Survey On Health Data Privacy Preserving Techniques and Features

Mrs. Ritu Raikwar, Prof. Dr. Manmohan Singh

Research Scholar, Professor
Department of Computer Science
IES Group of Institution, Bhopal, India
ritu.raikwar2009@gmail.com, kumar.manmohan4@gmail.com

Abstract— **The primary objective of Data Mining is to extract knowledge from vast datasets. This involves the application of data mining techniques to uncover meaningful information and patterns within extensive databases. In the realm of health science, machine learning has become increasingly indispensable, leveraging its ability to derive valuable insights from high-dimensional data. However, this often necessitates the amalgamation of research and patient data from various institutions and hospitals—a challenging feat due to privacy constraints. In a recent survey, the paper explores methodologies proposed by different researchers, delving into data mining methods that aid in information extraction. The paper also discusses features employed by these approaches to address data privacy concerns. Furthermore, it elaborates on various data mining techniques and provides a comprehensive examination of evaluation parameters for comparing privacy-preserving methods.**

Index Terms—Data mining, Healthcare records, Information Extraction, Association Rule.

INTRODUCTION

In the present era, healthcare has emerged as a top-priority human concern, marked by the generation, storage, and frequent utilization of healthcare-related data. Among the pivotal components of healthcare systems, Electronic Health Records (EHR) play a vital role, offering numerous advantages to stakeholders. EHR facilitates patient access to medical records, reducing the need for expensive tests, radiology, and redundant imaging. Whether a patient receives treatment in different medical centers or geographically distant locations, physicians across these centers can access records seamlessly through EHR. Another noteworthy benefit includes access to a patient's medication history, aiding physicians in prescribing new drugs. Additionally, EHR supports the use of patients' medical records for research, contributing to the exploration of novel treatment methods.

Data Mining, as a process of extracting meaningful insights from extensive datasets, involves discovering valuable knowledge from large volumes of data stored in databases or archives. Through data mining, consistent patterns, interesting facts, or advanced information can be derived by examining databases from various perspectives. Mined information may encompass clusters, rules, patterns, or classification models. High-utility pattern mining, a prominent research area in pattern mining, focuses on enhancing the efficiency of mining algorithms. This paper introduces an improved strategy applied to the EFIM algorithm, aiming to eliminate non-candidate items from the global and local header tables early on, thereby reducing the search space and enhancing efficiency.

Comprehensive information about an individual often contains private details, and the careless handling of such data can result in an immediate breach of privacy. Privacy, in this context, refers to the condition of being shielded or secluded from the view or presence of others. When data mining intersects with privacy concerns, it implies safeguarding an individual's information from unauthorized access by others. Privacy is not violated as long as personal information is not misused; however, once sensitive information is disclosed, preventing its misuse becomes challenging. Preserving privacy is crucial for preventing information leakage and ensuring the effective utilization of extensive data. This involves maintaining the confidentiality of electronic data throughout the data mining process. Privacy preservation is a major consideration for the success of the data mining process, with Privacy Preserving Data Mining (PPDM) aiming to protect individuals' personal data or classified knowledge without compromising the full utilization of the required data.

LITERATURE SURVEY

McGhin et al. discussed the role of Blockchain in healthcare in their paper [], addressing challenges in system security, interoperability, data sharing, and mobility within Electronic Health Records (EHR). The paper explains how Blockchain can effectively manage these challenges. The authors introduced several platforms for implementing Blockchain in healthcare, including Gem Health Network, OmniPHR, Medrec, Inclusive Social Networking System (PSN), and Virtual Resources.

In another paper [], Wang et al. proposed a secure data-sharing scheme incorporating data deduplication and sensitive information hiding in cloud-assisted electronic medical systems. To safeguard sensitive information privacy and enhance deduplication efficiency, the authors replaced patient-sensitive information with wildcards before encrypting entire electronic medical records. Authorized researchers can decrypt and access the records under the condition that sensitive information is hidden. The paper also classifies diagnostic information into different types based on the duplicate ratio, allowing authorized researchers to selectively download data. The proposed scheme is resilient against brute-force attacks and single-point-of-failure attacks.

