International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 06 | June - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

SAFEGUARDING HEALTH CARE INFORMATION USING AN ALGORITHM FROM DATA MINING

Chithra S. Prasad[1], Dr. Smita C Thomas[2], Saranya Raj S[3]

[1]M. Tech Student, Computer Science & Engineering, Mount Zion College of Engineering, Kadamanitta, Kerala, India
[2]Assistant Professor, Computer Science & Engineering, Mount Zion College of Engineering, Kadamanitta, Kerala, India
[3]Assistant Professor, Computer Science & Engineering, Mount Zion College of Engineering, Kadamanitta, Kerala, India

---

Abstract - The privacy of healthcare information is an important part of encouraging data custodians to give accurate records so that mining may proceed with confidence. The application of association rule mining in healthcare data has been widespread to this point in time. Most applications focus on positive association rules, ignoring the negative consequences of particular diagnostic techniques. When it comes to bridging divergent diseases and drugs, negative association rules may give more helpful information than positive ones. This is especially true when it comes to physicians and social organizations (e.g., a certain symptom will not arise when certain symptoms exist). Data mining in healthcare must be done in a way that protects the identity of patients, especially when dealing with sensitive information. However, revealing this information puts it at risk of attack. Healthcare data privacy protection has lately been addressed by technologies that disrupt data (data sanitization) and reconstruct aggregate distributions in the interest of doing research in data mining. In this study, metaheuristic-based data sanitization for healthcare data mining is investigated in order to keep patient privacy protected. It is hoped that by using the Tabu-genetic algorithm as an optimization tool, the suggested technique chooses item sets to be sanitized (modified) from transactions that satisfy sensitive negative criteria with the goal of minimizing changes to the original database. Experiments with benchmark healthcare datasets show that the suggested privacy preserving data mining (PPDM) method outperforms existing algorithms in terms of Hiding Failure (HF), Artificial Rule Generation (AR), and Lost Rules (LR).

Key Words: Machine Learning, Deep Learning

1.INTRODUCTION

Privacy-preserving publishing of micro data has been studied extensively in recent years. Micro data contain records each of which contains information about an individual entity, such as a person, a household, or an organization. Several micro data anonymization techniques have been proposed. Preservation of privacy in data mining has emerged as an absolute prerequisite for exchanging confidential information in terms of data analysis, validation, and publishing. Ever-escalating internet phishing posed severe threat on widespread propagation of sensitive information over the web. Conversely, the dubious feelings and contentions mediated unwillingness of various information providers towards the reliability protection of data from disclosure often results utter rejection in data sharing or incorrect information sharing. This project provides a panoramic overview on new perspective and systematic interpretation of a list published literatures via their meticulous organization in subcategories. The fundamental notions of the existing privacy preserving data mining methods, their merits, and shortcomings are presented. The current privacy preserving data mining techniques are classified based on distortion, association rule, hide association rule, taxonomy, clustering, associative classification, outsourced data mining, distributed, and k-anonymity, where their notable advantages and disadvantages are emphasized. This careful scrutiny reveals the past development, present research challenges, future trends, the gaps and weaknesses. Broadly, the privacy preserving techniques are classified according to data distribution, data distortion, data mining algorithms, anonymization, data or rules hiding, and privacy protection. Intensive research findings over the decades revealed that the existing privacy preserving data mining search approaches are still suffer from major incompleteness including the distributed clients' data to multi semi honest providers, the overhead of computing global mining, incremental data privacy issue in cloud computing, integrity of mining result, utility of data, scalability and overhead performance. Thus, a strong, efficient, and scalable model is essential to surmount these shortcomings. Furthermore, proper anonymization of data is needed to protect the privacy of each client prior to publish. The connection between personal data and personal identification should be vanished. Such an anonymization must not only satisfy underlying privacy requirements but also safeguard the utility of the data.

2. EXISTING SYSTEM

Healthcare process data can include a large number of sensitive variables and highly changeable process behaviors that pose extra privacy issues. The healthcare industry must conform to strict data privacy standards. Privacy protection for such data while maintaining its usefulness for process mining is an ongoing concern in healthcare. For example, encryption does not provide enough privacy protection when used to optimize the value of data for process mining. The accuracy of results may be compromised if methods that conform to more severe privacy rules (such as generalization) are used. In literature, major research has used anonymity, data masking, data perturbation, and cryptography for data privacy. Using dynamic data masking, we are able to achieve format-preserving masking and anonymization without having to manually copy data or remove values tasks which can not only delay analysis, but can weaken the utility of data and introduce the risk of human error.

