ISSN 2277-3061

A review on Privacy Preservation and Collaborative Data Mining

Amit Kumar
M.Tech
(Computer Science & Engineering.)
TIT, Bhopal
amit.nnsharma@gmail.com

Abstract

Privacy preservation is major issue in current data transmission over internet and cloud network. For the integrity and security of data various methods are used such as cryptography, data transformation, Steganography, watermarking and many more method. In consequence of all these method some data mining technique is used. The data mining technique provide Varity of algorithm for privacy preservation. The collaborative data mining technique used different agent method for the integrity of security of data during transmission. Issues about privacy-preserving data mining have emerged globally, but still the main problem is that non- sensitive information or unclassified data, one is able to infer sensitive information that is not supposed to be disclosed. Data collection is a necessary step in data mining process. Due to privacy reasons, collecting data from different parries becomes difficult. In this paper presents the review of privacy persevering technique used data mining.

Key words

Privacy Preservation; Data Mining; Association Rule; SMC

Council for Innovative Research

Peer Review Research Publishing System

Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY

Vol. 14, No. 12

www.ijctonline.com, editorijctonline@gmail.com

INTRODUCTION

In current decade the integrity and security of data over internet is challenging task. For the integrity and security of data used various techniques such as cryptography, Steganography and watermarking. In the journey of integrity and security privacy and preservation technique are used in terms of mining operation in domain knowledge. The process of mining used various algorithm such as clustering technique, classification technique and rule mining. The rule mining technique is strong approach of data mining used for privacy preservation[1,3]. Data mining techniques have been developed successfully to extracts knowledge in order to support a variety of domain areas marketing, weather forecasting, medical diagnosis, and national security. But it is still a challenge to mine some kinds of data without violating the data owners 'privacy .For example, how to mine patients 'private data is an ongoing

problem in health care applications. As data mining become more pervasive, privacy concerns are increasing. Though many types of preserving individual information have been developed, there are ways for circumventing these methods. For example, in order to preserve privacy, passenger information records can be de-identified before the records are shared with anyone who is not permitted directly to access the relevant data[4,5]. This can be done by removing from the dataset unique identity fields, such as name and passport number. Even though if this information is deleted, there are still other forms of information both personal and behavioural (e.g. date of birth, zip code, gender, number of children, number of calls, number of accounts) that, when connected with other available datasets, could easily recognize subjects. To avoid these types of violations, we require various data mining algorithms for privacy preserving[7]. Data mining a non-trivial extraction of novel, implicit, and actionable knowledge from large data sets is an evolving technology which is a direct result of the increasing use of computer databases in order to store and retrieve information effectively .It is also known as Knowledge Discovery in Databases (KDD) and enables data exploration, data analysis, and data visualization of huge databases at a high level of abstraction, without a specific hypothesis in mind. The working of data mining is understood by using a method called modelling with it to make predictions [8]. Data mining techniques are results of long process of research and product development and include artificial neural networks, decision trees and genetic algorithms. This retrieval of data as and when needed contributes the technology of data mining. Data mining can be viewed as a result of the natural evolution of information technology. Section-I gives the introduction of the privacy preservation and data mining[9]. Section-II gives the information of related work in privacy preservation. In section III discuss the method of privacy preservation. In section IV discuss problem formulation finally, in section-V conclusion and future scope.

II RELATED WORK

In this section discuss the related work in the field of privacy preservation using data mining technique and some other technique. The data mining technique offers various algorithms for the process of privacy preservation. Association rule mining play an important role in privacy preservation. Here discuss some work along their authors.

S KumaraSwamy and Manjula S H etl. [1] A Key Distribution-Less Privacy Preserving Data Mining system in which the publication of local association rules generated by the parties is published. The association rules are securely combined to form the combined rule set using the commutative RSA algorithm. The combined rule sets established are used to classify or mine the data. The results discussed in this paper compare the accuracy of the rules generated using the C4.5 based KDLPPDM system and the C5.0 based KDLPPDM system using receiver operating characteristics curves (ROC).

