International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 1

Privacy Preserving Data Mining for Healthcare Record: A Survey of Algorithms

Musavir Hassan
Research Scholar, University of Kashmir Hazratbal Srinagar

Muheet Ahmad Butt
Assistant Professor, University of Kashmir Hazratbal Srinagar

Majid Zaman
Assistant Professor, University Kashmir Hazratbal Srinagar

ABSTRACT

Due to the wide deployment of sensitive information on the internet, privacy preserving data mining has been studied extensively in recent years. The emerging privacy concern has become a major obstacle in storing and sharing of medical data. The proliferation of medical data can be highly useful, but it must be performed in a way that preserves patient'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, however, the privacy concerns and data utilization requirements on different parts of medical data may be quite different. In this paper, we present a survey of the state-of-the-art algorithms that have been proposed for publishing medical data in a privacy preserving way. We review algorithms like; Randomization, k-anonymization, and distributed privacy-preserving data mining etc., derive insights on their operation, and highlight their advantages and disadvantages. We also provide discussion of the computational and hypothetical boundaries associated with privacy-preservation over high dimensional data sets.

Keywords: *PPD (Privacy preserving Data mining), electronic health records (EHRs), PPDDM (Privacy preserving Distributed Data mining), EMR (Electronic Medical Record), and CPR (Computer-based Patient Record).*

I. Introduction

In recent years, electronic health records (EHRs) have been commonly adapted at many health care facilities in an attempt to improve the quality of patient care and increase the productivity and efficiency of health care delivery. An Electronic Health Record (EHR) is a digital version of a patient’s medical history. It is a longitudinal record of patient health information generated by one or several encounters in any healthcare providing setting. The term is often used interchangeably with EMR (Electronic Medical Record) and CPR (Computer-based Patient Record). It encompasses a full range of data relevant to a patient’s care such as demographics, problems, medications, physician’s observations, vital signs, medical history, immunizations, laboratory data, radiology reports, personal statistics, progress notes, and billing data. The EHR system automates the data management process of complex clinical environments and has the potential to streamline the clinician’s workflow. These EHRs can not only aid in various daily health care operations but also help in accurately diagnosing diseases, if utilized appropriately.

A. Privacy Preserving Data mining

Privacy preserving data mining is an area of data mining that seeks to protect sensitive information from unsolicited or unsanctioned disclosure. It consists of those 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 [1, 2].The block diagram of PPDM is shown in figure

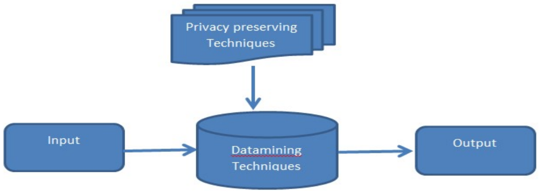

Figure 1: Block diagram of PPDM

B. Privacy Preserving Data Mining In Electronic Health Record

A PPDM technique is highly useful in healthcare systems to minimise the disclosure of sensitive personal information. Thus provides ability to make analysis of large data sets of medical research systems to gain knowledge and to find out curative measures of fatal diseases without breaking privacy of patients. To achieve privacy constraints privacy preserving data mining uses various techniques. One such technique includes perturbation approach in which the confidential information regarding any patient's record is perturbed using random process. It apparently distorts sensitive data values by changing them by adding, subtracting or any other mathematical formula. Several other privacy preserving data mining algorithms are used to build various classification models for predictive measures and for finding association rules among various attributes of data set with a significant accuracy [2]. EHR enables the good availability of medical data and provides various features such as "Home health monitoring system" in which a patient could use the services of monitoring system to measure and analyse his disease statistics from his own place every day. For each patient, the EHRs were comprehensively collected from the following primary sources of information [3]:

