A Survey on Privacy Preserving Data Mining Techniques

Cina Mathew
Assistant Professor
Kristu Jyothi College of management and Technology

***Abstract*:- The emerging privacy concern has become a major obstacle in storing and sharing of data. The proliferation of data can be useful, but it must be performed in a way that preserves user's privacy. This is not straightforward, because the proliferated data need to be protected against several privacy threats. Various algorithms have been designed for privacy-preserving data mining, that can be classified into three categories i.e., privacy by policy, privacy by statistics, and privacy by cryptography. We review algorithms like; Randomization, k-anonymization, and distributed privacy-preserving data mining etc., derive insights on their operation, and compare their advantages and disadvantages. We also provide a study of the computational and hypothetical boundaries involved with privacy-preservation over high dimensional data sets.**

Keywords: PPDM, Anonymization, Perturbation, Cryptography

I. INTRODUCTION

Recent years have seen unprecedented growth in applicability of Computer Science in day-to-day activities. Organizations, community and individuals show an augmented trend of storing their data in cloud. The huge amount of data collected can be used for analyzing trends of markets and individual or society. Data mining activities involve extracting knowledge from this massive pool of data. The sensitive information about the individuals may be disclosed creating ethical or privacy issues. Many individual therefore don't share their data publicly, creating data unavailability. Privacy of individual should not be compromised under any case. PPDM has gained popularity so as to address the privacy concerns while data mining is being carried out [1].

II. PRIVACY PRESERVING DATA MINING [PPDM]

Privacy preserving data mining is an area of data mining that is used to protect sensitive information from unsolicited or unsanctioned disclosure. It consists of techniques and methodologies of data mining, which would be used to fulfil privacy constraint and it also maintains the utilization of data for data mining. Privacy preserving data mining is solely based on description of privacy that defines the different attributes of data. It depicts which attribute is sensitive and hence required to ensure confidentiality constraint [2, 3]. The block diagram of PPDM is shown in figure;

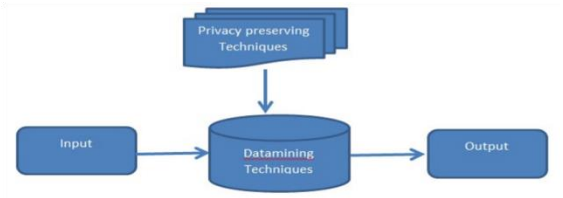

Figure 1: Blockdiagram of PPDM

III. PRIVACY PRESERVING DATA MINING (PPDM) METHODS

In this section we focus on number of methods that have recently been proposed for privacy preserving data mining. A survey on several privacy preserving data mining technologies are studied in [5] and the pros and cons of these technologies are analysed. In this paper, we analyse an overview of the state-of-the-art in privacy preserving data mining. In order to perform the privacy preservation most methods for computations use some form of transformation on the data. Typically, such methods reduce the granularity of representation in order to reduce the privacy. This reduction in granularity results in some loss of effectiveness of data and mining algorithms. This is the natural trade-off between information loss and privacy. Methods such as k-anonymity, l-diversity, t-closeness, classification, association rule mining are all designed to prevent identification to preserve the privacy of sensitive information. The Application of several techniques for preserving privacy on experimental dataset is illustrated in [6] and their effects on the results are revealed.

A. Anonymization Algorithms

Anonymization methods have emerged as an effective means to achieve privacy preservation. In these methods some part of the original data, for instance, through generalization, compression, etc., is transformed and let the transformed data cannot be combined with other information to reason about any personal privacy information. The implementation of privacy preservation mainly concentrates on two aspects: (1) How to ensure that the data been used without privacy disclosure? (2) How to make the data to be better utilized? So, the problem to be solved urgently is a trade-off between privacy preservation and data utilization.

B. Perturbation Techniques

Data Perturbation introduces random perturbation to individual values to preserve privacy before data are published. These techniques are statistically based methods that seek to protect confidential data by adding random noise to confidential, numerical attributes, thereby protecting the original data. Data Perturbation techniques are not encryption techniques, where the data is first modified, then (typically) transmitted, and then received, ‘decrypted’ back to the original data. But the intent of these techniques is to allow authentic users the capability to access important aggregate statistics (such as mean, correlations, etc.) from the entire database while ‘protecting’ the individual identify of a record.

C. Distributed Privacy Preservation

In many cases, individual entities may wish to derive aggregate results from data sets which are partitioned across these entities. For this purpose, Privacy preserving distributed data mining is used that aims to design secure protocols which allow multiple parties to conduct collaborative data mining while protecting the privacy of their data. Such partitioning may be horizontal (when the records are distributed across multiple entities) or vertical (when the attributes are distributed across multiple entities). In this the individual entities may consent to limited information sharing with the use of a variety of protocols and may not desire to share their entire data sets. The whole effect of such methods is to preserve privacy for each individual entity, while deriving aggregate results over the entire data.

