Survey on Privacy Preserving Data Mining Techniques using Recent Algorithms

Rajesh N.
Research Scholar
School of Computer Science and Engg,
Bharathiyar University
Coimbatore, Tamilnadu, India.

Sujatha K.
Research Scholar
Computer Science and Engg,
Reva University
Bangalore,India

A. Arul Lawrence Selvakumar, PhD
Professor & Head
Department of Computer Science & engineering
Rajiv Gandhi Institute of Technology
Bangalore, India.

ABSTRACT

The privacy preserving data mining is playing crucial role act as rising technology to perform various data mining operations on private data and to pass on data in a secured way to protect sensitive data. Many types of technique such as randomization, secured sum algorithms and k-anonymity have been suggested in order to execute privacy preserving data mining. In this survey paper, on current researches made on privacy preserving data mining technique with fuzzy logic, neural network learning, secured sum and various encryption algorithm is presented. This will enable to grasp the various challenges faced in privacy preserving data mining and also help us to find best suitable technique for various data environment.

Keywords

Privacy Preserving Data Mining (PPDM), Privacy Preserving Data Publishing (PPDP), Secure Multiparty Computation (SMC), Cryptographic & Secured Sum Computation Algorithms.

1. INTRODUCTION

There is a tremendous increase in the research of data mining. Data mining is the process of extraction of data from large repositories. The most important extent in research community is Privacy preserving data mining (PPDM). It is very much necessary to maintain a ratio between privacy protection and knowledge discovery. The goal is to hide sensitive item sets so that the adviser cannot extract the modified database. To solve such problems there are various algorithms presented by various authors universal. The major goal of this survey paper is to understand the existing privacy preserving data mining methods and to achieve efficiency. In the recent years there is a great growth in the research of data mining and it's a process of pulling out of data from large set of database. One of the most crucial topics in research society is privacy preserving data mining (PPDM). It is essential to mange a data between privacy protection and knowledge discovery technique. The actual aim is hide sensitive data set from the authorized people extract and modification of dataset from large database. To solve a recent trend problem there are various techniques such as algorithm is going to be presented in this paper with its each technique drawback and beneficial. The primary aim of this paper is to grasp the existing privacy preserving data mining techniques to achieve efficiency. Data mining is the technique of analyzing the data set from different perspectives and get the useful information and specially to discover the knowledge is the ultimate aim of the data mining technique. In recent years the large information or day to day activities and carried out as well as is be transferred over the internet as well as any type of social media. Different privacy preservation deserves the huge attention over information and PPDM is a crucial technique in data mining where mining algorithms are incorporated. The importance of PPDM changes from different perspective because while express the data, the individual's identity and other crucial details should not get disclosed. Even though information will be lost due to privacy preservation surly damages the data utility. So PPDM, set of scales the substitution between data utility and privacy preservation by using various unique technique.

2. PRIVACY PRESERVING DATA MINING

Privacy preserving data mining is going to be achieved in different ways specifically by using randomization methods, cryptography algorithms and anonymization methods. A modern survey on are being used on various methods using privacy preserving data mining are found [1] in which reviews major PPDM techniques based on merits and demerits on recent trends in PPDM. The current scenario privacy preserving data mining [2] propose some future research directions for research people. In [15] all techniques of PPDM is studied and analyzed and from the analysis of [3] cryptography, Random data perturbation methods does better than other existing methods and specially cryptography is the best technique for encrypt the sensitive data of large data set.

3. PRIVACY PRESERVING DATA MINING (PPDM) METHODS

Many techniques have recently been proposed for privacy preserving data mining of multidimensional data set. Many privacy preserving data mining technologies are examined in [19] clearly and the benefits and drawbacks are analyzed such as k- anonymity, l-diversity, t-closeness, classification, association rule mining are proposed and designed to prevent identification to preserve the primary sensitive information and several application of several techniques for preserving privacy on testing dataset are expressed [22].

In recent situation many number of methods have been proposed for modifying or transformation of data to preserving privacy which are much needed and an effective but without compromising security to hide the sensitive data. This paper express a complete detailed survey on recent algorithms which are proposed for achieving a privacy preserving data mining using fuzzy logic, Neural networks, and other asymmetric encryption methods and also comparisons are made to know the best to do further research.

3.1 Anonymization Algorithms

Anonymization methods have an important tool to preserve privacy when releasing sensitive data set from larger volume of data. Most survey says common type of attack for Anonymization Algorithms is based on PPDM and PPDP is presented in [6] and their data privacy are explained and [7] novel technique is called slicing is proposed, which protect the data set.