Lee et al. [] introduced a utility-preserving anonymization method for privacy-preserving data publishing (PPDP). To maintain data utility, the method consists of a utility-preserving model, counterfeit record insertion, and a catalog of counterfeit records. The anonymization algorithm employs a full-domain generalization algorithm. The evaluation compares the proposed method with existing approaches, considering information loss measured through various quality metrics and the error rate of analysis results.

Khanna et al. [] outlined a simple federated learning algorithm implementing differential privacy to ensure privacy when training a machine learning model on data distributed across different institutions. The authors tested their model by predicting breast cancer status from gene expression data, achieving a level of accuracy and precision comparable to a single-site non-private neural network model while enforcing privacy.

In another contribution [], Onesimu et al. proposed an attribute-focused privacy-preserving data publishing scheme. The scheme includes a fixed-interval approach to protect numerical attributes and an improved l-diverse slicing approach to safeguard categorical and sensitive attributes. Extensive experiments with real-world datasets demonstrated that the proposed scheme's classification models on anonymized datasets yield approximately 13% better accuracy than benchmarked algorithms.

DATA FEATURES FOR PRIVACY PRESERVING

A. Data Distribution: Currently, privacy protection data mining algorithms operate either on centralized data or distributed data. Distributed data encompasses horizontally partitioned data, where different database records exist in different sites, and vertically partitioned data, where each database record's attribute values are stored in different sites [12].

B. Data Distortion: This technique involves modifying the original database record before release to achieve privacy protection objectives [13]. Data distortion methods include perturbation, blocking, merging or aggregation, swapping, and sampling. These methods entail altering attribute values or transforming the granularity of attribute values.

C. Data Mining Algorithms: Privacy-preserving data mining algorithms encompass classification mining, clustering, association rule mining, Bayesian networks, and others.

D. Data or Rules Hidden: This technique involves concealing original data or rules governing the original data. Due to the complexity of reconstructing hidden rules from the original data, some individuals have proposed heuristic methods to address this issue.

E. Privacy Protection: To safeguard privacy, careful data modification is essential to maintain high data utility. This can be achieved for various reasons, such as [14]: modifying data based on adaptive heuristic methods and selectively altering values rather than all values to minimize information loss. Encryption technologies, like secure multiparty computation, ensure safety by allowing each site to know only their input and nothing about others. Data reconstruction methods can then reconstruct the original data distribution from random data.

TECHNIQUES OF PRIVACY PRESERVING MINING

Cryptographic Approaches:
In numerous scenarios, multiple entities may desire to share aggregated private data without divulging sensitive information on their end [16]. For instance, distinct superstores holding sensitive sales data might wish to collaborate in understanding aggregate trends without revealing the specifics of individual stores. Achieving this necessitates secure cryptographic protocols for sharing information across these different entities.

The Randomization Technique:
The randomization technique employs data distortion methods to generate private representations of records [17]. In most cases, individual records cannot be retrieved, only aggregate distributions can be reconstructed. These distributions serve data mining purposes, and two types of perturbation exist within the randomization method:
Additive Perturbation: Randomized noise is added to data records, allowing recovery of overall data distributions from the randomized records. Data mining and management algorithms are designed to work with these distributions.
Multiplicative Perturbation: This involves using random projection or random rotation techniques to perturb the records.

Quantification of Privacy:
Measuring the security of various privacy-preservation methods hinges on how the underlying privacy is quantified. Privacy quantification aims to measure the risk of disclosure for a given level of perturbation.

Utility-Based Privacy-Preserving Data Mining:
Most privacy-preserving data mining methods entail a transformation that diminishes the effectiveness of underlying data when applied to data mining methods or algorithms. The tradeoff between privacy and accuracy is inherent, impacted by the specific algorithm employed for privacy preservation. The challenge is to maintain maximum utility of the data without compromising underlying privacy constraints. The paper addresses the design of utility-based algorithms for effective use in specific data mining problems.

Utility-Oriented Pattern Mining:
A general definition of Utility-Oriented Pattern Mining (UPM) is provided, utilizing utility theory and various mining techniques to discover interesting patterns leading to utility maximization and high benefit in business or other tasks [18]. UPM is classified into following categories such as high-utility itemset mining (HUIM), high-utility association rule mining (HUARM), high-utility sequential pattern mining (HUSPM), high-utility sequential rule mining (HUSRM), and high-utility episode mining (HUEM).

