A Novel Privacy Preservation Model For E-Healthcare System

Nidhi G Pandya[1],Ms. Gaytari S Pandi [2]
[1]Student,Department of Computer Engineering,L. J. I. E. T, Ahmedabad, Gujarat, India
[2]Head of Post Graduate Department ,L. J. I. E. T, Ahmedabad, Gujarat, India

Abstract

In this era of data digitization, data mining is essential for getting valuable information. However, privacy and security issues remain major barriers during this process. Since medical records are related to human subjects, privacy protection is taken more seriously than other data mining tasks. As required by the Health Insurance Portability and Accountability Act (HIPAA), it is necessary to protect the privacy of patients and ensure the security of the medical data.Variety of different approaches and algorithms of privacy preservation are available but they suffer from various types of attacks and information loss. So, for making privacy preservation techniques more strong the developers uses different PPDM techniques like, generalization, randomization, k-anonymity, cryptography, etc. In this report We are trying to describe the PPDM model for E-Healthcare data. we have proposed a model/technique that hide multiple sensitive attribute and try solve problem of high dimensionality.

Keywords— *Healthcare data, Privacy preserving, Sensitive data, Security, K-anonymity,MCAR*

.1.INTRODUCTION

Data mining is a knowledge discovery process of analyzing large databases to find useful patterns. Privacy preserving data mining is a broad research area for protecting sensitive data or knowledge. It has numerous Applications such as marketing, business, medical analysis, product control, engineering design, bioinformatics and scientific exploration, etc. The knowledge discovered from a database can be expressed in patterns such as decision trees, clusters or association rules.

Most of the privacy-preserving data mining techniques apply a transformation which reduces the usefulness of the underlying data when it is applied to data mining techniques or algorithms. In preserving privacy of data, the problem is how securely results are gained but not with data mining result but. As a simple example, suppose some hospitals want to get useful aggregated knowledge about a specific diagnosis from their patients' records while each hospital is not allowed, due to the privacy acts, to disclose individuals' private data. Therefore, they need to run a joint and secure protocol on their distributed database to reach to the desired information[1].

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].

A. BACKGROUND:- History of Privacy Preservation

As technology has advanced, the way in which privacy is protected and violated has changed with it. In the case of some technologies, such as the printing press or the Internet, the increased ability to share information can lead to new ways in which privacy can be breached. It is generally agreed that the first publication advocating privacy in the United States was the article by Samuel Warren and Louis Brandeis, "The Right to Privacy", 4 *Harvard Law Review* 193 (1890).New technologies can also create new ways to gather private information. However, in 2001 in s*Kyllo v. United States* (533 U.S. 27) it was decided that the use of thermal imaging devices that can reveal previously unknown information without a warrant does indeed constitute a violation of privacy.

B CLASSIFICATION OF PRIVACY PRESERVATION TECHNIQUES

1) **The randomization method:**

   The randomization method is a technique for privacy-preserving data mining in which noise is added to the data in order to mask the attribute values of records. The noise added is sufficiently large so that individual record values cannot be recovered. Therefore, techniques are designed to derive aggregate distributions from the perturbed records. Subsequently, data mining techniques can be developed in order to work with these aggregate distributions.

2) **The k-anonymity model and l-diversity:**

   The *k*-anonymity model was developed because of the possibility of indirect identification of records from public databases. This is because combinations of record attributes can be used to exactly identify individual records. In the *k*-anonymity method, we reduce the granularity of data representation with the use of techniques such as generalization and suppression. This granularity is reduced sufficiently that any given record maps onto at least *k* other records in the data. The *l*-diversity model was designed to handle some weaknesses in the *k*-anonymity model since protecting identities to the level of *k*-individuals is not the same as protecting the corresponding sensitive values, especially when there is homogeneity of sensitive values within a group. To do so, the concept of intra-group diversity of sensitive values is promoted within the anonymization scheme.

3) **Distributed privacy preservation:**

   In many cases, individual entities may wish to derive *aggregate results* from data sets which are partitioned across these entities. Such partitioning may be horizontal (when the records are distributed across multiple entities) or vertical (when the attributes are distributed across multiple entities). While the individual entities may not desire to share their entire data sets, they may consent to limited information sharing with the use of a variety of protocols.

