PRIVACY PRESERVING IN DATA MINING

B.R.Sanjita[1] , Nipunika.A[2], Rohita Desai[3]
Sreenidhi Institute of Technology and Science, Medchal, Telangana.

Abstract : The collection and analysis of data are continuously growing due to the continuous introduction of data from various sources. This huge information has led to the advent of Big Data, Data Science and various other concepts. The analysis of such information is fostering businesses and contributing beneficially to the society in many different fields. However, this storage and flow of possibly sensitive data poses serious privacy concerns. Methods that allow the knowledge extraction from data, while preserving privacy, are known as privacy-preserving data mining (PPDM) techniques. Data mining produces a large amount of data that needs to be analyzed in order to extract useful information from it and gain knowledge. This data is vulnerable to data hackers and rough employees to take advantage of the situation and misuse the data. The privacy preservation is an important concern in data mining as secrecy of sensitive information must be maintained while sharing the data among different un-trusted parties. Privacy preserving data mining (PPDM) protects the privacy of sensitive data without losing the usability of the data. Various techniques have been introduced under PPDM to achieve this goal. This study describes about various techniques of privacy preserving in data mining. It also analyzes their advantages and limitations and comes up with a conclusion that a single technique does not exceed all the parameters such as performance, data utility, level of uncertainty, resistance to data mining algorithms and complexity.

IndexTerms **- Data mining, Privacy Preserving, Big Data.**

I. INTRODUCTION

Data mining is the process of extracting patterns (knowledge) from big data sets that can then be represented and interpreted. A pattern is defined as an expression to describe a subset of data (itemset), or a model applicable to a subset. Since data mining methods involve pattern discovery and extraction, pattern recognition techniques are often used.

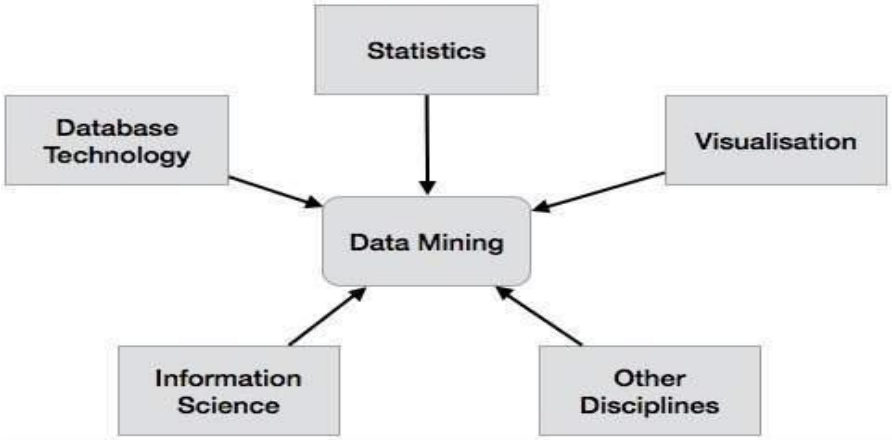

Fig 1: Data mining

Data collection and data mining techniques are applied to several application domains. Some of these domains require handling, and often publishing sensitive personal data (e.g. medical records in health care services), which raises the concern about the disclosure of private information.

Privacy means to have an appropriate usage. Extraction of the data must be privacy preserved in order to maintain the integrity and confidentiality of data. Data mining is the extraction of vast interesting patterns or knowledge from huge amount of data. The initial idea of privacy-preserving data mining PPDM was to extend traditional data mining techniques to work with the data modified to mask sensitive information. The key issues were how to modify the data and how to recover the data mining result from the modified data. Privacy-preserving data mining considers the problem of running data mining algorithms on confidential data that is not supposed to be revealed even to the party running the algorithm (Abou-el-ela Abdou Hussien, 2013).

II. PROBLEM STATEMENT

With the proliferation of devices connected to the Internet and connected to each other, the volume of data collected, stored, and processed is increasing everyday, which also brings new challenges in terms of the information security. In fact, the currently used security mechanisms such as firewalls and DMZs cannot be used in the Big Data infrastructure because the security mechanisms should be stretched out of the perimeter of the organization's network to satisfy the user/data mobility requirements and the policies of BYOD (Bring Your Own Device). Considering these new scenarios, the pertinent question is what security and privacy policies and technologies are more adequate to satisfy the current top Big Data privacy and security demands (Cloud Security Alliance, 2013). These challenges may be organized into four Big Data aspects such as infrastructure security (e.g. secure distributed computations using Map Reduce), data privacy (e.g. data mining that preserves privacy/granular access), data management (e.g. secure data provenance and storage) and, integrity and reactive security (e.g. real time monitoring of anomalies and attacks).

