Gagan Sharma, 10:6
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

Survey on Privacy Preserving Mining Techniques for Health Care Data Analysis

M.Tech. Scholar Gagan Sharma, Dr. Sadhana Mishra
CSE Department
LNCT College Bhopal,MP,India

Abstract- Nowadays, valuable data are generated and collected rapidly from numerous rich data sources. Following the initiatives of open data, many organizations including health, education, etc. are willing to share their data such as open data regarding parking violations. While there have been models to preserve privacy of sensitive personal data like patient data for health informatics, privacy of individuals who violated regulations should also be protected. Hence, this article, presented a model for supporting privacy-preserving big data analytics on temporal open data. Paper has performed a survey on recent methodology proposed by different researcher. Some of data mining methods were also describe in the paper which help in information extraction. Classification of the privacy preserving mining techniques was also introduced, in the article.

Keywords- Data mining, Information Extraction, Association Rule.

I. INTRODUCTION

In the current day and age, data collection is ubiquitous. Collating knowledge from this data is a valuable task. If the data is collected and mined at a single site, the data mining itself does not really pose an additional privacy risk; anyone with access to data at that site already has the specific individual information [1].

While privacy laws may restrict use of such data for data mining, controlling such use is not really within the domain of privacy-preserving data mining technology. The technologies discussed in this paper are instead concerned with preventing disclosure of private data: mining the data when researchers aren't allowed to see it. If individually identifiable data is not disclosed, the potential for intrusive misuse (and the resultant privacy breach) is eliminated [2].

Electronic health records (EHRs) are widely used by a variety of healthcare organizations in an effort to enhance patient care and enhance the efficiency of healthcare delivery. In complex clinical environments, the EHR system accelerates the clinician's workflow by automating the data management process. When utilized effectively, these EHRs not only facilitate many routine health care tasks, but also help in the accurate identification of diseases. Individuals' access to their medical records is facilitated by EHRs. In addition, they come with a home health monitoring system that allows patients to measure and evaluate their symptoms every day [3].

Privacy and security are two terms used interchangeably under different contexts. But both are related to each other and at the same time entirely separate issues. The three fundamentals of security are Confidentiality. Integrity and Availability [3, 4]. In context of Census data, security can be termed as the facility for controlling person-specific access information, protect it from unauthorized disclosure, modification, loss or destruction of his information.

Security can be accomplished through controls based on operational and technical knowhow. In contrast privacy is very specific. It can be termed as a right of an individual to keep his/her personal information from being disclosed. Privacy can be accomplished through policies and procedures. Person's personal information which may lead to his identification may not be disclosed under ethical grounds.

II. TYPES OF PRIVACY PRESERVING

There are a number of techniques are present to preserve the privacy. These are classified as in this figure2 [5].

Centralized/Data Publishing Scenario: Centralized scenario is also known as the data publishing scenario. The data can be published publically in its original form. Even though it is also possible that there is no encryption is done in the format. Some types of alterations have to be applied before disclosing the data to maintain the privacy of data of individuals.

The techniques used in this scenario like Neural Network based, Fuzzy based and Anonymization based are discussed here. Hayden Wimmer and Loreen Powell [4] provide the framework of PPDM using neural network. They discussed the different PPDM techniques and their effects of on machine learning algorithms. Firstly, they read a data set file and provided the classification using machine learning algorithm. Secondly, they read the same data file and they applied a privacy structure on it and applied the same machine learning algorithm on it.

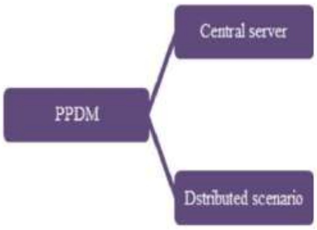

Fig 1. Classification of PPDM.

b) Distributed Privacy Preserving Data Mining: Privacy preserving data mining permits the distribution of personal and private data for the purpose of analysis. Recently there are several techniques and methods have been elaborated for the same. In most of the techniques some type of alteration is performed to the data to provide the protection to the privacy. Three approaches are used in distributed privacy preserving data mining.
These are Perturbation model, Cryptographic protocols and Anonymization. Perturbation is also a most popularly used technique in PPDM and it is basically used for electronic health record (EHR) [5]. In data perturbation the data is distorting using the noise. Both additive noise and multiplicative noise can be used for the purpose of distortion. The major challenge that is faced in data perturbation is that the balancing the ratio of protection of privacy and the quality of data. Both of these factors are k contradictive factors.

