ISSN:0975-9646

Survey on Recent Algorithms for Privacy Preserving Data mining

S.Selva Rathna[#1], Dr. T. Karthikeyan[*2]

[#]*Manonmaniam Sundaranar University*
Tirunelveli, Tamilnadu, India
[*]*P.S.G.Arts and Science College*
Coimbatore, Tamilnadu, India

Abstract— **Privacy preserving Data mining is an emerging technology which performs data mining operations in centralized or distributed data in a secured manner to preserve sensitive data. A number of techniques such as randomization, secured sum algorithm and k-anonymity have been suggested in order to perform privacy-preserving data mining. In this paper, a survey on recent researches made on Privacy preserving data mining techniques with Fuzzy logic, neural network learning, secured sum and encryption algorithms is presented. This will enable to understand the challenges faced in Privacy preserving data mining and also helps to identify best techniques suitable for various data environment.**

Keywords— **Multi level Trust Privacy Preserving Data mining (MLT-PPDM), Neural Network Learning (NNL), Non negative Matrix Factorization (NMF), Probabilistic Neural Network (PNN) Privacy Preserving Data Mining (PPDM), Privacy Preserving Data Publishing (PPDP), Secure Multiparty Computation (SMC)**

I. INTRODUCTION

In recent years, a number of techniques have been proposed for modifying or transforming data to preserve privacy which are effective without compromising security. This paper presents a detailed survey on recent algorithms developed for achieving Privacy preserving Data mining using Fuzzy logic, Neural network and other cryptographic methods. In Section II, recent researches related to PPDM are listed out. In Section III, recent researches on PPDM methods like anonymization, perturbation, Secured sum computation, encryption, fuzzy logic, neural network learning and other techniques are described. Section IV presents comparison of various techniques and algorithms developed recently and also shows their advantages.

II. PRIVACY PRESERVING DATA MINING

Privacy preserving data mining can be achieved in various ways by use of randomization techniques, cryptographic algorithms, ananymization methods etc. A recent survey on some of the techniques used for privacy-preserving data mining may be found in [1] which reviews main PPDM techniques based on a PPDM framework and compare the advantages and disadvantages of different PPDM technique and discuss the open issues and future research trends in PPDM. [2] describes the current scenario of Privacy preserving data mining and propose some future research directions. In [3] all state of art techniques of PPDM is studied. From the analysis of [3], Cryptography and Random Data Perturbation methods perform better than the other existing methods. Cryptography is best technique for encryption of sensitive data and Data Perturbation will help to maintain sensitivity of data.

III. PRIVACY PRESERVING DATA MINING (PPDM) METHODS

A number of methods have recently been proposed for privacy preserving data mining of multidimensional data records. Several privacy preserving data mining technologies are studied in [8[clearly and the merits and shortcomings of these technologies are analysed. Methods such as *k*-anonymity, *l*-diversity, *t*-closeness, classification, association rule mining are all designed to prevent identification to preserve the underlying sensitive information. the Application of several techniques for preserving privacy on experimental dataset are illustrated in [4] and their effects on the results is revealed.

A. Anonymization Algorithms

Anonymization methods have emerged as an important tool to preserve individual privacy when releasing privacy sensitive data sets. A survey on most of the common attacks techniques for anonymization-based PPDM & PPDP is presented in [5] and their effects on Data Privacy are explained. A new approach for building classifiers using anonymized data by modeling anonymized data as uncertain data is propsed in [6]. In [7], a novel technique called slicing is proposed, which preserves better data utility than generalization and can be used for attribute disclosure protection and membership disclosure protection.

B. Perturbation Algorithms

Perturbation-based PPDM approach introduces random perturbation to individual values to preserve privacy before data are published. In [9] the use of truncated non-negative matrix factorization (NMF) with sparseness constraints for data perturbation is investigated. [In 10], the possibility 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 [11] which is 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.

C. Cryptographic & Secured Sum Computation Algorithms

A new privacy preserving collaborative protocol is shown in [19] with light weight overhead which uses a new similarity measure approach. An innovative encryption algorithm with an efficient aggregation operator is proposed in [20] an Efficient Conjunctive Query (ECQ) scheme is used to achieve zero data and query privacy leakage in [21]. A secure k-means data mining approach in the distributed environment is proposed in [22] by combining the advantages of both RSA public key cryptosystem and homomorphism encryption scheme, a model of hierarchical management on the cryptogram is put forward in the algorithm [23].