Each of these aspects faces the following security challenges, according to CSA:

- Infrastructure Security
    1. Secure Distributed Processing of Data

2. Security Best Actions for Non-Relational Data-Bases

- Data Privacy
1. Data Analysis through Data Mining Preserving Data Privacy
2. Cryptographic Solutions for Data Security
3. Granular Access Control
- Data Management and Integrity
1. Secure Data Storage and Transaction Logs
2. Granular Audits
3. Data Provenance
- Reactive Security
1. End-to-End Filtering & Validation
2. Supervising the Security Level in Real-Time

Solution

While trying to take the most of Big Data, in terms of security and privacy, it becomes mandatory that mechanisms that address legal requirements about data handling, need to be met. Secure encryption technology must be employed to protect all the confidential data Personally Identifiable Information (PII), Protected Health Information (PHI) and Intellectual Property (IP) and careful cryptographic material (keys) access management policies, need to be put in place, to ensure the correct locking and unlocking of data – this is particularly important for data stored. In order to be successful these mechanisms need to be transparent to the end-user and have low impact of the performance and scalability of data. To achieve these needs, we have various techniques involved in PPDM. PPDM algorithms or techniques proposed are various ways through which privacy can be produced. All techniques have their own uniqueness. The algorithms may not provide privacy in all aspects or all possible attacks but can be used for certain attacks.

III. DESIGN AND METHODOLOGY

Existing techniques mainly focus on preserving private information in different stages of a data mining process. In this paper, we review main PPDM techniques according to a PPDM framework that has three layers: Data Collection Layer (DCL), Data Pre-Process Layer (DPL) and Data Mining Layer (DML). The framework was constructed according to the stages in the data mining process, from data collection, pre-process, to final data mining procedure. The PPDM framework contains three layers: Data Collection Layer (DCL), Data Pre-Process Layer (DPL) and Data Mining Layer (DML), the first layer DCL contains a huge number of data providers that provide original raw data that could contain some sensitive information. The privacy-preserving data collection can be carried out during the data collection time. All the data collected from the data providers will be stored and processed in the data warehouse servers in DPL.