III. LITERATURE SURVEY

In [6] privacy of Internet of Health (IOH) data was done in three modules. First was the LSH (Locality-Sensitive Hashing) into multi-source IoH data fusion and integration so as to secure the sensitive information of patients hidden in the past IoH data.

Second was the IoH data without patient privacy after LSH process, we bring forth a similar IoH data record search method for subsequent IoH data mining and analyses, so as to balance the IoH data availability and privacy. Finally based on a dataset collected by real-world users, we validate the advantages of the proposed work in this paper, through a set of pre-designed experiments.

In [7] author focus on *k*-nearest neighbor (*k*NN) in this study to realize classification. Although several studies have already attempted to address the privacy problems associated with *k*NN computation in a cloud environment, the results of these studies are still inefficient. In this paper, we propose a very efficient and privacy-preserving *k*NN classification (P*k*NC) over encrypted data. While the amount of computation (encryptions/decryptions and exponentiations) and communication of the most efficient *k*NN classification proposed in prior studies is bounded by $O(kln)$, that of the proposed P*k*NC is bounded by $O(ln)$, where l is the domain size of data and n is the number of data.

In [8] author propose an efficient protocol to evaluate whether an item set is frequent or not under the encrypted mining query on supermarket transactions. To improve the mining efficiency, we design a blocking algorithm. In this algorithm, we separate the encrypted transactions into blocks and only calculate bilinear pairings on cipher texts of part blocks instead of all cipher texts, which helps cut down the computation cost of the mining process. Finally, we evaluate the performance of our protocol by conducting theoretical analyses and simulator experiments in the aspects of computation cost, security, correctness, and running time.

In [9] author proposed a mixture-model-based label propagation algorithm against malicious adversaries with corruption abilities. Privacy constraints in this paper are mainly focused on individual privacy, which means no individual data value should be disclosed and no information can be traced back to a specific site. In addition, another constraint should be included is that no site except *P*0 shall gain the information on the task.

In [10] author work on the medical research data for the improvement by the collaborative association rule mining on vertically partitioned healthcare data. Privacy of patients must be preserved during this collaboration. Paper further proposed an efficient approach for privacy preserving association rule mining in the vertically partition healthcare data for discovering the correlation related to disease and preserving the privacy of the patients. Finally analyze the proposed scheme with the medical examination data and outpatient data. The analysis of results shows that the association between diseases and symptoms discovered using the collaborative mining as well as privacy of the patients is preserved.

IV. TECHNIQUES OF PRIVACY PRESERVING MINING

Data Perturbation: In this technique available data is modified before it is passed to Data Mining. There are number of ways to modify the data like swapping, adding noise etc. but after modification quality of the released data is maintained. In case of noise addition technique, Data owner add some random number (noise) to input data.

This random number is generally drawn from a normal distribution with zero mean and a small standard deviation, which preserves statistics of original data. Then data owner share this noisy data for Data Mining task. Data owner also shares distribution of the noise added to the original data. By using this distribution and noisy data, data miner reconstruct original data set's distribution. But data miner cannot retrieve actual data values. This enabled a Data Mining algorithm to construct a much more accurate result without revealing actual data [2].

1. Anonymization-Based Method:

This technique presented in the literature as means of achieving privacy [11]. Anonymization is widely used by researchers owing to the lower communication and computation costs of the same as compared to their cryptographic counterparts. Information loss is one of the important issues in an anonymization-based approach. The information loss calculates the difference between the original databases and the anonymized databases. The information loss increases with the increase in the level of the generalization and/or suppression method. Generally, the information loss should be less to achieve higher data utility [12].

2. K-anonymity:

This is a key concept that was introduced to address the risk of re-identification of anonymized data through linkage to other datasets. K-Anonymity is able to prevent identity disclosure, i.e., a record in the k-anonymized data set cannot be mapped back to the corresponding record in the original data set. However, in general, it may fail to protect against attribute disclosure. Some critics of k-anonymization take issue with the fact that achieving a re-identification risk of zero is impractical or impossible [13, 14].

3. Condensation Technique:

Condensation approach compresses and packs the raw input data into multiple groups or clusters. Each group or cluster has constraint which is defined for it in terms of its size. This size is referred as the level of that privacy preserving approach. Greater is the level, the greater will be the amount of privacy. This size is chosen in a way so as to preserve k-anonymity. After condensing data into clusters, statistics of data in each group is analyzed and maintained separately for each cluster. This statistics from each cluster is used further to generate pseudo data for corresponding clusters. In the process of data mining, data owner publish this pseudo data instead of original data. Various data mining tasks use this pseudo data as input. In this way actual data remains hidden from other parties [15].

