowledge Discovery in Databases and Data Mining*, pp. 217-228.

Evfimievski, A.; Srikant, R.; Agrawal, R.; Gehrke, J. (2004): Privacy preserving mining of association rules. *Information System*, vol. 29, pp. 343-364.

Fung, B.; Wang, K.; Yu, P. (2005): Top-down Specialization for Information and Privacy Preservation. *Proceeding of the 21st IEEE International Conference on Data Engineering*, pp. 205-216.

Guo, L.; Guo, S.; Wu, X. (2017): Privacy preserving market basket data analysis. *Proceeding of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pp. 103-114.

Han, J.; Kamber, M. (2012): *Data Ming: Concepts and Techniques*. Beijing: China Machine Press.

Hou, M. G.; Wei, R.; Wang, T. G.; Cheng, Y.; Qian, B. Y. (2018): Reliable medical recommendation based on privacy-preserving collaborative filtering. *Computers, Materials & Continua*, vol. 56, no. 1, pp. 137-149.

Huang, Z.; Du, W.; Chen, B. (2005): Deriving private information from randomized data. *Proceeding of the ACM SIGMOD Conference on Management of Data*, pp. 37-48.

Ioannidis, I.; Grama, A.; Atallah, M. J. (2012): A secure protocol for computing dot-products in clustered and distributed environments. *Proceeding of the 31st International Conference on Parallel Processing*, pp. 379-384.

Kantarcioglu, M.; Clifton, C. (2004): Privacy-preserving distributed mining of association rules on horizontally partitioned data, *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, pp. 1026-1037.

Kargupta, H.; Datta, S.; Wang, Q.; Sivakumar, K. (2003): On the privacy preserving properties of random data perturbation techniques. *Proceeding of the 3rd International Conference on Data Mining*, pp. 99-106.

Lefevre, K.; Dewittd, J.; Ramakrishnan, R. (2005): Incognito: efficient full-domain k-anonymity. *Proceeding of the ACM SIGMOD International Conference on Management of Data*, pp. 49-60.

Li, N. H.; Li, T. C. (2007): t-Closeness: Privacy beyond k-anonymity and l-diversity. *Proceeding of the IEEE 23rd International Conference on Data Engineering*, pp. 106-115.

Li, Y.; Li, J. B.; Chen, J. W.; Lu, M. C.; Li, C. Y. (2018): Seed selection for data offloading based on social and interest graphs. *Computers, Materials & Continua*, vol. 57, no. 3, pp. 571-587.

Lindell, Y.; Pinkas, B. (2000): Privacy preserving data mining. *Proceeding of the Advances in Cryptology*, pp. 36-54.

Machanavajjhala, A.; Gehrke, J.; Kifer, D. (2007): l-Diversity: privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 3.

Rizvi, S. Haritsa, J. (2012): Maintaining data privacy in association rule mining. *Proceeding of the 28th International Conference on Very Large Data Bases*, pp. 682-693.

Rizvi, S. J.; Haritsa, J. R. (2012): Maintaining data privacy in association rule mining. *Proceeding of the 28th VLDB Conference*, pp. 1-12.

Sweeney, L. (2002): K-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557-570.

Sweeney, L. (2002): Achieving k-anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 571-588.

Truta, T.; Vinay, B. (2006): Privacy protection: p-sensitive k-anonymity property. *Proceeding of the 22nd International Conference on Data Engineering Workshops*, pp. 94-103.

Vaidya, J.; Clifton, C. (2014): Privacy preserving naive Bayes classifier for vertically partitioned data. *Proceeding of SIAM International Conference on Data Mining*, pp. 522-526.

Vaidya, J.; Clifton, C. (2002): Privacy preserving association rule mining in vertically partitioned data. *Proceeding of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 639-644.

Vaidya, J.; Clifton, C. (2002): Privacy preserving association rule mining in vertically partitioned data. *Proceeding of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 639-644.

Vaidya, J.; Clifton, C. (2013): Privacy-preserving k-means clustering over vertically partitioned data. *Proceeding of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 206-215.

Verykios, V.; Bertino, E.; Fovino, I.; Theodoridis, Y. (2004): State-of-the-art in privacy preserving data mining. *ACM SIGMOD Record*, vol. 33, pp. 50-57.

Wang, M. J.; Wang, J.; Guo, L. H.; Harn, L. (2018): Inverted XML access control model based on ontology semantic dependency. *Computers, Materials & Continua*, vol. 55, no. 3, pp. 465-482.

Wang, X.; Xiong, C.; Pei, Q. Q.; Qu, Y. Y. (2018): Expression preserved face privacy protection based on multi-mode discriminant analysis. *Computers, Materials & Continua*, vol. 57, no. 1, pp. 107-121.

Warner, S. (1965): Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63-69.

Wong, R. C.; Li, J. Y.; Fu, A. W. (2006): (a, k)-Anonymity: An enhanced k-anonymity model for privacy-preserving data publishing. *Proceeding of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 754-759.

Xiao, X. K.; Tao, Y. F. (2006): Personalized privacy preservation. *Proceeding of the ACM Conference on Management of Data*, pp. 229-240.

Xiao, X. K.; Tao, Y. F. (2007): M-invariance: Towards privacy preserving re-publication of dynamic datasets. *Proceeding of the ACM Conference on Management of Data*, pp. 689-700.

Yao, A. C. (1986): How to generate and exchange secrets. *Proceeding of the 27th IEEE Symposium on Foundations of Computer Science*, pp. 162-167.

Zhang, N. (2006): Privacy-preserving data mining. *Texas A&M University*, pp. 19-25.