- **Demographic information** such as year of birth, gender, weight, and geographical location of each patient.
- **Diagnosis information** consists of the ICD9 Codes.
- **Allergy and immunization** consists of a list of allergies and vaccination records.
- **Laboratory information** consists of lab test observations for lab panels, and the lab test results received from lab facilities.
- **Medication and prescription** consists of the medication history, where each medicine is identified by National Drug Code (NDC), and prescription records.
- **Patient smoking status** is an ordinal status variable maintained on a yearly basis.
- **Transcripts** consist of visited document records including allergy, medication, and diagnosis information. In this system internet is used to provide connectivity between user's personal computer and hospital system and fed the health data from hospital to user system and vice-versa. The data can be represented in the form of various disease attributes such as patient's blood pressure, diabetic measures such as glucose level and other measures of chronic diseases. This system fed health data on the web application which is running on patient's system and transfer to the hospital's system. The transferred data is used by healthcare attendant or Doctor for monitoring and diagnosis of health condition of patient. Researcher in healthcare sector requires these patient's health data to find out new discovery in preventive and curative of deadly diseases. This individual patient's health data could be integrated and transformed in large data sets which could be used to perform various researches in medical field and also for data mining purpose. The integrated data which is collectively made from individual information could be used for sharing among various research institutes and hospitals. But the problem in data sharing is that it would result in privacy violation of patients. To solve this problem Hospital management could use PPDM. In PPDM system, hospital received patient health record from "Home Monitoring Online system". It saves one copy to original database and sends another copy to randomizer. Randomizer possesses functionality to randomize or perturbation of those attributes of health record which could be disclosed the identity of any patient. The perturb data is then saved to database which is developed for research and mining purpose and which contains data with satisfying privacy constraints.

II. 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 [4] and the pros and cons of these technologies are analysed. In this paper, we will study an overview of the state-of-the-art in privacy preserving data mining. In order to perform the privacy preservation most methods for privacy 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 management or 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 underlying sensitive information. The Application of several techniques for preserving privacy on experimental dataset is illustrated in [5] 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 a problem that the academia and industry need to be solved urgently is a trade-off between privacy preservation and data utilization.

B. Perturbation Algorithms

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 all PPDM techniques are tabulated in Table 1.

Technique	Advantages	Limitations
Anonymization technique of PPDM	Data owner's sensitive or private data are to be secreted.	More information loss, Linking attack
Perturbation technique of PPDM	Preserves various attributes independently.	Information loss and Cannot regenerate original data values.
Randomized Response technique of PPDM	It provides good efficiency. Simple and useful for keeping the individual information secretly.	Loss in individual's information. Not much good for database containing several attributes.
Cryptography technique of PPDM	Data transformation is accurate and protected. Provides better privacy and data utility.	It is particularly hard to scale if multiple parties are involved.

Table1: Advantages and limitations of all PPDM techniques

III. 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	A technique through which sensitive data can be encrypted. There is also a proper toolset for algorithms of cryptography.	This approach is especially difficult to scale when more than a few parties are involved. Also it does not hold good for large databases.
2	L. Sweeney[7]	2002	K- Anonymity	A record from a dataset cannot be distinguished from at least k-1 records whose data is also in the dataset.	K- Anonymity Approach is able to preserve privacy.
3	J. Vaidya and C. Clifton[8]	2002	Association Rule	Distribution of data vertically into segments.	Distribution Based Association Rule Data Mining provides privacy.
4	Hillol Kargupta, Souptik Datta, Qi Wang and Krishnamoorthy Sivakumar[9]	2003	Data Perturbation	They tried to preserve data privacy by adding random noise, while making sure that the random noise still preserves the “signal” from the data so that the patterns can still be accurately estimated.	Randomization-based Techniques are used to generate random matrices.

5	CharuC.Aggarwa, Philip S. Yu[10]	2004	Condensation Approach	This approach works with pseudo-data rather than with modifications of original data, this helps in better preservation of privacy than techniques which simply use modifications of the original data.	The use of pseudo-data no longer necessitates the redesign of data mining algorithms, since they have the same format as the original data.
6	A. Machanavajjhala, J. Gehrke, D. Kifer and M. Venkitasubramaniam [11]	2006	L-Diversity Algorithm	If there are ‘l’ ‘well represented’ values for sensitive attribute then that class is said to have L-Diversity.	It is better than KAnonymity in preserving Data mining.
7	Slava Kisilevich, Lior Rokach, Yuval Elovici, Bracha Shapira[12]	2010	Anonymization	Anonymization is a technique for hiding individual’s sensitive data from owner’s record. K-anonymity is used for generalization and suppression for data hiding.	Background Knowledge and Homogeneity attacks of K-Anonymity Algorithm do not preserve sensitivity of an individual.
8	P.Deivanai, J. Jesu Vedha Nayahi and V.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.