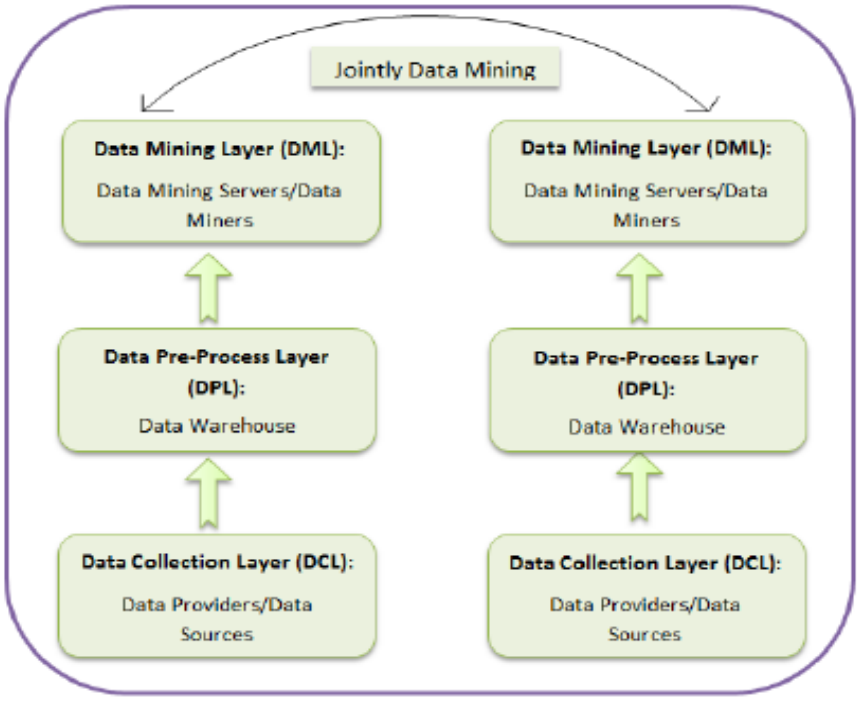

Fig. 2: A PPDM Framework

The second layer DPL contains data warehouse servers that are responsible for storing and pre-processing the collected raw data from the data providers. The raw data stored in the data warehouse servers could be aggregated in sum, average etc., or pre-computed using privacy-reserving methods in order to make the data aggregation or fusion process more efficient. The privacy preservation in this layer concerns two aspects. One is privacy-preserving data pre-processing for later data mining, and the other is the security of data access. In this paper, we focus on the privacy reserving data pre-computing methods, whereas the secure data access control is beyond the scope of this paper. The third layer DML consists of data mining servers and/or data miners located mostly in the Internet for conducting actual data mining and providing mining results. In this layer, privacy preservation concerns two aspects. One is to improve or optimizing data mining methods to enable privacy-preserving features. The other is collaborative data mining based on the union of a number of data sets owned by different parties without revealing any private information.

I. PPDM ATTACKS

Various attacks are addressed from a privacy-preserving perspective. In the following subsections the most common attacks are discussed.

i. Background Knowledge Attack

Recently, Xiao and Tao, two authors introduced Anatomy as an alternative anonymisation technique to generalization. Anatomy releases all the quasi-identifier and sensitive data directly into two separate tables. The original table is decomposed into two tables, the quasi-identifier table (QIT) and the sensitive table (ST). The QIT table and the ST table are then released. The authors also proposed an anatomizing algorithm to compute the anatomized tables. The algorithm first hashes the records into buckets based on the sensitive attribute, i.e., records with the same sensitive values are in the same bucket. Then the algorithm iteratively obtains the buckets that currently have the largest number of records and selects one record from each of the buckets to form a group. Each remaining record is then assigned to an existing group.

ii. Unsorted Matching Attack

This attack is based on the order in which tuples appear in the released table. While we have maintained the use of a relational model, and so the order of tuples cannot be assumed, in real-world use this is often a problem. It can be corrected of course, by randomly sorting the tuples of the solution. Otherwise, the release of a related table can leak sensitive information.

Solution: Random shuffling of rows.

iii. Complementary Release Attack

It is more common that the attributes that constitute the quasi-identifier are themselves a subset of the attributes released. As a result, when a k-minimal solution, which we will call table T is released, it should be considered as joining other external information. Therefore, subsequent releases of generalizations of the same privately held information must consider all of the released attributes of T a quasi-identifier to prohibit linking on T, unless of course, subsequent releases are themselves generalizations of T.

Solution:

1) Other data holders may release some data that can be used in this kind of attack. Generally, this kind of attack is hard to be prohibited completely.

iv. Temporal Attack

Data collections are dynamic. Tuples are added, changed, and removed constantly. As a result, releases of generalized data over time can be subject to a temporal inference attack.

Solution: Subsequent releases must use the already re- leased table.

v. Homogeneity Attack and Background Knowledge Attack

In this subsection we present two major attacks, the homogeneity attack and background knowledge attack], along with unsorted matching attack, complementary release attack and temporal attack, and we show that how they can be used to compromise a k-anonymous dataset (Abou-el-ela Abdou Hussien, 2013).

II. PPDM TECHNIQUES:

Multiple techniques have been developed to deal with privacy concerns in data mining, while attempting to preserve data utility. The goal of PPDM is to lower the risk of misuse of sensitive data and produce the same result as that produced in the absence of such privacy preserving techniques**.** Privacy preserving data mining proposes a few techniques to perform the data mining tasks in a privacy-preserving way. These techniques generally fall into the following categories of data modification techniques, cryptographic methods, randomization and perturbation-based techniques.

Based on the dimensions, PPDM techniques are of five types:

1. Anonymization based PPDM
2. Perturbation based PPDM
3. Randomized Response based PPDM
4. Cryptography based PPDM
5. Condensation based PPDM

Let's look at each technique in detail.