This technique is referred as condensation because of its approach of using condensed statistics of the clusters in order to generate pseudo-data. In this approach, a greater amount of information is lost because of the condensation of a larger number of records into a single statistical group entity.

Following figure shows steps in followed in the condensation approach.

V.CONCLUSION

Machine learning is increasingly used in the most diverse applications and domains, whether in healthcare, to predict pathologies, or in the financial sector to detect fraud. However, when it contains personal information, full access may be restricted due to laws and regulations aiming to protect individuals' privacy.

Therefore, data owners must ensure that any data shared guarantees such privacy. Removal or transformation of private/ sensitive information is among the most common techniques. Paper has discussed a variety of techniques and algorithms used for maintaining privacy or securing private and sensitive information. Still, there can be improvements in the defined algorithms. Better and improved algorithms must be defined which provides more security.

REFERENCES

[1] José Luis Fernández-Alemán, Inmaculada Carrión Señor, Pedro Ángel Oliver Lozoya, Ambrosio Toval, "Methodological Review-Security and Privacy in electronic health records: A systematic literature review", Journal of Biomedical Informatics (2013).

[2] Yan Zhao, Ming Du, Jiajin Le, Yongcheng Luo, "A Survey on Privacy Preserving Approaches in Data Publishing", First International Workshop on Database Technology and Applications, 2009.

[3] Xun Yi, Yanchun Zhang, "Privacy-preserving distributed association rule mining via semi-trusted mixer", Data & Knowledge Engineering 63 (2007) 550–567.

[4] Hayden Wimmer, Loreen Powell, A Comparison of the Effects of K-Anonymity on Machine Learning Algorithms, (IJACSA) International Journal of Advanced Computer Science and Applications, 5(11) (2014) 155-160.

[5] Alpa Shah and Ravi Gulati, Privacy Preserving Data Mining: Techniques, Classification and Implications - A Survey, International Journal of Computer Applications (0975 – 8887) 137(12) (2016) 40-46.

[6] Q. Zhang, B. Lian, P. Cao, Y. Sang, W. Huang and L. Qi, "Multi-Source Medical Data Integration and Mining for Healthcare Services," in IEEE Access, vol. 8, pp. 165010-165017, 2020.

[7] J. Park and D. H. Lee, "Parallelly Running k-Nearest Neighbor Classification Over Semantically Secure Encrypted Data in Outsourced Environments," in IEEE Access, vol. 8, pp. 64617-64633, 2020.

[8] C. Ma, B. Wang, K. Jooste, Z. Zhang and Y. Ping, "Practical Privacy-Preserving Frequent Itemset Mining on Supermarket Transactions," in IEEE Systems Journal, vol. 14, no. 2, pp. 1992-2002, June 2020.

[9] Z. Li, L. Yang and Z. Li, "Mixture-Model-Based Graph for Privacy-Preserving Semi-Supervised Learning," in IEEE Access, vol. 8, pp. 789-801, 2019.

[10] Nikunj Domadiya, Udai Pratap Rao. "Privacy Preserving Distributed Association Rule Mining Approach on Vertically Partitioned Healthcare Data". Procedia Computer Science Volume 148, 2019.

[11] K. Oishi, Y. Sei, Y. Tahara, and A. Ohsuga, "Semantic diversity: Privacy considering distance between values of sensitive attribute," Comput. Secur., vol. 94, Jul. 2020, Art. no. 101823.

[12] S. Darwish, M. Madbouly, and M. El-Hakeem, "A database sanitizing algorithm for hiding sensitive multi-level association rule mining,'' Int. J. Comput. Commun. Eng., vol. 3, no. 4, p. 285, 2014.

[13] M. N. Dehkordi, K. Badie, and A. K. Zadeh. ''A novel method for privacy preserving in association rule mining based on genetic algorithms,'' J. Softw., vol. 4, no. 6, pp. 555–562, Aug. 2009.

[14] R. Crawford, M. Bishop, B. Bhumiratana, L. Clark, and K. Levitt, ''Sanitization models and their limitations," in Proc. Workshop New Secur. Paradigms, 2006, pp. 41–56.

[15] Suchitra Shelke, Prof. Babita Bhagat. "Techniques for Privacy Preservation in Data Mining". International Journal of Engineering Research & Technology, Vol. 4 Issue 10, October-2015.