4) **Downgrading Application Effectiveness:**

   In many cases, even though the data may not be available, the output of applications such as association rule mining, classification or query processing may result in violations of privacy. This has lead to research in downgrading the effectiveness of applications by either data or application modifications. Some examples of such techniques include association rule hiding, classifier downgrading, and query auditing.

2. RELATED WORK

In An Efficient Approach For Privacy Preserving in Data Mining, Manish Shannal et al. proposed an efficient approach for privacy preservation in data mining. This technique protects the sensitive data with less information loss which increase data usability and also prevent the sensitive data for various types of attack. Data can also be reconstructed using our proposed technique.In this proposed method, first we apply

randomization on original data and then after randomization we categorize the sensitive attribute values into high sensitive and low sensitive class. Secondly we apply k-anonymization on those tuples who belongs to high sensitive class and those tuples who belongs to low sensitive remain as it is. So it reduces the information loss and improves the usability of data. The combination of anonymization with randomization technique is made difficult for the attacker to attack on database[1]

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.[2]

In Privacy Preserving in Data Mining Using Hybrid Approach**,** Savita Lohiya et.al. focus on Data sharing that done between two organizations is common in many application areas like business planning or marketing. When data are to be shared between parties, there could be some sensitive data which should not be disclosed to the other parties. Also medical records are more sensitive so, privacy protection is taken more seriously. As per requirement by the Health Insurance Portability and Accountability Act (HIPAA), it is necessary to protect the privacy of patients and ensure the security of the medical data. we propose a method called Hybrid approach for privacy preserving. First we randomizing the original data. Then we apply generalization on randomized or modified data. This technique protect private data with better accuracy, also it can reconstruct original data and provide data with no information loss, makes usability of data[4] .

In Task Independent Privacy Preserving Data Mining on Medical Dataset , E. Poovammal and M. Ponnavaikko et.al. K anonymization algorithm works only on the QI attribute and its improved methods such as L-diversity and t-closeness works on sensitive attribute in the QI group. But our technique transforms only the sensitive attribute(s) and the transformed table can be published for any type of mining task. Any data mining algorithm without any modification can be applied to the transformed table and the accuracy of results/patterns/rules will be as good as the original table. Our transformation procedure is based on the data type of sensitive data. If it is numerical data type, transformation is performed by categorical membership values and if categorical by mapping values[5].

3. PROBLEM DESCRIPTION

K-anonymity only prevents association between individuals and tuples instead of association between individuals and their sensitive values. Since, this method places no constraint on the sensitive values, it may result in homogeneity attack. In result of K-anonymity and Randomization an adversary who knows Quasi Identifier values can guess the actual disease of patient with 50% probability. It is possible that dataset has more than one sensitive attribute, and attacker can leak record by using any of them sensitive attribute value knows. But patient may be willing to disclose his actual disease instead of being linked with other disease. So we proposed K- anonymity and Multiple Sensitive Attribute hiding approach.