Murat Kantarcioglu and Wei Jiang etl. [2] first develop key theorems, then base on these theorems, we analyze certain important privacy-preserving data analysis tasks that could be conducted in a way that telling the truth is the best choice for any participating party. Even though privacy-preserving data analysis techniques guarantee that nothing other than the final result is disclosed, whether or not participating parties provide truthful input data cannot be verified.

Tamir Tassa etl. [3] proposed a protocol for secure mining of association rules in horizontally distributed databases. Their protocol, like theirs, is based on the Fast Distributed Mining (FDM) algorithm, which is an unsecured distributed version of the Apriori algorithm. The main ingredients in our protocol are two novel secure multi-party algorithms one that computes the union of private subsets that each of the interacting players hold, and another that tests the inclusion of an element held by one player in a subset held by another.

S. Sasikala, and Nathira Banu etl. [4] They have considered, for the first time, the issue of providing efficiency in privacy preserving mining. Our goal was to investigate the possibility of simultaneously achieving high privacy, accuracy and efficiency in the mining process. they first showed how the distortion process required for ensuring privacy can have a marked negative side-effect of hugely increasing mining runtime. Then, we presented our new K-Means LBG algorithm that is specifically designed to minimize this side-effect through the application of symbol specific distortion.

Shipra Agrawal and Jayant R. Haritsa etl. [5] present FRAPP, a generalized matrix-theoretic frame-work of random perturbation, which facilitates a systematic approach to the design of perturbation mechanisms for privacy-preserving mining. Specifically, FRAPP is used to demonstrate that (a) the prior techniques differ only in their choices for the perturbation matrix elements, and (b) a symmetric positive definite perturbation matrix with minimal condition number can be identified, substantially enhancing the accuracy even under strict privacy requirements.

Somayyeh And Mohammed Reza etl. [6] presents a Framework for classification and evaluation of the privacy preserving data mining techniques for distributed data scenario. Based on our framework the techniques are divided into three . major groups, namely Secure Multiparty Computation based techniques, Secret Sharing based techniques and Perturbation based techniques. Also in proposed framework, seven functional criteria will be used to analyze and analogically

evaluation of the techniques in these three major groups. The proposed framework provides a good basis for more accurate comparison of the given techniques to privacy preserving distributed data mining.

III METHOD OF PRIVACY PRESERVATION

Process of review and study find data mining techniques indicate that these methods can be classified based on the conditions of privacy protection into three principle groups of Secure Multiparty Computation based techniques, Secret Sharing based techniques and Perturbation based techniques[10].