3.2 Perturbation Algorithms

Perturbation based PPDM method expresses random perturbation to individual values to preserve privacy before data are published. In [8] the use of truncated non negative matrix factorization with sparseness rules of data perturbations are discussed [10]. The most possible of using multiplicative random projection matrices for privacy preserving distributed data mining for computing statistical aggregates like the inner product matrix, correlation coefficient matrix, and Euclidean distance matrix from distributed privacy sensitive data is explored. The scope of perturbation-based PPDM to Multilevel Trust (MLT-PPDM) is expanded in [9] which are robust against diversity attacks with respect to the privacy goal. In [12] a kind of privacy preserving classification mining method based on the random perturbation matrix is proposed which is suitable to the data of character type, Boolean type, classified type and digital type. It protects privacy adequately and has high accuracy in the mining results.

3.3 Cryptographic & Secured Sum Computation Algorithms

Modern privacy preserving collaborative rules is shown in [32] with light weight transparency which uses a new technique. A new approach encryption algorithm with an efficient approach is proposed in [20] an efficient conjunctive query scheme is being used to achieve the privacy preservation in [16]. Secure k-means data mining approach in the distributed environment is discussed in [17] by binding the merits of both RSA public key cryptosystem and homomorphism encryption technique. The purpose of security multiparty computation is to allow parties to carry out distributed computing tasks in secure way. In [13] a survey is made in the basis of paradigm and notations of secure shared calculation and evaluating the issue of efficiency and the problems involved in designing highly effective protocols. Different efficient basic secure building blocks are fast secure matrix multiplication (FSMP), Secure Scalar Product (SSP), and Secure Inverse of Matrix Sum (SIMS) is studied in [14]. Secure multi-party-data-ranking rules are proposed in [27] which are secure in the semi-honest design. A new approach [23] uses both actual and idyllic model to provide fair enough security and privacy. A procedure to compute the secured sum with zero leakage probability is provided in [21] and a procedure that is protected under the semi-honest adversarial model as well as stronger non-disruptive spiteful model is provided in [5].

3.4 Fuzzy based PPDM

A set of fuzzy-based mapping techniques is compared in [18] in terms of their privacy-preserving property and their ability to retain the same relationship with other fields. In [26], a method to extract global fuzzy rules from distributed data with the same attributes in a privacy-preserving manner is proposed. In [15], a fuzzy c-regression method is to generated synthetic data generation procedure which allows third parties to do statistical computations with a limited risk of disclosure. Fuzzy clustering approach can achieve data anonymization without significant loss of information because it effectively merges similar records into clusters where each record is not distinguishable from others after within-cluster merging. A study on intuitionistic fuzzy clustering is made in [28] and the applicability of fuzzy k-member clustering to privacy preserving pattern recognition is studied in [11]. K-member clustering is a basic technique for achieving k-anonymization, in which data samples are summarized so that any sample is indistinguishable from at least k - 1 other sample. A fuzzy variant of k-member clustering is proposed in [4] with the goal of improving the quality of data summarization with k-anonymity. This method is also applied to collaborative filtering. In [29] a secure framework for privacy preserving fuzzy co-clustering is proposed for handling both vertically and horizontally distributed co-occurrence matrices. A method to hide fuzzy association rule is proposed in [24] using modified apriori algorithm in order to identify sensitive rules to be hidden.

3.5 PPDM with Neural Network Learning (NNL)

To understand the design of Bayesian network on distributed various data set is addressed in the paper [31] and [25] clearly. A typical simple privacy preserving set of rules for understand the arguments of Bayesian network of vertically divided databases or data set with better performance, complete privacy and accuracy is presented with a clear picture[34]. A probabilistic neural network (PNN) board machine for Peer to Peer data mining is discussed in [12]. The l-diversity concepts are combined with k-anonymity is discussed in [30]. The background information cannot be exploited to effectively attack the privacy of information and the data disclosure probability and information loess are possibly kept minor [35]. From this paper gives clear picture about the secure privacy algorithm.

4. COMPARISON OF RECENT RESEARCHES ON PPDM

Table 1 shows the all available PPDM methods for data mining to secure the data set.

When we are transferring or exchanging the data set with fair enough security and also these methods ensures the various approaches [33] which are being used to obtain the cryptosystem.The following Table 1 shows the PPDM methods for supporting cryptography, Fuzzy logic and neural networks.

Table 1. Ppdm Methods

PPDM Methods	Techniques		
	Crypto graphy	**Fuzzy Logic**	**Neural Network**
Random perturbation	√	√	
k-anonymity	√	√	√
Horizontally partitioned Distribution	√	√	√
Vertically partitioned Distribution	√	√	√
Clustering	√	√	√
Classification	√	√	√
Association Rule Mining	√	√	√
Secured sum Computation	√	√	√
Aggregation	√	√	

5. CONCLUSIONS

In this paper, broad survey has been done on various privacy preserving data mining algorithms Fuzzy logic, Cryptography and Neural network learning techniques is made. Various merits of various algorithms shown in Table 1 explore to identify the algorithms which have good performance in terms of privacy and utility. This survey also helps researchers to understand the vital roles played by Fuzzy logic, neural network, Cryptography and secure sum computation methods in various PPDM methods and also to identify PPDM algorithms which are yet to be developed with better performance. It will lead to further researches to develop new and effective PPDM algorithms with high degree of privacy and lesser information loss. So from the above survey clearly express the problem of privacy preservation concern is security issue on data set over the media for sensitive typical data set.