The aim of secure multiparty computation is to enable parties to carry out distributed computing tasks in a secure manner. In [13], a survey is made on the basic paradigms and notions of secure multiparty computation and reviewing the issue of efficiency and the difficulties involved in constructing highly efficient protocols. Various efficient fundamental secure building blocks such as Fast Secure Matrix Multiplication (FSMP), Secure Scalar Product (SSP), and SecureInverse of Matrix Sum (SIMS) is studied in [14]. Secure multi-party multi-data ranking protocol is proposed in [15] which is secure in the semi-honest model. An innovative protocol [16] which uses both actual and idyllic model to provide more security and privacy. A protocol to compute the secured sum with zero leakage probability is provided in [17] and a protocol that is secure under the semi-honest adversarial model as well as stronger non-disruptive malicious model is provided in [18]

D. Fuzzy based PPDM

A set of fuzzy-based mapping techniques is compared in [24] in terms of their privacy-preserving property and their ability to retain the same relationship with other fields. In [25], a method to extract global fuzzy rules from distributed data with the same attributes in a privacy-preserving manner is proposed. In [26], 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 [27] and the applicability of fuzzy k-member clustering to privacy preserving pattern recognition is studied in [28].

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 samples. A fuzzy variant of k-member clustering is proposed in [29] with the goal of improving the quality of data summarization with k-anonymity. This method is also applied to collaborative filtering. In [30] a secure framework for privacy preserving fuzzy co-clustering is proposed for handling both vertically and horizontally distributed cooccurrence matrices. A method to hide fuzzy association rule is proposed in [31] using modified apriori algorithm in order to identify sensitive rules to be hidden.

E. PPDM with Neural Network Learning(NNL)

Learning the structure of Bayesian network on distributed heterogeneous data is addressed in [32] and [33]. A simple privacy-preserving protocol for learning the parameters of Bayesian network on vertically partitioned databases with better performance, full privacy, and complete accuracy is presented in [34]. A probabilistic neural network (PNN) committee machine for Peer-to-Peer datamining is described in [35]. The l-diversity principle is combined with k-anonymity concepts in [36], so that background information cannot be exploited to successfully attack the privacy of data and the data disclosure probability and information loss are possibly kept negligible.

IV. COMPARISON OF RECENT RESEARCHES ON PPDM

Table I shows the researches and algorithms available for various PPDM methods with Cryptography, Fuzzy logic and Neural network learning.

TABLE I
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	√	√	√
Regression		√	
Aggregation	√	√	

TABLE III
ADVANTAGES OF VARIOUS PPDM ALGORITHMS

Techniques used	Reference &Year	Advantage
k-Means algorithm	[22] 2014	Secure k-means data mining approach with correctness even in distributed environment.
Efficient Conjunctive Query (ECQ) scheme for Query Control	[21] 2013	Achieve conjunctive query without data and query privacy leakage
Slicing Technique for Attribute disclosure protection	[7] 2012	Better data utility than generalization and can be used for attribute and membership disclosure protection. It can handle high-dimensional data.
Privacy preserving collaborative protocol for Similarity measure	[19] 2012	Uses a new similarity measure and has light weight overhead.
Random perturbation matrix for Classification Mining	[12] 2010	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.
RSA public key and homomorphism encryption scheme for finding Association rules	[23] 2009	Effective privacy preserving distributed mining algorithm with RSA public key cryptosystem & homomorphism encryption scheme.
Building classifiers using anonymized data.	[6] 2009	Propose collecting all necessary statistics during anonymization and releasing these together with the anonymized data.
Non-negative matrix factorization (NMF) for perturbation	[9] 2007	The use of truncated non-negative matrix factorization (NMF) with sparseness constraints for data perturbation
Perturbation -based PPDM to Multilevel Trust (MLT-PPDM)	[11] 2006	Prevents diversity attacks by properly correlating perturbation across copies at different trust levels and robust against diversity attacks with respect to the privacy goal.

. Table II shows the advantages of algorithms available for PPDM methods like Anonymization, Pattern recornition etc.

In Table III, the advantages of various algorithms recently developed using Fuzzy logic, Neural network and Cryptography techniques for Secured Multi party Sum computation (SMC) is listed.

.

TABLE IIII
ADVANTAGES OF VARIOUS SMC ALGORITHMS

Techniques used	Reference &Year	Advantage
Actual and idyllic model Secured Multi-party sum	[16] 2013	By using actual and idyllic models, more security and privacy is provided.
Distributed changing neighbours k-secure sum protocol	[17] 2013	Zero leakage probability.
Multiplicative random projection matrices for Aggregation	[10] 2012	Computes statistical aggregates and is directly related to clustering, principal component analysis, and classification.
Secure multi-party multi-data ranking protocol for Secured Multi-party sum	[15] 2011	Secure in the semi-honest model and supports privacy-preserving sequential pattern mining solution.
Semi-honest adversarial model and Non-disruptive malicious mode for Secured Multi-party sum	[18] 2010	Secure under the semi-honest adversarial model and non-disruptive malicious model. This protocol exchanges O(N log n) messages for the non-disruptive malicious model, where as other protocols requires O(N)
Secured data comparison between the encrypted data	[20] 2010	Provide a robust and efficient aggregation operator to provide aggregation result with greater data privacy.
Fuzzy c-regression method for Secured sum Computation	[26] 2009	Generation of Synthetic data allows third parties to do statistical computations with a limited risk of disclosure
Bayesian network using K2 algorithm for Secured Computation	[33] 2009	Various Secured Computation algorithms are provided with improved resistance against colluding attack and can even be used over public channels.