4. PROPOSED WORK

4.1 Proposed System Flow-Diagram

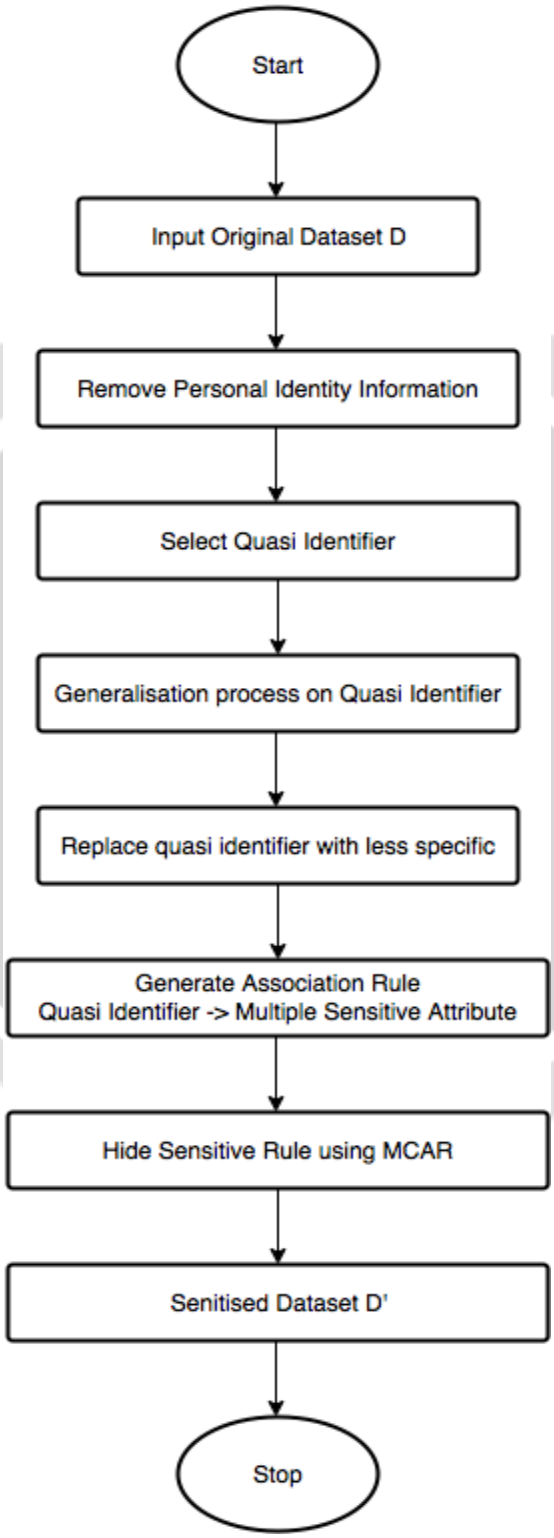

Figure 1: Proposed System Flow Diagram

Proposed solution steps describe as below:

1) First step is start the process and get patient's data as input
2) After that in next step remove personal identity information of patient
3) Now, select Quasi Identifier based on certain parameters, like (age, language, gender) for patient.
4) Apply Generalization process on selected Quasi Identifier.
5) Now, replace that Quasi Identifier with less specific.
6) Generate association rule for that quasi-Identifier using Multiple Sensitive Attribute (Quasi Identifier → Multiple sensitive Attribute).
7) Now, Hide sensitive rule using Multi Class Association Rule which has confidence >= Minimum Confidence Threshold.
8) Finally system generates the sanitized Datasets.