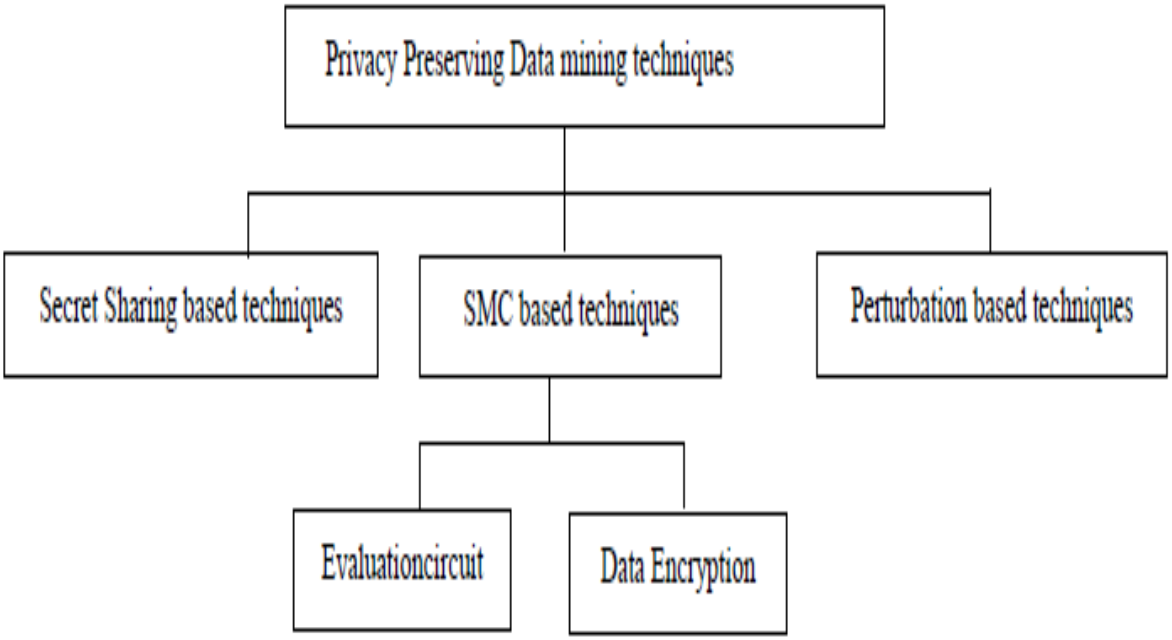

Figure 1: PPDM Techniques classification framework.

RANDOMIZATION METHOD

Randomization method is a popular method in current privacy preserving data mining studies. It masks the values of the records by adding noise to the original data. The noise added is sufficiently large so that the individual values of the records can no longer be recovered[11]. 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. The randomization method is more efficient. However, it results in high information loss.

THE ANONYMIZATION METHOD

Anonymization method aims at making the individual record be In distinguishable among a group records by using techniques of generalization and suppression[12]. The representative anonymization method 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 advanced methods have been proposed, such as, p-sensitive k-anonymity, (a, k)-anonymity, l-diversity, t-closeness, M-invariance, Personalized anonymity, and so on. The anonymization method can ensure that the transformed data is true, but it also results in information loss in some extent.

THE ENCRYPTION METHOD

Encryption method mainly resolves 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, therefore, protecting 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. The encryption method can ensure that the transformed data is exact and secure, but it is much low efficient[13]

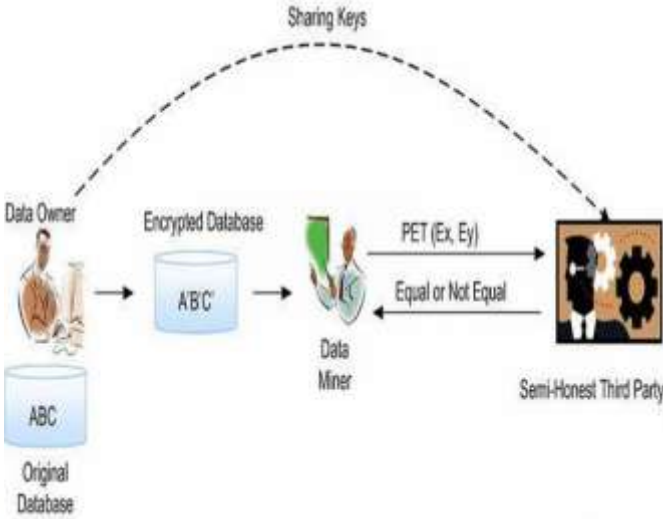

Figure 2: PPDM Techniques based on secret share generation technique.

IV PROBLEM FORMULATION

Privacy preserving play an important role in data hiding and data security. Now in conventional technique of data hiding required cryptography technique. But now a day's data mining important tools for data privacy preserving. Data mining is a set of automated techniques used to extract hidden or buried information from large databases. The term data mining refers to the nontrivial extraction of valid, implicit, potentially useful and ultimately understandable information in large databases with the help of the modern computing devices[14,15]. In the last few decades, many successful applications in data mining have been reported from varied sectors such as marketing, finance, medical diagnosis, banking, manufacturing and telecommunication. Apart from the benefits of using data per se the mining of these datasets with the existing data mining tools can reveal invaluable knowledge that was unknown to the data holder beforehand. The extracted knowledge patterns can provide insight to the data holders as well as be invaluable in tasks such as decision making and strategic business planning. As a valuable technique, data mining is developing and is flourishing. But, at the same time, serious concerns have grown over individual privacy in data collection, processing and mining. In concern of data mining application as privacy preserving various techniques are used such as association rule mining, clustering technique and classification technique. And also used some data mixed technique for adaptive noise data in original data. Matrix decomposition is big role in privacy preserving in data mining classification. The types of matrix decomposition are horizontal vertical and diagonal of index data of privacy. In data mining application the utility of third party has been removed. In the process of matrix decomposition singular and multiple values are involved. The singular value decomposition prevents the loss of mixed data and extracted data in decomposition of matrix.