6. REFERENCES

[1] A. Hussien, N. Hamza and H. Hefny, 2013 , Attacks on Anonymization-Based Privacy-Preserving: A Survey for Data Mining and Data Publishing, Journal of Information Security, Vol. 4 No. 2, 2013, pp. 101-112. doi:10.4236/jis.2013.42012.

[2] Cano I, Torra V. 2009 Generation of synthetic data by means of fuzzy c-Regression . IEEE International Conference on Fuzzy Systems, 2009. FUZZ-IEEE, pp: 1145 – 1150.

[3] Grljevic O, Bosnjak Z, Mekovec R. 2011, Privacy preserving in data mining - Experimental research on SMEs data, IEEE 9th International Symposium on Intelligent Systems and Informatics (SISY), 2011 , pp-477 – 481.

[4] Honda K, Kawano A, Nots A, Ichihashi H., 2012, A fuzzy variant of k-member clustering for collaborative filtering with data anonymization, Fuzzy Systems (FUZZ-IEEE), 2012 IEEE International Conference on, pp: 1-6.

[5] Hasan O, Bertino E, Brunie L. 2010, Efficient privacy preserving reputation protocols inspired by secure sum, Privacy Security and Trust (PST), 2010 Eighth Annual International Conference on, pp: 126 – 133.

[6] Inan A, Richardson TX, Kantarcioglu M, Bertino E., 2009, Using Anonymized Data for Classification, IEEE 25th International Conference on Data Engineering, 2009. ICDE '09. , pp : 429-430.

[7] Jian Wang, Yongcheng Luo ; Yan Zhao ; Jiajin Le, 2009, A Survey on Privacy Preserving Data Mining, First International Workshop on Database Technology and Applications, 2009 , pp: 111-114.

[8] Jiang, J. and Umano, M. 2014, Privacy preserving extraction of fuzzy rules from distributed data with different attributes, Joint 7th International Conference on and Advanced Intelligent Systems (ISIS), 15th International Symposium on Soft Computing and Intelligent Systems (SCIS), 2014, pp : 1180-1185.

[9] Kabir S.M.A, Youssef A.M, Elhakeem, A.K., 2007, On Data Distortion for Privacy Preserving Data Mining, Canadian Conference on Electrical and Computer Engineering, 2007. CCECE 2007. , pp : 308 – 311.

[10] Kun Liu , Kargupta H, Ryan J., 2006, Random projection-based multiplicative data perturbation for privacy preserving distributed data mining, Knowledge and Data Engineering, IEEE Transactions on (Volume:18 , Issue: 1), pp: 92 – 106.

[11] Kikuchi H., Aoki Y, Terada M, Ishii K., 2012, Accuracy of Privacy-Preserving Collaborative Filtering Based on Quasi-homomorphic Similarity, 9th International Conference on Ubiquitous Intelligence & Computing and 9th International Conference on Autonomic & Trusted Computing (UIC/ATC), 2012,pp : 555- 562.

[12] Kasugai H, Kawano A, Honda, K, Notsu A. 2013, A study on applicability of fuzzy k-member clustering to privacy preserving pattern recognition, IEEE International Conference on Fuzzy Systems (FUZZ), 2013, pp:1-6 .

[13] Kokkinos Y, Margaritis K. 2013, Distributed privacy-preserving P2P data mining via probabilistic neural network committee machines, Fourth International Conference on Information, Intelligence, Systems and Applications (IISA), 2013, pp: 1-4.

[14] Li Yaping Chen Minghua, Li Qiwei, Zhang, Wei. 2012, Enabling Multilevel Trust in Privacy Preserving Data Mining, Knowledge and Data Engineering, IEEE Transactionson (Volume:24, Issue: 9), pp: 1598 – 1612.

[15] Liu Wen, Luo Shou-shan, Wang Yong-bin, Jiang Zhen-tao, 2011, A Protocol of Secure Multi-party Multi-data Ranking and Its Application in Privacy Preserving Sequential Pattern Mining, 2011 Fourth International Joint Conference on Computational Sciences and Optimization (CSO), pp: 272 – 275.

[16] Malik M.B, Ghazi M.A, Ali R. 2012, Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects, Third International Conference on Computer and Communication Technology (ICCCT), pp: 26 – 32.

