International Institution for Technological Research and Development
Volume 1, Issue 1, 2015

A Survey on Privacy Preserving Data Mining Techniques in Health Sector

Mrs. Madhushree B
L.J. Institute of Engineering and Technology
Mentoring Masters in Computer Engineering
Gujarat Technological University
Ahmedabad-Gujarat-India.
Email: bmadhushree.lj@gmail.com
Telephone: (+91)7405285827

Nidhi G Pandya
L.J. Institute of Engineering and Technology
Mentoring Masters in Computer Engineering
Gujarat Technological University
Ahmedabad-Gujarat-India.
Email: pandya.nidhi91@gmail.com
Telephone: (+91) 8153945646

Abstract

Data mining is a process of extracting useful knowledge from large data sets. The typical process of data collection and data dissemination result in a possible risk of privacy threats and attacks. Some private information about individuals, businesses and organizations has to be suppressed before it is shared or published. In recent years, privacy preserving data mining has been studied extensively, because of the wide proliferation of sensitive information on the internet. In this paper we discussed about the recent trends involved in privacy preservation in medical domain.

Keywords: *Privacy preservation, Electronic Medical Records(EMR), eHealth, security.*

1. INTRODUCTION

In recent years, data mining has been viewed as a threat to privacy because of the wide spread proliferation of electronic data maintained by corporations. This has leads to increased concerns about the privacy of the underlying data. In the last few decades a number of approaches and techniques such as classification, association rule mining have been proposed for modifying or transforming the data in such a way so as to preserve the privacy. Preservations of individuals information is an essential for the data owners to ensure his privacy. Privacy plays an important role in data publishing. Data mining process allows a company to use large amount of data to develop correlations and relationships among the data to improve the business efficiency. Therefore privacy preserving data mining has become important field of research. The Data Mining technology can develop these analyses on its own, using commix of statistics, artificial intelligence, machine learning algorithms, and data stores. In order to face the challenging risk, some researchers have been proposed as a remedy of this awkward situation, which target at accomplishing the balance of data utility and information privacy when publishing dataset. The ongoing research is called Privacy Preserving Data Publishing. Balancing the privacy of the data as per the legitimate need of the user is the major problem. The original data is modified by the sanitization process to conceal sensitive knowledge before release so the problem can be addressed. Privacy preservation of sensitive knowledge is addressed by several researchers in the form of association rules by suppressing the frequent item sets. As the data mining deals with generation of association rules, the change in support and confidence of the association rule for hiding sensitive rules is done. A new concept named „not altering the support‟‟ is proposed to hide an association rule. Confidentiality issues in data mining. A key problem that arises in any en masse collection of data is that of confidentiality. The need for privacy is sometimes due to law (e.g., for medical databases) or can be motivated by business interests. The irony is that data mining results rarely violate privacy. The objective of data mining is to generalize across populations, rather than reveal information about individuals.[1]

2. PRIVACY REQUIREMENTS

Privacy is an important concern while disclosing various categories of electronic data including business dataand medical data for data mining. Privacy can be interpreted in two ways. For instance, privacy is so crucial with respect to medical data, since it contains sensitive information like type of disease. Especially for doing medical data mining the original data should be available for making accurate predictions otherwise lead to impractical solutions. Any kind of disclosure related to the person- specific information leads to many problems including ethical issues. Therefore extra care should be taken to protect privacy of individuals before publishing such data. On the other hand, the privacy can be interpreted as preventing unwanted disclosure of information while performing data mining on aggregate results. Thus, privacy can be addressed at various levels in the process of data mining. For entire database security both privacy and security measures are needed. For better understanding of the concept of privacy, we would like to distinguish between the two related issues security and privacy according to HealthCare data. And the remaining sub sections provide an introduction to privacy issues and privacy policies.

2.1. Security Vs Privacy

Even though the two terms, security and privacy are synonymously used, these can be treated as two related, but separate issues: *i) Security* is defined as the mechanism for protecting the entire HealthCare data including the ability to control access to patient information, safeguard from unauthorized disclosure, alteration, loss or destruction of patient information. Security is typically accomplished through operational and technical controls. The three fundamental security goals are Confidentiality, Integrity and Availability. And *ii) Privacy* is a more specific term which is defined as the right of an individual to keep his/her individual health information from being disclosed. Privacy is typically accomplished though policies and procedures. With this understanding it is clear that security is necessary, but not sufficient for addressing privacy. Today several known PPDM techniques are available and these are extensively studied in literatures.