Table IV shows the advantages of algorithms available for PPDM methods using Fuzzy logic for performing association rule mining, patter recognition, classification, collaboration filtering, clustering etc

Table V shows the advantages of algorithms available for PPDM methods using neural network learning for environments where data is distributed horizontally or vertically.

TABLE IV
ADVANTAGES OF VARIOUS PPDM USING FUZZY ALGORITHMS

Techniques used	Reference &Year	Advantage
Fuzzy based Association Rule Hiding	[25] 2014	Extracts global fuzzy rules from distributed data.
Fuzzy co-clustering for Collaborative filtering	[30] 2014	A secure framework for privacy preserving fuzzy co-clustering for handling both vertically and horizontally distributed co-occurrence matrices.
Fuzzy k-member clustering for Pattern recognition	[28] 2013	Perform supervised pattern recognition keeping a certain anonymization level.
Fuzzy k-member clustering for k-anonymity & Collaborative filtering	[29] 2012	Uses a fuzzy variant of k-member clustering with the goal of improving the quality of data summarization with k-anonymity
Modified apriori algorithm with Fuzzy Data for Association rule hiding	[31] 2011	A method to hide fuzzy association rule using modified apriori algorithm to extract rules and identify sensitive rules. Provides efficient information hiding with minimum side effects.
Fuzzy based clustering	[27] 2008	Intuitionistic fuzzy clustering for its application to privacy.
Fuzzy classifier based on an aggregation operator for Pattern Recognition	[37] 2008	This classifier with a better learning and uses a PI-type membership function, product aggregation reasoning rule (operator).

TABLE V
ADVANTAGES OF VARIOUS PPDM USING NEURAL NETWORK ALGORITHMS

Techniques used	Reference &Year	Advantage
Probabilistic neural network (PNN) committee machine for Peer-to-Peer datamining	[35] 2013	It performs weight based ensemble selection of best peer members to find class-specialized Peer modules in P2P systems.
l-diversity with k-anonymity	[36] 2010	Kohonen Self Organizing Feature Maps (SOMs) is used to successfully maintain privacy of data and negligible data disclosure probability & information loss.
Bayesian network Learning for Distribution Data	[32] 2005	Effective and privacy-preserving version of the B&BMDL algorithm to construct the structure of a Bayesian network with complete accuracy, full privacy, ideal universality, and better performance for both binary and non-binary discrete data.
Bayesian network Learning on vertically partitioned databases	[34] 2005	This simple solution provides better performance, full privacy, and complete accuracy for learning Bayesian networks (both structure and parameters) on vertically partitioned data with very little overhead.

V. CONCLUSIONS

In this paper, a broad survey on various privacy preserving data mining algorithms developed recently using Fuzzy logic, Cryptography and Neural network learning techniques is made. Advantages of various algorithms shown in Table II, Table III, Table IV and Table V helps to identify the algorithm 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.

ACKNOWLEDGMENT

The authors wish to acknowledge all those who have supported in this work.

REFERENCES

[1] 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.

[2] 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 .

[3] 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.

[4] 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.

[5] 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

[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] 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

[8] 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

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

[11] 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.

[12] 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

[13] Yehuda Lindell, and Benny Pinkas, 2009, Secure Multiparty Computation for Privacy-Preserving Data Mining, The Journal of Privacy and Confidentiality (2009), pp : 59-98

[14] 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

[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] 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

[17] 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

[18] 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

[19] 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

[20] 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

[21] 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

[22] 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

[23] Gui Qiong, Cheng Xiao-hui, 2009, A Privacy-Preserving Distributed Method for Mining Association Rules Artificial Intelligence and Computational Intelligence, 2009. AICI '09. International Conference on Volume: 4 DOI: 10.1109/AICI.2009. 486 pp: 294 – 297

[24] 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

[25] 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

[26] 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

[27] 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

[28] 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

[29] Honda, K. ; Kawano, A. ; Notsu, 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

[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] 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

[32] [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] 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

[34] 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:1196

[35] 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

[36] 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

[37] Ghosh, A., Meher, S. K., & Shankar, B. U. (2008). A novel fuzzy classifier based on product aggregation operator. Pattern Recognition, 41(3), 961-971

[38] Marinai, S., Gori, M., & Soda, G. (2005). Artificial neural networks for document analysis and recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(1), 23-35.

[39] C. Gentry. Fully homomorphic encryption using ideal lattices. In Proceedings of the 41st ACM Symposium on Theory of Computing – STOC 2009, pages 169–178. ACM, 2009

[40] C.L. Blake and C.J. Merz, UCI Repository of machine learning databases, http://www.ics.uci.edu/ learn/MLRepository.html, Irvine, CA: University of California, Department of Information and Computer Science, 1998.

[41] PRIVACY-PRESERVING DATA MINING: Models and Algorithms, Advances in Database System- Volume 34s, edited by Charu C. Aggarwal and Philip S. Yu; ISBN : 978-0-387-70991-8