[17] Mi Wen, Rongxing Lu ; Jingshen Lei ; Xiaohui Liang , 2013, ECQ: An Efficient Conjunctive Query scheme over encrypted multidimensional data in smart grid, Global Communications Conference (GLOBECOM), 2013 IEEE, 796 – 801.

[18] Mittal D, Kaur D, Aggarwal A. 2014 , Secure Data Mining in Cloud Using Homomorphic Encryption IEEE International Conference on Cloud Computing in Emerging Markets (CCEM), 2014, pp : 1 – 7.

[19] Mukkamala R., Ashok V.G, 2011 Fuzzy-based Methods for Privacy-Preserving Data Mining Eighth International Conference on Information Technology: New Generations (ITNG), pp: 348 – 353.

[20] Inan A, Richardson TX, Kantarcioglu M., Bertino E., 2009, Using Anonymized Data for Classification, IEEE 25th International Conference on Data Engineering, 2009. ICDE '09. , pp : 429-430.

[21] Pathak F.A.N , Pandey S.B.S., 2013, An efficient method for privacy preserving data mining in secure multiparty computation, Nirma University International Conference on Engineering (NUiCONE), 2013, pp: 1 – 3 Pathak, F.A.N. , Pandey, S.B.S., 2013.

[22] Pathak F.A.N , Pandey S.B.S, 2013, Distributed changing neighbors k-secure sum protocol for secure multiparty computation, Nirma University International Conference on Engineering (NUiCONE), 2013, pp: 1 – 3.

[23] Shweta Taneja, Shashank Khanna, Sugandha Tilwalia, Ankita, 2014, A Review on Privacy Preserving Data Mining: Techniques and Research Challenges, International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, pp: 2310-2315.

[24] Shu Qin Ren , Khin Mi Mi Aung ; Jong Sou Park, 2010, A Privacy Enhanced Data Aggregation Model, Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on, pp: 985 – 990.

[25] SathiyaPriya K, Sadasivam G.S, Celin N. 2011, A new method for preserving privacy in quantitative association rules using DSR approach with automated generation of membership function, World Congress on Information and Communication Technologies (WICT), 2011, pp: 148-153.

[26] Samet S, Miri A., 2009, Privacy-Preserving Bayesian Network for Horizontally Partitioned Data International Conference on Computational Science and Engineering, 2009. CSE '09. (Volume:3), pp: 9-16.

[27] Tiancheng Li, Ninghui Li, Jian Zhang, Ian Molloy, "Slicing: A New Approach for Privacy Preserving Data Publishing", *IEEE Transactions on Knowledge & Data Engineering*, vol.24, no. 3, pp. 561-574, March 2012, doi:10.1109/TKDE.2010.236.

[28] Teo S.G ,Lee V, Shuguo Han, 2012, A Study of Efficiency and Accuracy of Secure Multiparty Protocol in Privacy-Preserving Data Mining, 26th International Conference on Advanced Information Networking and Applications Workshops (WAINA), pp: 85-90.

[29] Torra, V, Miyamoto S, Endo Y. Domingo-Ferrer, 2008, On intuitionistic fuzzy clustering for its application to privacy,J. FUZZ-IEEE (IEEE World Congress on Computational Intelligence). IEEE International Conference on Fuzzy System DOI: 10.1109/FUZZY.2010.5584186, pp 1042 –1048.

[30] Tanaka D, Oda T, Honda K, Notsu A, 2014, Privacy preserving fuzzy co-clustering with distributed cooccurrence matrices Joint 7th International Conference on Soft Computing and Intelligent Systems (SCIS), 2014 and 15th International Symposium on Advanced Intelligent Systems (ISIS), pp: 700-705.

[31] Tsiafoulis, S.G. Zorkadis, V.C., 2010, A Neural Network Clustering Based Algorithm for Privacy Preserving Data Mining, International Conference on Computational Intelligence and Security (CIS), 2010, pp: 401-405.

[32] Wang Hongmei , Zhao Zheng , Sun Zhiwei, 2005, Privacy-preserving Bayesian network structure learning on distributed heterogeneous data,.11th Pacific Rim International Symposium on Dependable Computing, 2005. Proceedings, DOI: 10.1109/PRDC.2005.49.

[33] Xiaolin Zhang, Hongjing Bi, 2010, Research on privacy preserving classification data mining based on random perturbation, Information Networking and Automation (ICINA), 2010 International Conference on (Volume:1), pp : V1-173 - V1-178

[34] Xueyun Li, Zheng Yan and Peng Zhang, 2014, A Review on Privacy-Preserving Data Mining, IEEE International Conference on Computer and Information Technology (CIT), 769 – 774.

[35] Zhiqiang Yang, Wright R.N. 2005, Improved Privacy-Preserving Bayesian Network Parameter Learning on Vertically Partitioned Data, 21st International Conference on Data Engineering Workshops, 2005. PP: 91-96.