The advantages and limitations of some of the PPDM techniques are tabulated in Table 1.

Technique	Advantages	Limitations
Anonymization Technique	Secrecy of data are preserved.	More information loss
Perturbation Technique	Preserves various attributes independently.	Original data values cannot be regained.
Distributed Data Mining	It is an efficient technique. Simple and supports large databases.	Minimal information loss.
Cryptography Technique	Data encryption and decryption using keys is accurate and improves security.	Complexity and number of keys are proportional.

Table1: Advantages and limitations of PPDM techniques

IV. COMPARISON OF RECENT RESEARCHES ON PPDM

Table 2 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 which are being used to obtain the cryptosystem.

S. No	Authors	Year of Publication	Technique Used for PPDM	Approach	Result and Accuracy
1.	Y.Lindell, B.Pinkas [6]	2000	Cryptographic Technique	Sensitive data are encrypted in different levels using keys.	The complexity increases when more than a few keys are involved. Also, it does not hold good for large databases.
2	L. Sweeney[7]	2002	K- Anonymity	Information about an individual contained in a release cannot be distinguished from at least k-1 individual’s information.	Privacy is Preserved at greater levels.
3	J. Vaidya and C. Clifton[8]	2002	Association Rule	Data are vertically distributed into segments.	Ensures privacy.
4	HillolKargupta, Souptik Datta, Qi Wang and Krishnamoorthy Sivakumar[9]	2003	Data Perturbation	Data Privacy is preserved by adding random noise.	Randomization Techniques are used to generate random matrices.

5	Charu C Aggarwal, Philip S. Yu[10]	2004	Condensation Approach	Condenses the data into multiple groups of predefined size. The different records are not distinguishable.	The use of pseudo-data no longer requires to redesign the data mining algorithms, since they have the original format.
6	SlavaKisilevich, Lior Rokach, Yuval Elovici, BrachaShapira[12]	2010	Anonymization	Anonymization uses generalization and suppression for data hiding.	Background knowledge and Homogeneity attacks of K-Anonymity algorithm do not preserve sensitivity of an individual.
7	P.Deivanai, J. JesuVedhaNayahi andV.Kavitha1[3]	2011	Hybrid Approach	Hybrid Approach is a combination of different techniques which combine to give an integrated result.	It uses Anonymization and suppression to preserve data.
8	George Mathew, Zoran Obradovic[14]	2011	Decision Tree	An approach which is technical, methodological and should give judgemental knowledge.	A graph-based framework for preserving patient's sensitive information.
9	M. N. Kumbhar and R. Kharat[16]	2012	Association Rule By Horizontal and Vertical Distribution	Different approaches in the field of Association rule is reviewed.	The performance of all models is analyzed in terms of privacy, security and communications.
10	Savita Lohiya and LataRagha[17]	2012	Hybrid Approach	A combination of K-Anonymity and Randomization.	It has more accuracy and original data can be regained.
11	George Mathew, ZoranObradovic[19]	2012	Distributed Privacy Preserving	Provides an algorithm to collaboratively build a better decision-making model	It improves the overall accuracy of a classification model
12	Shweta Taneja, Shashank Khanna, SugandhaTilwalia, Ankita[21]	2014	Cryptography, Anonymization, Perturbation	A tabular comparison of work done by different methods.	Cryptography and Random Data Perturbation methods perform better than the other existing methods.
13	M. Antony Sheela, K. Vijayalakshmi[24]	2017	Partition Based Perturbation	Applied techniques on the vertically partitioned data.	When the threshold value is reached,the individual data is changed.
14	JalpeshVasa, PanthiniModi[25]	2018	t-closeness	Anonymization based techniques used to preserve privacy by reducing the granularity.	Wasn't that perfect, so opted differential privacy.

V. CONCLUSION

Privacy is the major concern to protect the sensitive data in today's world. People are very much anxious about their sensitive information which they don't want to share. In this paper our survey focuses on the existing literature present in the field of Privacy Preserving Data Mining. The primary objective of PPDM is promoting algorithm to hide sensitive data or offer privacy in data mining. From our analysis, we have found that that there is no single PPDM technique in existence that outshines every other technique with relation to each possible criterion such as use of data, performance, difficulty, compatibility with procedures for data mining, and so on. All methods perform in a different way depending on the type of data as well as the type of application or domain. But still from our analysis, we can conclude that Distributed data mining and Random Data Perturbation methods perform better than the other existing methods.

VI. REFERENCES

[1] Alpa Shah and Ravi Gulati, "Privacy Preserving Data Mining: Techniques, Classification and Implications - A Survey", International Journal of Computer Application, Vol. 137 – No 12, March 2016, 40-46.

[2] AlShwaier and A. Z. Emam, "Data Privacy OnEHealth Care System", International Journal of Engineering, Business and Enterprise Applications, (2013).

