International Journal of Scientific Research and Engineering Development-– Volume X Issue X, Year

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Privacy Preserving technology of data mining- A Review

Seema Kongali*, Firozsab Nadaf**,Mr.M.S.Emmi***

*(Department of Master of Computer Applications, VTU/KLS Gogte Institute of Technology, Belagavi
Email: seemakongali@gmail.com)

**(Department of Master of Computer Applications, VTU/KLS Gogte Institute of Technology, Belagavi
Email: firozsnadaf@gmail.com)

***(Department of Master of Computer Applications, VTU/KLS Gogte Institute of Technology, Belagavi
Email: msemmi@git.edu)

----------**********----------

Abstract:

The academic community focused on safety and knowledge discovery recently established a fresh category of data mining techniques called as Privacy preserving data mining. The goal of PPDM is to secure sensitive information while simultaneously extracting pertinent insights from massive amounts of data. We describe a general PPDM framework and the frequencies of the various approaches utilized in this paper. Additionally, a set of measures and a conceptual framework for comparing the effectiveness of particular PPDM algorithms are included.

Keywords — **Secure Multiparty Computation(SMC), privacy-preserving data mining, and data mining.**

----------**********----------

I. INTRODUCTION

Extraction or mining of knowledge from vast volumes of data is known as data mining. The practice of extracting valuable knowledge from vast amounts of data that have been stored in data warehouses, databases, or other information repositories is known as data mining. It is an area of study that combines methods from various academic fields, including statistics, machine learning, recognition of patterns, database as well as data warehouse technology, information retrieval, and positional or temporal analysis of data. Data mining is particularly prone to abuse because of its promise to quickly uncover useful, obscure information from massive databases. As a result, data mining and privacy may collide.

Data mining used to extract private information is referred to as privacy [9]. On the other hand, the sensitive and private data that is stored in massive and dispersed data warehouses is at risk due to the overwhelming processing capability of intelligent algorithms. The gathering and processing of enormous amounts of personal data, including criminal histories, shopping patterns, credit and medical histories, and driving records, has been made possible by recent advancements in information technology. for a variety of purposes, such as national security, law enforcement, and medical research. The public is becoming more concerned about people's privacy, nevertheless. The right to control data Undoubtedly, such data is very beneficial regarding oneself is frequently understood

as constituting one's right to privacy. Issues with general privacy are secondary uses of the personal data.

Data mining with a focus on privacy means protecting individual data from data mining technologies. using network and database technology. Data mining also results in the inevitable leaking of privacy because it operates straight on the beginning of data collection. Therefore, the primary study area for privacy-preserving data mining is how to prevent the leakage of sensitive or private knowledge while still getting reliable findings from data mining.

II. THE PRIVACY PRESERVING GOALS.

The objective of privacy preservation is to extract the raw data affecting privacy without leaking it. These two factors are the main sources of current technology realization[2].

1) To prevent the leak of private information, critical raw data in databases, such as labels, certificate numbers, addresses, and hobbies, can be changed or removed. In other words, accurate results can be obtained by applying data mining techniques without touching private data.

2) Rule algorithms can be used to exclude sensitive rules from data mining findings. In other words, aim to prevent someone with bad intentions from obtaining potentially sensitive rules in the mining process.

III. PRIVACY PRESERVING IN CLUSTERING.

It is a difficult task to protect people's privacy when data are exchanged for clustering. The tricky part is figuring out how to keep clustering from revealing the similarities between the items under investigation while still protecting the underlying data values[9].

A series of geometric data transformation algorithms (GDTMs) that affect numerical properties by scaling, rotations, translations, or by combining all of the aforementioned transformations were revisited by Stanley et al. This approach is meant to define privacy-preserving clustering in situations where data owners must ensure both the validity of the clustering findings and adhere to privacy standards. The authors also provide a detailed, comprehensive, and sophisticated overview of techniques for security clustered by data transformation[9].

A. Advantages:

The entire document should be in Times New Roman or Times font. Type 3 fonts must not be used. Other font types may be used if needed for special purposes.

Recommended font sizes are shown in Table 1.

1. Geometric Dimensional transformation of data methods (GDTMs) that alter private numerical attributes in clustering analysis to adhere to privacy protection.
2. Consumers are able to make use of their own tools, hence the privacy constraint must be implemented before the data mining via data transformation process.