V CONCLUSION AND FUTURE WORK

In this paper presents the review of privacy preservation technique based on data mining approach. The data mining approach provide the Varity of algorithm for the process of privacy preservation technique. In the process of review seen that the mining algorithm used some hybrid method and technique for the preservation of data. The transformation technique is very good approach for privacy preservation, but the problem are still remain in concern of complexity noise and loss of data. In future use single point vector decomposition method for the process of privacy preservation in data mining. The single point vector decomposition technique is very efficient process of technique. The decomposition technique used a single selection point for data and reduces the loss of data and increases the security strength of privacy preservation.

REFERENCES

[1] S KumaraSwamy,Manjula S H, K R Venugopal, Iyengar S S, L M Patnaik “Association Rule Sharing Model for Privacy Preservation and Collaborative Data Mining Efficiency” IEEE, Proceedings of 2014 RAECS UIET, 2014. Pp 12-18.

[2] Murat Kantarcioglu and Wei Jiang “Incentive Compatible Privacy-Preserving Data Analysis” IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Vol-25, 2013. Pp 1323-1335.

[3] Tamir Tassa “Secure Mining of Association Rules in Horizontally Distributed Databases” 2011. Pp 1-18.

[4] S. Sasikala, and Nathira Banu “Privacy Preserving Data Mining Using Piecewise Vector Quantization (PVQ)”,IJARCST 2014. Vol-2, Pp 302-306.

[5] Shipra Agrawal , Jayant R. Haritsa, B. Aditya Prakash “FRAPP: a framework for high-accuracy privacy-preserving mining” Springer, 2005. Pp 1-39.

[6] Somayyeh And Mohammed Reza “Classification And Evaluation The Privacy Preserving Distributed Data Mining techniques” Journal of Theoretical and Applied Information Technology, 2012. Pp 204-210.

[7] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding “ Data Mining with Big Data” 2011. Pp 1-26.

[8] Nirali R. Nanavati and Devesh C. Jinwala “Privacy Preserving Approaches for Global Cycle Detections for Cyclic Association Rules in Distributed Databases” IEEE, 2011. Pp 368-371.

[9] Ms.R.Kavitha, Prof.D.Vanathi “A Study Of Privacy Preserving Data Mining Techniques” International Journal of Science and Applied Information Technology, Vol-3, 2014. Pp 71-78.

[10] Kumaraswamy S, Manjula S H, K R Venugopal and L M Patnaik “ A Data Mining Perspective in Privacy Preserving Data Mining Systems” International Journal of Current Engineering and Technology, 2014. Pp 704-717.

[11] Julien Freudiger, Shantanu Rane, Alejandro E. Brito and Ersin Uzun “Privacy Preserving Data Quality Assessment for High-Fidelity Data Sharing” ACM, 2014. Pp 1-9.

[12] W. Jiang and B.K. Samanthula “A Secure and Distributed Framework to Identify and Share Needed Information” Proc. IEEE Int’l Conf. Privacy, Security, Risk and Trust, 2011.

[13] W. Jiang and B.K. Samanthula“N-Gram Based Secure Similar Document Detection” Proc. 25th Ann. WG 11.3 Conf. Data and Applications Security and Privacy, 2011.

[14] M. Kantarcioglu and O. Kardes “Privacy-Preserving Data Mining in the Malicious Model” Int’l J. Information and Computer Security, Vol-2, 2009, Pp353-375.

[15] M. Kantarcioglu and R. Nix “Incentive Compatible Distributed Data Mining” Proc. IEEE Int’l Conf. Soc. Computing/IEEE Int’l Conf. Privacy, Security, Risk and Trust, 2010, Pp 735-742.