2.2. Privacy Issues

The privacy issue varies according to the data in use and the context it is used. But, the most important issue is how to provide privacy while preserving information (that is without loss of information).The methods like attribute removal, data hiding, and data compression can be applied on the data set to provide privacy, but will lead to information loss. Another important issue is regarding the computational overhead. Complex procedures like cryptographic techniques create additional overhead both technical and computational. The main parameter that affects the feasibility of implementing a secure protocol based on the generic constructions is the size of the best combinatorial circuit that computes the function that is evaluated. For a distributed environment, when the number of parties becomes bigger, the communication and computational cost grow exponentially. The PPDM algorithm that addresses all these issues is still a myth. Even though no such generic solutions are available to address all privacy issues, some research has focused on finding efficient protocols for specific problems that balance privacy, data utility and computational feasibility.

2.3 Privacy Policies

To ensure privacy the researcher has to address various privacy breaches (attacks) which need a high level of attention. Privacy breach happens when one’s exact privacy information are directly linked to him. Since it is difficult to identify all types of attacks periodically, the privacy providers can follow certain kind of policies provided by different countries such as HIPAA of US, Data Protection Act of UK. Federal Health Insurance Portability and Accountability Act (HIPAA) of US sets the floor on privacy rights which means states are free to adopt more stringent medical privacy laws but states cannot pass any law that takes away HIPAA rights.

3. RELATED WORK

In SCOOP-The Social Collaboratory for Outcome Oriented Primary care Morgan Price Morgan Price et al. have been proposed Primary care research networks (PCRN) to overcome slow, knowledge discovery and translation in primary care clinics who used computer-based Electronic Medical Record (EMR) systems. Primary care research networks (PCRN) helps in taking major design decisions and finding their underlying rationales. Their current and future work will focus on three main aspects (1) develop and integrate privacy enhancing technologies to provide formal guarantees on specific confidentiality properties by Allowing peer-to-peer communication among the Endpoints can enable the use of data exchange protocols that provide formal privacy guarantees. (2) scale the network quantitatively (more clinics) as well as qualitatively (more data types, EMR types and delivery models). As a result, the network will become more heterogeneous both in terms of data quality as well as system/node connectivity and improved approach to failed / unresponsive Endpoints. (3) Develop tool support for authoring and validating research questions built h-Query which will be a visual query composer, its capabilities to date are very limited and researchers often have to resort to encoding their queries directly using.[2]

In Privacy and eHealth-enabled Smart Meter Informatics Georgios Kalogridis et.al. Proposes sensor data mining algorithms that help infer health/well-being related lifestyle patterns and anomalous (or privacy-sensitive) events it also solved centralized (database) health data privacy issues. Algorithms enable a user-centric context awareness at the network edge, which can be used for decentralized eHealth decision making and privacy protection by design. The main hypothesis of this work involves the detection of atypical behaviors from a given stream of energy consumption data recorded at eight houses over a period of a year for cooking, microwave, and TV activities. This method brings appliance monitoring, privacy, and anomaly detection together within a healthcare context, which is readily scalable to include other health-related sensor streams. it helps to will help close the gap among nationwide eHealth instrumentation, health indications, atypical events, and their connection to privacy analytics.[3]

In A New Model for Privacy Preserving Sensitive Data Mining M. Prakash solve the issue of protecting privacy in micro data publishing. Publishing data about individuals without revealing sensitive information about them is an important problem. k-anonymity and I-Diversity has been previously used mechanism for protecting privacy but mechanisms are insufficient to protect the privacy issues like Homogeneity attack, Skewness Attack, Similarity attack and Background Knowledge Attack so A new privacy measure called "(n, t)-proximity" is proposed which is more flexible model it achieves more privacy and less utility.[4]

In Integrity Preservation and Privacy Protection for Medical Images with Histogram-Based Reversible Data hiding et.al Hsiang-Cheh Huang discussed the integrity preservation of medical image, and the data protection and authentication for the medical data. With reversible data hiding, the medical images and medical records of the same patient can be authenticated, and proper treatment can be performed by medical doctors. The proposed scheme is suitable for X-ray or CT medical images, and it has the potential to be integrated into the databases for managing the medical images in the hospital. In *the reversible data hiding scheme that data, including patients' private information and the diagnosis data, can be hidden into the medical image by some means. Later on, the medical image containing data might be retrieved while necessary, and both the original image and the hidden data can be perfectly recovered.[5]*