The cryptographic approach is especially difficult to scale when more than a few parties are involved. It also does not address the question of whether disclosing the final data mining results may violate the privacy of individual records. The perturbation approach does not reconstruct the original data values. New algorithms have been developed to reconstruct the original data distribution. In general, every technique has its own demerits, i.e.information loss, privacy breach, and low data utility.

3. IDENTIFICATION OF PROBLEM DOMAIN

The studies that the integrated health care system has evolved into a critical component of the current health information system. Medical personnel may gain from data mining since it allows them to expand their practicability by sharing and evaluating results with others. This type of information demands a higher level of privacy, which ensures that the association rules remain safe even when data owners use a shared cloud. PPDM has become a significant concern in recent years due to its ability to conceal not just private information but also enable the discovery of essential information using various data mining methods. PPDM is a so-called NP-hard issue. The reason for this is that traditional PPDM algorithms are primarily concerned with concealing sensitive information to the greatest extent feasible. This development might have significant unexpected consequences in terms of missing and artificial costs. Since both side effects are taken into the account, it is difficult to choose the best technique.

4. USE CASE DIAGRAM

The unified modeling language (UML) is a standard language for specifying, visualizing, and documenting the artifacts of software system, as well as for business modeling and other non-software systems. The UML represent the collection of the best engineering practices that have proven successful in the modeling of large and complex systems. The UML is a very important part of developing object - oriented software and the software development process.

To model a system the most important aspect is to capture the dynamic behavior. To clarify a bit in details, dynamic behavior means the behavior of the system when it is running or operating. So only static behavior is not sufficient to model a system rather dynamic behavior is more important than static behavior. In UML there are five diagrams available to model dynamic nature and use case diagram is one of them. The use case diagram is dynamic in nature there should be some internal or external factors for making the interaction. These internal and external agents are known as actors. So, use case diagrams are consists of actors, use cases and their relationships. The diagram is used to model the system/subsystem of an application. A single use case diagram captures a particular functionality of a system.

The purposes of use case diagrams can be as follows:

- Used to gather requirements of a system
- Used to get an outside view of a system
- Identify external & internal factors of system

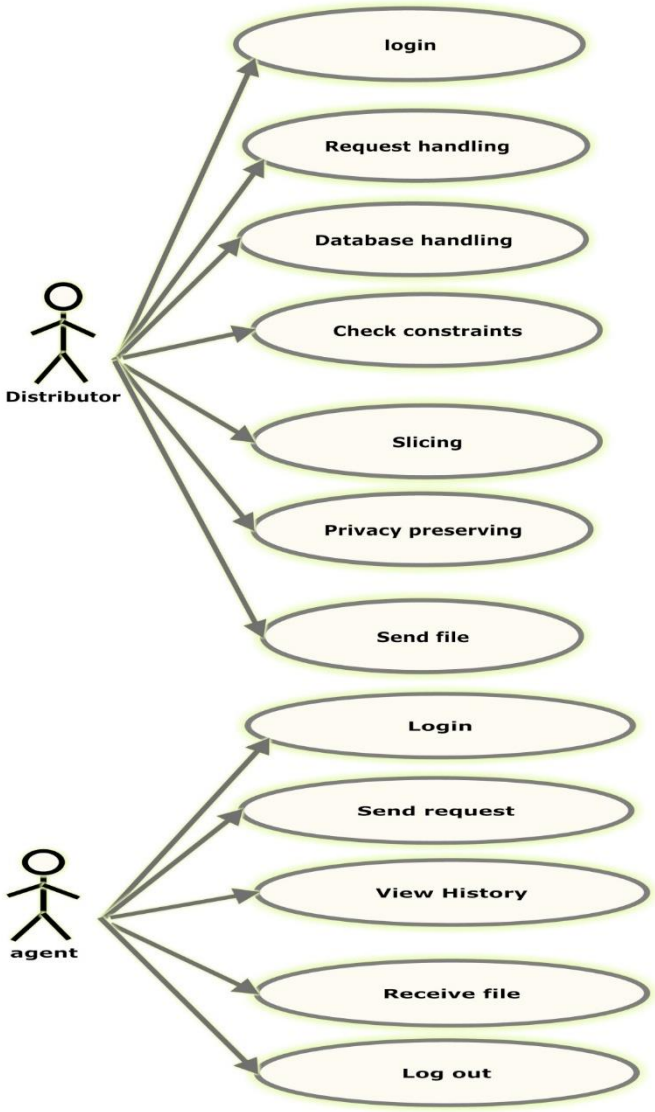

Fig -1: Use Case Diagram

5. MODULES

- **Dataset Preprocessing**

The correlation of discrimination prevention and privacy preservation is dealt in the proposed architecture. Here, the distributor will register the agent and he will decide the level of priority to be given to the agent. For each level, different privacy preservation rule sets are designed which determine the degree of data to be hided or preserved. The distributor will accept the data request from the registered agent only.