B. Disadvantages:

1.Individuals' data must be shared because clustering is too sophisticated to safeguard the data values from grouping similarity between the objects under investigation.

Clustering with privacy protection for centralized data To adjust sensitive data of individuals without impacting clustering, Oliveira employed geometric modification methods based on translating data change, scaling data perturbation, and rotation data perturbation.

In Agarwal proposed When data is provided for data mining analysis, personal information about individuals should be kept private.

Varki's provides a list of numerous data distortion techniques for keeping the crucial data in data analysis while protecting privacy in data mining.

To get a low-rank approximation of the original dataset, Zhang et al.'s suggested matrix decomposition algorithms, which were tested on

several datasets. A hybrid method for privacy-preserving clustering that is based on a fuzzy approach and randomized rotate perturbation for centralized databases was reported by Naga Lakshmi.

IV. PPDM FRAMEWORK

Fig. 1 depicts the PPDM framework. Data (mainly transactional) is gathered by one or more organizations and saved in the appropriate databases during the data mine or discovering knowledge from databases (KDD) process. The data is subsequently translated into an analytically-friendly format, stored in sizable data warehouses, and subjected to data mining algorithms to provide information or knowledge. The model must be improved with the goal of safeguarding privacy. Privacy restrictions cannot be implemented in stages; they must be considered throughout the entire data mining process, from the collecting of data through the creation of knowledge. The graphic below shows three stages where privacy issues are addressed[4].

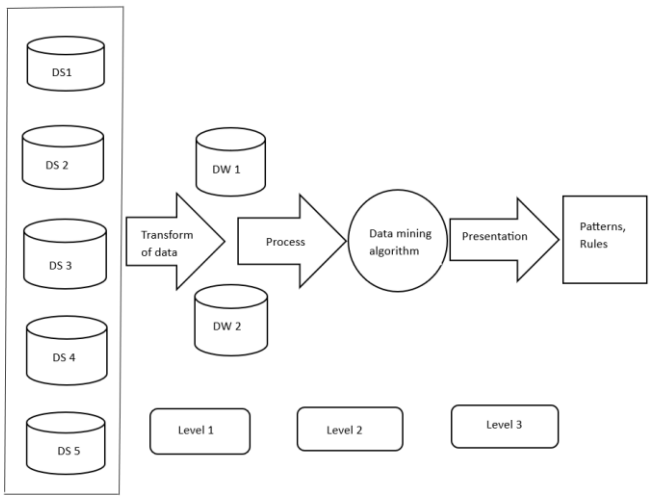

Fig 1. Framework of PPDM

At level 2, several techniques are used to clean up information from data warehouses so that it can still be revealed to shady data miners. At this level, techniques like blocking, suppression, perturbation, modification, generalization, and sampling are used. Then, for the purpose of knowledge/information discovery, the data processing techniques are applied to the processed data. Even the information mining processes have been altered to maintain privacy while still achieving the objectives of data mining. At stage 3, the data mining algorithms' newly revealed knowledge is examined for its sensitivity to potential disclosure.

V.PPDM TECHNIQUES

PPDM has been the subject of substantial investigation in recent years. Privacy-preserving data mining has drawn a lot of attention as a field of study in data mining and statistical databases, and numerous studies have been conducted there.
The vast majority of current strategies can be divided into two groups[4].
Methodologies that

1. safeguard sensitive data throughout the mining process and
2. safeguard sensitive data mining outputs (i.e., knowledge that was derived during the mining process) are both important.

The list of the five criteria that can be used to categorize various PPDM Techniques is as follows:

a. Data distribution,
b. Data manipulation,
c. Data mining methods, and finally,
d. Data exploitation Hiding rules or data
e. Protecting privacy

Different PPDM approaches can be grouped into the five groups below on the basis of these dimensions.
1. PPDM based on anonymization
2. PPDM based on perturbation
3. PPDM based on Randomized Response
4.PPDM based on the condensation technique
5. PPDM based on cryptography
In the subsections that follow, we go into further depth on each of these.

A. Anonymization based PPDM

The following categories of attributes make up the fundamental structure for information in a table.