[3] Xu, Yang, Tinghuai Ma, Meili Tang, and Wei Tian. "A survey of privacy preserving data publishing using generalization and suppression." Appl. Math 8, no. 3, pp. 1103-1116, (2014).

[4] Y.Li, B.Vinzamuri, C.K.Reddy, Constrained elastic net based knowledge transfer for health care information exchange, Data Mining Knowl. Discov. 29 (4) (2015) 1094–1112.

[5] Jian Wang, YongchengLuo ; Yan Zhao ; Jiajin Le, 2009,A Survey on Privacy Preserving Data Mining, First International Workshop on Database

[6] 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.

[7] Y. Lindell, B.Pinkas, “Privacy preserving data mining”, in proceedings of Journal of Cryptology, 5(3), 2000.

[8] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy”, in proceedings of Int'l Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 2002.

[9] J. Vaidya and C. Clifton, “Privacy preserving association rule mining in vertically partitioned data”, in The Eighth ACM SIGKDD International conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, CA, July 2002, IEEE 2002.

[10] H. Kargupta and S. Datta, Q. Wang and K. Sivakumar, “On the Privacy Preserving Properties of Random Data Perturbation Techniques”, in proceedings of the Third IEEE International Conference on Data Mining, IEEE 2003.

[11] C. Aggarwal , P.S. Yu, “A condensation approach to privacy preserving data mining”, in proceedings of International Conference on Extending Database Technology (EDBT), pp. 183–199, 2004. 746

[12] A. Machanavajjhala, J.Gehrke, D. Kifer and M. Venkitasubramaniam, "I-Diversity: Privacy Beyond k-Anonymity", Proc. Int'l Con! Data Eng. (ICDE), p. 24, 2006

[13] SlavaKisilevich, LiorRokach, Yuval Elovici, BrachaShapira, “Efficient Multi-Dimensional Suppression for K-Anonymity”, inproceedings of IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 3. (March 2010), pp. 334-347, IEEE 2010.

[14] P.Deivanai, J. JesuVedhaNayahi and V.Kavitha,” A Hybrid Data Anonymization integrated with Suppression for Preserving Privacy in mining multi party data” in proceedings of International Conference on Recent Trends in Information Technology, IEEE 2011.

[15] G. Mathew, Z. Obradovic,” A PrivacyPreserving Framework for Distributed Clinical Decision Support”, in proceedings of 978-1-61284852-5/11/$26.00 ©2011 IEEE.

[16] A. Parmar, U. P. Rao, D. R. Patel, “Blocking based approach for classification Rule hiding to Preserve the Privacy in Database” , in proceedings of International Symposium on Computer Science and Society, IEEE 2011.

[17] M. N. Kumbhar and R. Kharat, “Privacy Preserving Mining of Association Rules on horizontally and Vertically Partitioned Data: A Review Paper”, in proceedings of 978-1-46735116-4/12/$31.00_c, IEEE 2012.

[18] S. Lohiya and L. Ragha, “Privacy Preserving in Data Mining Using Hybrid Approach”, in proceedings of 2012 Fourth International Conference on Computational Intelligence and Communication Networks, IEEE 2012.

[19] Martin Beck and Michael Marh¨ofer,” Privacy-Preserving Data Mining Demonstrator”, in proceedings of 16th International Conference on Intelligence in Next Generation Networks, IEEE 2012.

[20] George Mathew, ZoranObradovic, "Distributed Privacy Preserving Decision System for Predicting Hospitalization Risk in Hospitals with Insufficient Data", in proceedings of 2012 11th InternationalConference on Machine Learning and Applications

[21] Yuan Zhang , Sheng Zhong, "A privacy-preserving algorithm for distributed training of neural network ensembles", Neural Comput&Applic (2013) 22

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

[23] Abel N. Kho, John P. Cashy, Karthryn L. Jackson, Adam R. Pah, SatyenderGoel, JornBoehnke, John Eric Humphries, Scott Duke Kominers, Bala N. Hota, Shanon A. Sims, Bradley A. Malin, Dustin D. French, Theresa L. Walunas, David O. Meltzer, Erin O. Kaleba, Roderick C. Jones, Wiliam L. Galanter,"Design and implementation of a privacy preserving in electronic health record linkage tool in chicago" journal of American Medical Informatics Association,(2015) 22(5)

[24] V. Baby , N. Subhash Chandra , " Privacy-Preserving Distributed Data Mining Techniques: A Survey ", International Journal of Computer Applications (0975 – 8887) Volume 143 – No.10, June 2016

[25] M. Antony Sheela, K. Vijayalakshmi,"Partition Based Perturbation for Privacy Preserving Distributed Data Mining" CYBERNETICS AND INFORMATION TECHNOLOGIES ,2017Volume17, No 2

[26] Review of different privacy preserving techniques of PPDM