Then, appropriate database requested by the agent is chosen and required table is selected. The sensitive attributes which lead to discrimination are specified thereby. Then, the schema is retrieved and the resultant structured data is subjected to grouping. Based on the sensitive attributes, the rules are classified into potentially discriminatory and potentially non-discriminatory groups. Then according to the level assigned by the distributor to each agent, the related rule set is applied.

The rule sets are the various privacy preservation techniques designed on the basis of the degree of data to be hidden according to the type of agent. The transformed dataset is transmitted through the network and will reach to the agent who requests the data set. In the anonymization techniques like generalization, bucketization and slicing, the attribute partitioning is the first and foremost step. When releasing microdata, the sensitive details of an individual will be disclosed. There are two types of information disclosure: identity disclosure and attribute disclosure. Identity disclosure occurs when an individual is related to a particular record in the released table. Attribute disclosure occurs when new information about some individuals is revealed.
The released data helps in inferring the characteristics of an individual more precisely than it would be possible before the data release. Identity disclosure sometimes leads to attribute disclosure.

The main step of anonymization is removing of explicit identifiers (quasi identifiers).
A common anonymization approach is generalization which replaces quasi identifier values with values that are less specific but semantically consistent.
So that, more records will have the same set of quasi identifier values. But, even though k-anonymity protects against identity disclosure, it is insufficient to prevent attribute disclosure. A considerable amount of information loss for high dimensional data is the major drawback of generalization. A new notion of privacy was introduced in called l-diversity, which requires that the distribution of a sensitive attribute in each equivalence class has at least “l” well-represented values. The records are sorted based on the occurrence of sensitive attributes.

Then, group the similar records with set of buckets and analyses it. Combine the set of correlated attributes after diversity check is done. Packetization does not prevent l-diversity is limited in its assumption of adversarial knowledge. It is possible for an adversary to gain information about a sensitive attribute as long as the information about the global distribution of the attribute is known. This assumption generalizes the specific background and homogeneity attacks used to motivate l-diversity.

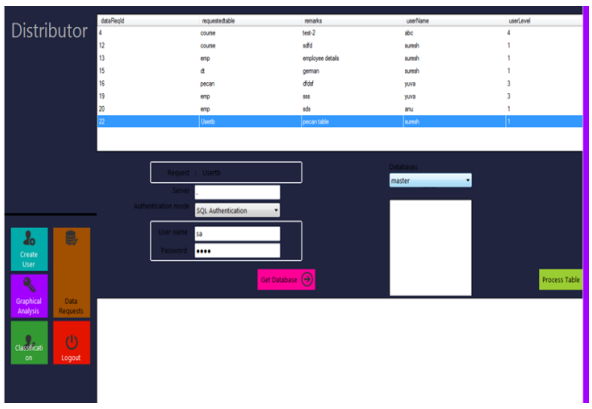

Fig -2: Dataset Preprocessing

- **Generalization Method**

First removes identifiers from the data and then partitions tuples into buckets. The two techniques differ in the next step. Generalization transforms the QI-values in each bucket into "less specific but semantically consistent" values so that tuples in the same bucket cannot be distinguished by their QI values. In order for generalization to be effective, records in the same bucket must be close to each other so that generalizing the records would not lose too much information.
Generalization is one of the commonly anonymized approaches, which replaces quasi-identifier values with values that are less-specific but semantically consistent.
Then, all quasi-identifier values in a group would be generalized to the entire group extent in the QID space. If at least two transactions in a group have distinct values in a certain column (i.e. one contains an item and the other does not), then all information about that item in the current group is lost. The QID used in this process includes all possible items in the log. Due to the high-dimensionality of the quasi-identifier, with the number of possible items in the order of thousands, it is likely that any generalization method would incur extremely high information loss, rendering the data useless.
In order for generalization to be effective, records in the same bucket must be close to each other so that generalizing the records would not lose too much information. However, in high-dimensional data, most data points have similar distances with each other.
To perform data analysis or data mining tasks on the generalized table, the data analyst has to make the uniform distribution assumption that every value in a generalized interval/set is equally possible, as no other distribution assumption can be justified.