9	George Mathew, Zoran Obradovic[14]	2011	Decision Tree	An approach which is technical, methodological and should give judgmental knowledge.	A graph-based framework for preserving patient's sensitive information.
10	Anita Parmar, Udai Pratap Rao, Dhiren R. Patel[15]	2011	Blocking Based Technique	Finding sensitive attribute and then they replace known sensitive values with unknown values ("?"). Finally the sanitized dataset is generated from which sensitive classification rules are no longer mined.	Unknown Values help in preserving privacy but reconstruction of original data set is quite difficult.
11	M. N. Kumbhar and R. Kharat[16]	2012	Association Rule By Horizontal and Vertical Distribution	Different approaches in the field of Association rule are reviewed.	The performance of all models is analyzed in terms of privacy, security and communications.
12	Savita Lohiya and Lata Ragha[17]	2012	Hybrid Approach	A combination of K-Anonymity and Randomization.	It has a better accuracy and original data can b reconstructed.
13	Martin Beck and Michael Marh¨ofer[18]	2012	Anonymizing Demonstrator	Making a demonstrator with user friendly interface and performs Anonymization.	Swapping and Recording can be applied to enhance the utility.
14	George Mathew, Zoran Obradovic[19]	2012	Distributed Privacy Preserving	DIDT algorithm to collaboratively build a better decision-making model	It improves the overall accuracy of a classification model

15	Yuan Zhang • Sheng Zhong[20]	2013	Distributed Privacy Preserving	privacy-preserving distributed algorithm for training neural network ensembles using AdaBoost.M2.	guarantees each party's data privacy in the semi-honest model. As the cost of privacy protection, the accuracy of ensemble learning is only slightly affected.
16	Shweta Taneja, Shashank Khanna, Sugandha Tilwalia, Ankita[21]	2014	PPDDM, Anonymization, Perturbation	A tabular comparison of work done by different authors is presented.	Cryptography and Random Data Perturbation methods perform better than the other existing methods.
17	Abel N Kho etal[22]	2015	Distributed Privacy Preserving	developed and distributed a software application that performs standardized data cleaning, preprocessing, and hashing of patient identifiers to remove all protected health information.	the software achieved a sensitivity of 96% and a specificity of 100%
18	V. Baby , N. Subhash Chandra,[23]	2016	Distributed Privacy Preserving		
19	M. Antony Sheela, K. Vijayalakshmi[24]	2017	Partition Based Perturbation	Perturbation and anonymization technique performed on the vertically partitioned data.	Makes each individual to perturb their data when the threshold value is reached.

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 conceal sensitive data or offer privacy. From our analysis, we have found that that there is no single PPDM technique in existence that outshines every other techniques with relation to each possible criteria 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

References

1) AlShwaier and A. Z. Emam, "Data Privacy On E-Health Care System", International Journal of Engineering, Business and Enterprise Applications, (2013).

2) 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).

3) 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.

4) 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

5) 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.

6) Y. Lindell, B.Pinkas, "Privacy preserving data mining", in *proceedings of Journal of Cryptology*, 5(3), 2000.

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

8) 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.

9) 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.

10) 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

11) 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

12) Slava Kisilevich, Lior Rokach, Yuval Elovici, Bracha Shapira, "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.

13) P.Deivanai, J. Jesu Vedha Nayahi 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.

14) G. Mathew, Z. Obradovic," A Privacy-Preserving Framework for Distributed Clinical Decision Support", *in proceedings of 978-1- 61284-852-5/11/$26.00 ©2011 IEEE.*

15) 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.

16) M. N. Kumbhar and R. Kharat, "*Privacy Preserving Mining of Association Rules on horizontally and Vertically Partitioned Data: A Review Paper*", in proceedings of 978-1-4673-5116- 4/12/$31.00_c, IEEE 2012.

17) 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.

18) 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.

19) George Mathew, Zoran Obradovic, "Distributed Privacy Preserving Decision System for Predicting Hospitalization Risk in Hospitals with Insufficient Data", in *proceedings of* 2012 11th International Conference on Machine Learning and Applications

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

21) Shweta Taneja, Shashank Khanna, Sugandha Tilwalia, 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

22) Abel N. Kho, John P. Cashy, Karthryn L. Jackson, Adam R. Pah, Satyender Goel, Jorn Boehnke, 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)

23) 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*

24) M. Antony Sheela, K. Vijayalakshmi,"Partition Based Perturbation for Privacy Preserving Distributed Data Mining" CYBERNETICS AND INFORMATION TECHNOLOGIES ,2017Volume 17, No 2