In Task Independent Privacy Preserving Data Mining on Medical Dataset E. Poovammal et.al implement a task independent technique which preserves the information, privacy and utility of the data. Algorithm is applied on the original data table to alter only the sensitive raw data before applying any mining methods. There is no information loss. Also any number of sensitive attribute can be handled. The complexity of the algorithm is of the order of the size of the table. It applicable to ensure privacy of patients and the security of the medical data.[6]
In A framework for privacy-preserving healthcare data sharing Lei Chen et al. focused on develop practical methods to balance health care data sharing and privacy protection introduce a framework for privacy preserving data sharing with the view of practical application in more comprehensive way. The framework focuses on three key problems of privacy protection during data sharing which are privacy definition and detection, privacy protection policy management, privacy preserving health care data sharing. System for privacy preserving electronic medical records sharing is showed as an application of the framework. In future plan to improve the framework by implemented all the components in a more complicated application and try to improve the efficiency for large dataset process.[7]

In Privacy Preserving Distributed Learning Clustering of HealthCare Data Using Cryptography Protocols Ahmed M. Elmisery et al. present a novel clustering algorithm for vertically partitioned data; they test the performance of that algorithm based on experiments and complexity analysis. Later they presented a private version of this protocol using protocols based on homo morphed encryption. Our protocol is robust against colluding attack.[8]

In Privacy-preserving range set union for rare cases in healthcare data J.Y. Chun et al. suggest a privacy-preserving ‘range set union’ protocol that can be used to find rare cases in the private medical datasets of individuals. They have suggested privacy-preserving range set union protocol PPRSU to find rare cases while preserving privacy. The range set unionRt1, t2 is a set of elements that at least t1 parties and at mostt2 parties have in their private sets. PPRSU can be used to make new set operations, as well as conventional set operations. PPRSU does not reveal any other information, except the information that could be inferred from the range set union and the size of each private set.[9]

4. Conclusion

After studying various technique in involved in privacy preservation in medical domain data mining we conclude that mining in medical field helps us to detect about various dieses reasons of dices and similarity. we also show methods to protect electronic health related data and security of published record in centralized and distributed database on different server. Efficiency and accuracy is still challenge in this domain and researcher need to work on technique which helps to improve efficiency and privacy mechanism of system.

Reference

[1]"A Survey on Privacy Preservation Recent Approaches and Techniques", International Journal of Innovative Research in Computer and Communication Engineering, vol. 2, no. 11, p. 8, 2015.

[2]Price, Morgan, Jens H. Weber, and Glen McCallum. "Scoop-the social collaboratory for outcome oriented primary care." In Healthcare Informatics (ICHI), 2014 IEEE International Conference on, pp. 210-215. IEEE, 2014.

[3]Kalogridis, Georgios, and Saraansh Dave. "Privacy and eHealth-enabled smart meter informatics." In e-Health Networking, Applications and Services (Healthcom), 2014 IEEE 16th International Conference on, pp. 116-121. IEEE, 2014.

[4]Prakash, Mangal, and G. Singaravel. "A new model for privacy preserving sensitive Data Mining." In Computing Communication & Networking Technologies (ICCCNT), 2012 Third International Conference on, pp. 1-8. IEEE, 2012.

[5]Huang, Hsiang-Cheh, and Wai-Chi Fang. "Integrity preservation and privacy protection for medical images with histogram-based reversible data hiding." In Life Science Systems and Applications Workshop (LiSSA), 2011 IEEE/NIH, pp. 108-111. IEEE, 2011.

[6]Poovammal, E., and M. Ponnavaikko. "Task independent privacy preserving data mining on medical dataset." In Advances in Computing, Control, & Telecommunication Technologies, 2009. ACT'09. International Conference on, pp. 814-818. IEEE, 2009.

[7]Chen, Lei, Ji-Jiang Yang, Qing Wang, and Yu Niu. "A framework for privacy-preserving healthcare data sharing." In e-Health Networking, Applications and Services (Healthcom), 2012 IEEE 14th International Conference on, pp. 341-346. IEEE, 2012.

[8]Elmisery, Ahmed M., and Huaiguo Fu. "Privacy preserving distributed learning clustering of healthcare data using cryptography protocols." In *Computer Software and Applications Conference Workshops (COMPSACW), 2010 IEEE 34th Annual*, pp. 140-145. IEEE, 2010.

[9]Chun, Ji Young, Dong Hoon Lee, and Ik Rae Jeong. "Privacy-preserving range set union for rare cases in healthcare data." IET Communications 6, no. 18 (2012): 3288-3293.