This significantly reduces the data utility of the generalized data. And also, because each attribute is generalized separately, correlations between different attributes are lost. In order to study attribute correlations on the generalized table, the data analyst has to assume that every possible combination of attribute value is equally possible. This is an inherent problem of generalization that prevents effective analysis of attribute correlations.

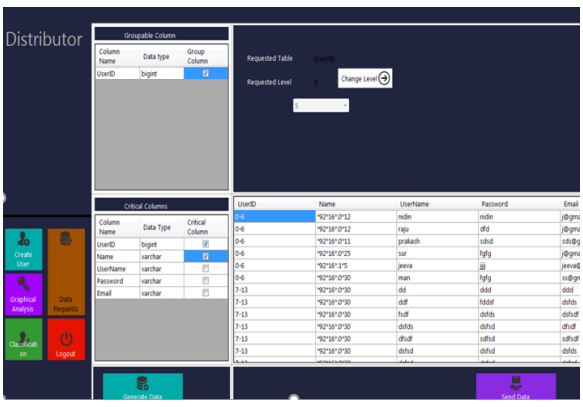

Fig -3: Generalization Method

- **Bucketization Method**

The effectiveness of slicing in membership disclosure protection. For this purpose, we count the number of fake tuples in the sliced data. We also compare the number of matching buckets for original tuples and that for fake tuples. The first, which we term bucketization, is to partition the tuples in T into buckets, and then to separate the sensitive attribute from the non-sensitive ones by randomly permuting the sensitive attribute values within each bucket. The sanitized data then consists of the buckets with permuted sensitive values.

In this we use bucketization as the method of constructing the published data from the original table T, although all our results hold for full-domain generalization as well. We now specify our notion of bucketization more formally. Partition the tuples into buckets (i.e., horizontally partition the table T according to some scheme), and within each bucket, we apply an independent random permutation to the column containing S-values. The resulting set of buckets, denoted by B, is then published.

For example, if the underlying table T, then the publisher might publish bucketization B. Of course, for added privacy, the publisher can completely mask the identifying attribute (Name) and may partially mask some of the other non-sensitive attributes (Age, Sex, Zip). For a bucket b ◌̈ B, we use the following notation. While bucketization has better data utility than generalization, it has several limitations. First, bucketization does not prevent membership disclosure. Because bucketization publishes the QI values in their original forms, an adversary can find out whether an individual has a record in the published data or not. Second, bucketization requires a clear separation between QIs and SAs. However, in many data sets, it is unclear which attributes are QIs and which are SAs. Third, by separating the sensitive attribute from the QI attributes, bucketization breaks the attribute correlations between the QIs and the SAs.

Bucketization first partitions tuples in the table into buckets and then separates the quasi identifiers with the sensitive attribute by randomly permuting the sensitive attribute values in each bucket.

The anonymized data consist of a set of buckets with permuted sensitive attribute values. In particular, bucketization has been used for anonymizing high dimensional data.