- Explicit Identifiers are a group of attributes that contain data that specifically identifies the record owner, such as name, social security number, etc.
- Sensitive traits are a group of traits that include private information about a specific person, like a condition or pay.
- Non-Sensitive traits are a group of traits that pose no issues even if they are revealed to unreliable parties.

Anonymization is a strategy used to conceal the sensitive information of record owners. Even the retention of sensitive data for analysis is presumed.

B. Perturbation based PPDM

It's intrinsic simplicity, effectiveness, and capacity to preserve statistical information, perturbation has a long history and is frequently employed in statistical disclosure control. To ensure that the statistical information derived from the perturbed data does not differ significantly from the statistic information derived from the original data, the original values are substituted for certain synthetic data values during perturbation. Since the records in the perturbed data do not match actual record owners, the attacker is unable to execute sensitive linkages or extract private information from the published data.

C. Randomized Response based PPDM

In essence, Warner presented randomized response as a statistical technique to address a survey problem. In Random respond, the data is jumbled so that the central location cannot determine if the data from a client contains true information or fraudulent information with probability above a pre-defined threshold. The information gathered from each individual user is jumbled, but if there are a lot of them, the combined data of all of them may be approximated very accurately.

The data providers randomize their data in the first stage, then send the randomised data to the data receiver.

D. Condensation approach based PPDM

Constrained clusters are created in the dataset using the condensation approach, and the statistical results of these clusters are then used to produce pseudo data. Because of its method of producing fake data by exploiting condensed statistics of clusters, it is known as condensation. It creates non-homogeneous size groups from the data, ensuring that each record is contained inside a collection whose size is at least equal to its level of anonymity. In order to build a synthetic data collection with the same aggregated distribution as the actual data, pseudo data is then generated from each group.

E. Cryptography based PPDM

Think about a situation where several medical institutions want to perform a cooperative study for certain mutual benefits without disclosing superfluous information. In this case, research on symptoms, diagnoses, and treatments based on different criteria needs to be done while simultaneously protecting people's privacy. These kinds of situations are known as distributed computing scenarios. Considering that the people participating in mining such tasks may be rivals or other parties with whom they have little confidence, privacy protection becomes a top priority. In situations where numerous parties work together to compute results or exchange non-sensitive mining findings, cryptographic techniques are useful since they prevent the publication of sensitive information.

VI. CONCLUSIONS

Some people use the original data to mine the database owner's privacy information patterns. The advantage of the database owner has been harmed. Before releasing the database, we should hide the private or sensitive data pattern of the database owner, includes the sensitive associations rule

information, in order to solve the problem of associate rule privacy preservation. When distributing data to mine helpful knowledge and information, the privacy disclosure problem about a person or business is inexorably exposed with the development of the processing and analysis of data technique, giving rise to the research field on privacy-preserving data mining. Recently, a number of techniques have been put out for mining multidimensional data records while protecting privacy.

VII. REFERENCE

1. Privacy Preserving Attribute Reduction Based on Rough Set https://ieeexplore.ieee.org/abstract/document/4771913/
2. Research on Privacy-Preserving Technology of Data Mining https://ieeexplore.ieee.org/abstract/document/5287756/
3. A Survey on Privacy Preserving Data Mining https://ieeexplore.ieee.org/abstract/document/5207803/
4. Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects https://ieeexplore.ieee.org/abstract/document/6394662/
5. Privacy 3arallel Algorithm for Mining Association Rules and Its Application in HRM https://ieeexplore.ieee.org/abstract/document/5368463/
6. A Fuzzy based Data Perturbation Technique for Privacy Preserved Data Mining https://ieeexplore.ieee.org/abstract/document/9077826/
7. A Privacy Preserving Jaccard Similarity Function for Mining Encrypted Data https://ieeexplore.ieee.org/abstract/document/5395869/
8. Three New Approaches to Privacy-preserving Add to Multiply Protocol and Its Application https://ieeexplore.ieee.org/abstract/document/4771997/
9. A Survey on Privacy Preserving Data Mining https://ieeexplore.ieee.org/abstract/document/7124885/
10. Study of Privacy Preserving Data Mining https://ieeexplore.ieee.org/abstract/document/5453697/