The k-Anonymity Method:
A significant method for privacy de-identification is the k-anonymity method [19]. It aims to reduce the granularity of data representation to the extent that a given record cannot be distinguished from at least (k − 1) other records, addressing the potential identification of records using pseudo-identifiers.

Mining Association Rules under Privacy Constraints:
Association rule mining is crucial in data mining, and the paper dedicates several chapters to this problem. Privacy-preserving association rule mining involves challenges in accurately determining association rules on perturbed data and ensuring that output association rules do not leak sensitive data, a problem known as contingency table privacy-preservation in the statistical community and association rule hiding in the database community.
Privacy-Preserving Models:
Horizontal Partitioning: In this method, the different sites may have different sets of records containing the same attributes.
Vertical Partitioning: In this method, the different sites may have different attributes of the same sets of records.

DATA MINING TECHNIQUES

Many of data mining approaches were developed and proposed by the researchers in last few decades [20, 21].
Decision Tree:
Decision tree classification involves learning decision trees from class-labeled training tuples. A decision tree is a tree-like structure with flowchart elements, where each internal node represents a test on an attribute, each branch signifies an outcome of the test, and each leaf node holds a class label. The advantages of decision tree includes simplicity, interpretability, handling of both numerical and categorical data, and robustness. Decision trees perform well with large datasets in a short time, enabling timely decision-making based on analysis.

Nearest Neighbor Classifier:
The k-nearest neighbors' algorithm (k-NN) classifies objects based on the closest training examples in the feature space. It is a type of instance-based or lazy learning and can be used for both classification and regression. The algorithm partitions the space into regions based on the locations and labels of training samples, assigning a point to a class based on the most frequent class label among the k nearest training samples. While Euclidean distance is commonly used, alternative metrics like the overlap metric can be employed for text classification.

Artificial Neural Network:
Neural networks are analytical techniques modeled after cognitive and neurological processes, capable of predicting new observations through a learning process. In data mining, the first step involves designing a specific network architecture with layers and neurons. The network undergoes a training phase where weights are adjusted iteratively to optimally predict sample data. Once trained, the network can generate predictions, representing patterns detected in the data.

Support Vector Machines:
Support Vector Machines (SVMs) are supervised learning methods for regression and classification. Viewing input data as two sets of vectors in an n-dimensional space, SVM constructs a separating hyper plane to maximize the margin between the two data sets. The margin is calculated using parallel hyper planes on each side of the separating hyper plane. The optimal hyper plane is found by using support vectors and margins, aiming to achieve good separation with lower generalization error.

Association Rule:
Association and correlation involve finding frequent item sets in large datasets. This information aids businesses in decision-making, such as catalog design, cross-marketing, and customer shopping behavior analysis. Association rule algorithms generate rules with confidence values less than one, but the sheer number of possible rules for a given dataset is often large, and a substantial proportion may have limited or no value.

EVALUATION PARAMETERS
Direct Discrimination Prevention Degree (DDPD).
This measure quantifies the percentage of discriminatory rules that are no longer discriminatory in the transformed dataset [19].

Direct Discrimination Protection Preservation (DDPP).
This measure quantifies the percentage of the protective rules in the original dataset that remain protective in the transformed dataset [22].

Data Loss: As proposed work provide privacy for the sensitive item set rules with minimum data loss. As in privacy data perturbation make data loss.

Originality: As change in original data is the way to provide privacy in mining. So algorithm that will maintain maximum originality after perturbation is major expectation.

Execution Time: Third parameter is to evaluate execution time of the algorithm that is time taken by the proposed method for execution. Algorithm time is expected after the evaluation of the direct and indirect rules.

CONCLUSION

The significance of data mining lies in its ability to uncover patterns, forecast trends, and discover knowledge across various business domains. Techniques and algorithms like classification and clustering play a crucial role in identifying patterns that contribute to informed decision-making about future business trends and growth. The paper conducts a comprehensive survey on data mining techniques, covering a diverse range of methods. It delves into various techniques and algorithms designed to preserve privacy and secure private and sensitive information. However, there is room for enhancement in the existing algorithms. It is essential to define superior and more advanced algorithms that offer heightened security measures.