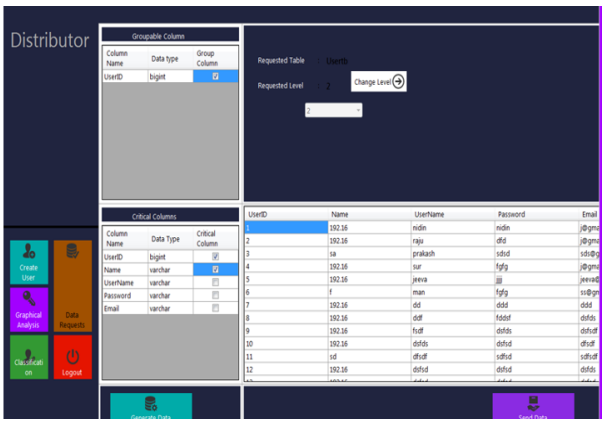

Fig -4: Bucketization Method

- **Slicing**

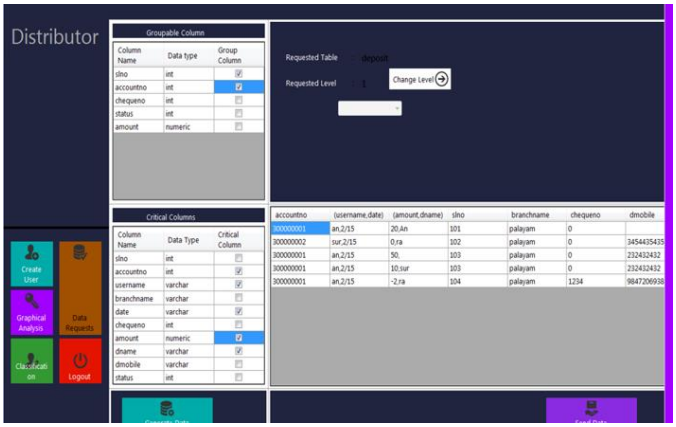

Fig -5: Slicing

6. CONCLUSIONS

The purpose of this study is to address the privacy issues associated with healthcare databases as a result of data mining technologies. We used a genetic optimization technique to hide negative sensitive association rules using a heuristic approach based on both distortion and restriction processes. The suggested solution is based on a strategy of concurrently decreasing the confidence in the sensitive rules. The technique makes the fewest possible modifications to the database and misses the fewest possible non-sensitive association rules, which is the ultimate goal of data sanitization. The proposed algorithm is a hybrid of the Apriori and integrated genetic-Tabu algorithms. Rather than mining negative association rules intuitively, the proposed methodology utilizes negative interestingness to describe and explain the success of negative association rules. By using a genetic-Tabu search method, the system lowers the mining process's search space. The main benefits of the algorithm are that (1) a simple heuristic method is used to choose the transactions and items to be cleaned; (2) a genetic algorithm is used to adjust the victim's choice of items; and (3) data availability is improved by hiding rules instead of items. The fitness function's efficiency has been evaluated in a variety of healthcare databases to determine whether it holds up when a variety of changes are made to the original database. From the simulation results, it is clear that the rules of the suggested technique have much higher support and confidence values while requiring much less processing time to reach the goals. Privacy-preserving data transformation techniques, log extensions, and process mining algorithms will all be examined in future work, as well as an empirical investigation of how these strategies affect healthcare logs. Furthermore, the suggested method will be tested on a variety of healthcare datasets, including those with varying features, to ensure that it is effective. For more privacy, negative association rules will be applied with differential privacy, local differential privacy, or a combination of both.

REFERENCES

1. H. Fatemidokht, M. K. Rafsanjani, B. B. Gupta, and C.-H. Hsu, ‘‘Efficient and secure routing protocol based on artificial intelligence algorithms with UAV-assisted for vehicular ad hoc networks in intelligent transportation systems,’’ IEEE Trans. Intell. Transp. Syst., vol. 22, no. 7, pp. 4757–4769, Jul. 2021

2. S. Huang, Z. Zeng, K. Ota, M. Dong, T. Wang, and N. N. Xiong, ‘‘An intelligent collaboration trust interconnections system for mobile information control in ubiquitous 5G networks,’’ IEEE Trans. Netw. Sci. Eng., vol. 8, no. 1, pp. 347–365, Jan. 2021.

3. van Leeuwen, M. Huang, A. Liu, N. N. Xiong, and J. Wu, ‘‘A UAV-assisted ubiquitous trust communication system in 5G and beyond networks,’’ IEEE J. Sel. Areas Commun., vol. 39, no. 11, pp. 3444–3458, Nov. 2021, doi: 10.1109/JSAC.2021.3088675.

4. C. Boudagdigue, A. Benslimane, A. Kobbane, and J. Liu, ‘‘Trust management in industrial Internet of Things,’’ IEEE Trans. Inf. Forensics Security, vol. 15, pp. 3667–3682, 2020.

5. M. Zhaofeng, W. Lingyun, W. Xiaochang, W. Zhen, and Z. Weizhe, ‘‘Blockchain-enabled decentralized trust management and secure usage control of IoT big data,’’ IEEE Internet Things J., vol. 7, no. 5, pp. 4000–4015, May 2020

6. S. Atiewi, A. Al-Rahayfeh, M. Almiani, S. Yussof, O. Alfandi, A. Abugabah, and Y. Jararweh, ‘‘Scalable and secure big data IoT system based on multifactor authentication and lightweight cryptography,’’ IEEE Access, vol. 8, pp. 113498–113511, 2020.

7. J. Shen, D. Liu, Q. Liu, X. Sun, and Y. Zhang, ‘‘Secure authentication in cloud big data with hierarchical attribute authorization structure,’’ IEEE Trans. Big Data, vol. 7, no. 4, pp. 668–677, Oct. 2021.

BIOGRAPHIES

Chithra S Prasad,currently pursuing MTech degree in Computer Science and Engineering from APJ Abdul Kalam Technological University, Kerala, India at Mount Zion College of Engineering, Kadamanitta, Kerala, India

Smitha C Thomas, received the MTech degree in Computer science and Engineering. She is currently working as Assistant Professor in the Department of Computer science and Engineering at Mount Zion College of Engineering, Kadamanitta, Kerala, India