5. Experimental Result and Analysis

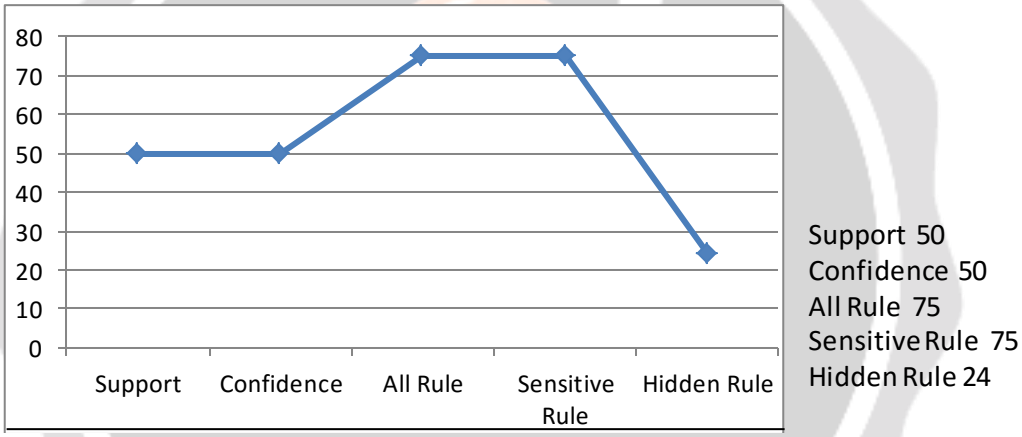

Chart 1 :- Sensitive Rule Hiding- Confidence 50 %

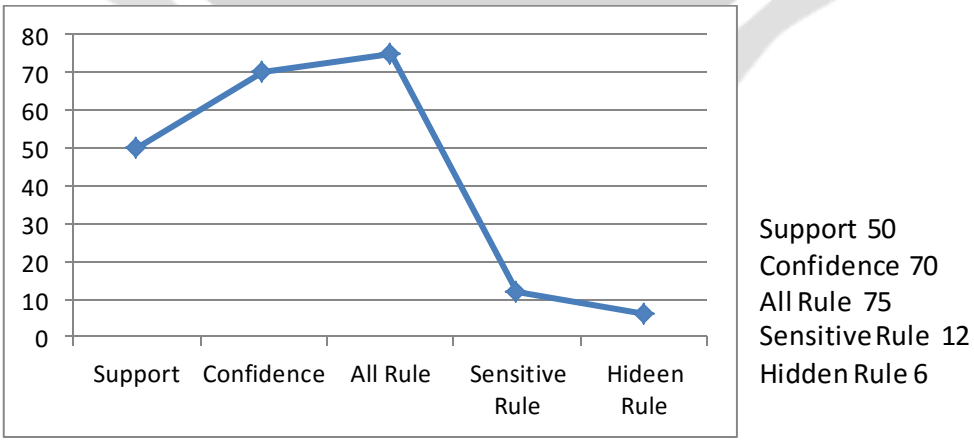

Chart 2 : Sensitive Rule Hiding- Confidence 75 %

6. CONCLUSION

One conclusion I can draw from this literature review that, currently privacy preserving in data mining is hot topic of research. Literature review clears that there are many privacy preserving techniques available but still they have shortcomings. Anonymity technique gives privacy protection of data but it suffers from homogeneity and background attack. The proposed approach employs K-anonymity and MCAR (Multi Class Association Rule) hiding. By using K-Anonymity and Multi Class Association Rule combined approach, it is possible to hide multiple sensitive attribute and provide more privacy to patient's sensitive data.

REFERENCES

1) Manish Sharma, Atul Chaudhary, Manish Mathuria, Shalini Chaudhary, Santosh Kumar,"An Efficient Approach For Privacy Preserving in Data Mining" IEEE International Conference on Signal Propagation and Computer Technology (ICSPCT) on, DOI: 978-1-4799-3140-8/14pp.244-249,IEEE AUG-2014.
2) Morgan Price, Jens H. Weber, and Glen McCallum,"SCOOP – The social Collaboratory for Outcome Oriented Primary Care", IEEE International Conference on Healthcare Informatics on, DOI:978-4799-5701-9/14pp.-210-215, IEEE SEP-2014.
3) Georgios Kalogridis and Saraansh Dave,"Privacy and eHealth-enabled Smart meter Informatics",IEEE HEALTHCOM 1st International Workshop on Secure and Privacy-Aware Information Management in eHealth on,DOI:978-1-4799- 6644-8/14pp.-116-121,IEEE AUG-2014.
4) Savita Lohiya, Lata Ragha,"Privacy Preserving in Data Mining Using Hybrid Approach",IEEE 4th International Conferences on Computational Intelligence and Communication Networks on,DOI:978-0-7695-4859-0/12pp.743-746,IEEE JAN- 2012.
5) E. Poovammal and M. Ponnavaikko ," Task Independent Privacy Preserving Data Mining on Medical Dataset", International Conference on Advances in Computing, Control, and Telecommunication Technologies on,DOI:978-0-7695- 3915-7/09pp.-814-818,IEEE July-2009.
6) M. Prakash, Dr. G. Singaravel,"A New Model for Privacy Preserving Sensitive Data Mining",IEEE 26-28 July,2012.
7) Hsiang-Cheh Huang and Wai-Chi Fang"Integrity Preservation and Privacy Protection for Medical Images with Histogram-Based reversible Data Hiding",IEEE/NIH Life Science Systems and Applications Workshop on,DOI:978-1-4577-0422-2-11pp.108-111,IEEE FEB-2011.
8) Jian Wang, Yong Cheng Lou, Yen Zha Jiajin Le, "A Survey on Privacy Preserving Data Mining", International Workshop on Database Technology and Application pp.111 -114, 2009.