REFERENCES

1. Ghalehsefidi, Narges J, Mohammad ND. A Hybrid Algorithm based on Heuristic Method to Preserve Privacy in Association Rule Mining. Indian Journal of Science and Technology. 2016 Jul; 9(27):1–10.
2. Benjamin CMF, Ke W, Rui C, Philip SY. Privacy-Preserving Data Publishing: A Survey of Recent Developments. ACM Computing Surveys. 2010 Jun; 42(4).
3. Charu CA, Philip SY. A General Survey of Privacy Preserving Data Mining Models and Algorithms, Springer US. 2008; 11–52.
4. Privacy. Available from: https://en.wikipedia.org/wiki/ Privacy. Date Accessed: 29/09/2016.
5. A New Model for Privacy Preserving Sensitive Data Mining. Available from: http://ieeexplore.ieee.org/ document/6396017/. Date Accessed: 26/07/2012.
6. Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects.
7. McGhin T, Choo K-KR, Liu CZ, He D (2019) Block chain in healthcare applications: research challenges and opportunities. J Netw Comput Appl 135:1–10.
8. Wang, Z., Gao, W., Yang, M. et al. Enabling Secure Data sharing with data DE duplication and sensitive information hiding in cloud-assisted Electronic Medical Systems. Cluster Computing 26, 3839–3854 (2023).
9. Lee, H., Kim, S., Kim, J.W. et al. Utility-preserving anonymization for health data publishing. BMC Med Inform Decis mak 17, 104 (2017).
10. A Khanna, V. Schaffer, G. Gürsoy and M. Gerstein, "Privacy-preserving Model Training for Disease Prediction Using Federated Learning with Differential Privacy," 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Glasgow, Scotland, United Kingdom, 2022, pp. 1358-1361.
11. J. A. Onesimu, K. J, J. Eunice, M. Pomplun and H. Dang, "Privacy Preserving Attribute-Focused Anonymization Scheme for Healthcare Data Publishing," in IEEE Access, vol. 10, pp. 86979-86997, 2022.
12. C C Aggarwal, P S Yu, “On Static and Dynamic Methods for Condensation-Based Privacy-Preserving Data Mining,” ACM TRANS DATABASE SYST, VOL. 33, NO. 1, 2008, DOI: 10.1145/1331904.1331906.
13. M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim and V. Verykios, “Disclosure Limitation of Sensitive Rules,”Proceedings of the IEEE Knowledge and Data Engineering Workshop, 1999, Pp. 45-52.
14. J Lin, Y Cheng, “Privacy Preserving Itemset Mining Through Noisy Items,” Expert Systems With Applications, Vol. 36, Mar. 2009, Pp. 5711-5717, Doi: 10.1016/J.Eswa.2008.06.052.
15. V.S. Verykios, E. Bertino, I.N. Fovino, L.P. Provenza, Y. Saygin And Y. Theodoridis, “State-Of-The-Art In Privacy Preserving Data Mining,” Acm Sigmod Record, Vol. 33, No. 1, 2004, Pp. 50-57.
16. Pinkas B.: Cryptographic Techniques for Privacy-Preserving Data Mining. Acm Sigkdd Explorations, 4(2), 2002.
17. Rizvi S., Haritsa J. Maintaining Data Privacy in Association Rule Mining. Vldb Conference, 2002.
18. Wensheng Gan, Jerry Chun-Wei Lin, Han-Chieh Chao, Vincent S. Tseng. "A Survey of Utility-Oriented Pattern Mining". Arxiv:1805.10511v2[Cs.Db] 16 Sep 2019.
19. Samarati P., Sweeney L. Protecting Privacy When Disclosing Information: K-Anonymity and Its Enforcement through Generalization and Suppression. IEEE Symp. On Security and Privacy, 1998.
20. Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. Sn Computer. Science 2, 160 (2021).
21. Rutvij H. Jhaveri, A. Revathi, Kadiyala Ramana, Roshani Raut, Rajesh Kumar Dhanaraj, "A Review On Machine Learning Strategies For Real-World Engineering Applications", Mobile Information Systems, Vol. 2022.