1. Anonymization based PPDM

Anonymization or data anonymization is a technique to remove or encrypt personal or sensitive information from a given data so that the person whom the data refers to remain anonymous. Therefore, anonymization based PPDM is an approach where identity or sensitive information about a person is hidden. In anonymization technique the explicit identifiers i.e. the identifiers which give sensitive and personal information about the record owner should be hidden or removed. But still there can be a risk of privacy intrusion when quasi identifier is linked to publicly available data. At certain times the data is required to be published in its original form publicly. The data may not be encrypted and perturbed, but still some sort of precaution should be taken before releasing the data in terms of anonymization. This is a kind of generalization of some attributes that protects against identity disclosure. anonymization can be achieved by methods like generalization, suppression, data removal, permutation, swapping etc. K-anonymity method is treated as the conventional anonymization method and many studies are based on k-anonymity. Quasi-Identifier is a combination of person specific sensitive attribute (say for example, age, disease and pin-code for census data). The author has proved that the removal of the quasi-identifier from dataset do not ensure data protection, still k – anonymity method is better choice for publishing data. A simple approach is to generalize fields which are part of quasi identifier. Anonymizing quasi-identifiers and sensitive attributes in datasets pose an information loss which is not desirable for mining. The authors of focus on

medical datasets and try to address the issues related to privacy requirements. Anonymization methods are also useful for addressing specific problems.
Condensation is effectively used for solving the classification problem.

Disadvantages:
This technique is not immune to two types of attacks namely:
(i) Homogeneity attack: An attack where all sensitive values are present in a single record. Hence it becomes very easy for the attacker to predict the sensitive values.
(ii) Background knowledge attack: In this attack, the attacker knows the background of the victim or has some sensitive data about the victim.

2. Perturbation techniques

Perturbation techniques employ a mechanism to distort data prior to data mining. A perturbed copy can be locally created by the individual contributor by adding noise. Once the local perturbed copy is generated the miner can reconstruct the perturbed version to obtain the original data distribution. The authors have tried to add Gaussian noise to generate perturbed version of dataset for decision tree classification. In same lines, authors have proposed an individually adaptable perturbation model. A multilevel privacy can be specified by the users. This opens a new venture in field of privacy preserving – Multi-level Trust PPDM (MLT PPDM). Based on the privacy settings a contributor specifies, the perturbed version of dataset will be generated. The authors have successfully proved with experiments the correctness of their approach for satisfying personal privacy. Another work offers the flexibility to the data owners to generate perturbed copies for arbitrary trust levels on demands. Perturbation methods can be classified into probability distribution category and fixed data perturbation.
The data distortion techniques like addition of noise, from some known distribution, randomization and condensation are applied. Perturbation methods are well suited in both central commodity-based computing as well distributed scenarios. A different type of perturbation called Geometric Data Perturbation (GDP) is based on service-oriented framework. With large number of users, aggregate information can be estimated with accuracy. This information can be used for decision-tree classification as the latter is based on aggregate values of a dataset.

3. Randomized Response based PPDM

Randomization has emerged as a useful technique for data disguising in privacy-preserving data mining. Randomized response method was first developed by S. L. Warner in 1965 and later modified by B. G. Greenberg in 1969. It's basically a research method which was used in survey interviews and it allowed respondents to respond to sensitive issues such as criminal behavior while maintaining confidentiality. In the randomized response technique, the data is scrambled either by adding noise or some random data to the original data such that central place cannot tell whether the data from a customer contain truthful information or false information. The information received from each user is scrambled. If the number of users is large than the aggregate information received from these users can be estimated accurately. Hence this technique can be useful for decision-tree classification as decision-tree classification is based on aggregate values. Randomized response model consists of two steps for the process of collecting data:
Step 1: The data providers randomize their data and transmit the randomized data to the data receiver.
Step 2: The data receiver reconstructs the original distribution of the data by using a distribution reconstruction algorithm.
Randomized response technique is very simple and does not require any knowledge of distribution of data in records. Hence this technique can be implemented during the time of data collection. It also does not require a trusted server to contain all the original records to perform anonymization process.

One weakness of randomized response based PPDM is that it treats all records equal irrespective of their local density. This leads to a problem where the records on the outer side are more susceptible to attacks as compared to records in more dense regions [3]. One solution to this could be by adding more and aggressive noise to the data but this will reduce the utility of the data. There are many data mining algorithms which were proposed under the randomized response technique.

4. Cryptography Based

If the parties distributed across multiple sites are legally prohibited from sharing their datasets, a mining model to be built must be able to maintain the privacy of contributing parties. Authors have discussed the efficiency and have demonstrated their relevance for PPDM. Examples to demonstrate secure sum computation of data mining algorithms are also discussed. Previous categories of PPDM allow disclose of data beyond the control of the data collection. Authors have addressed the problem of reconstructing missing values by building a data model where the parties are distributed and data is horizontally partitioned. A cryptographic protocol based on decision-tree classification is described by them. The implementations of these protocols consist of computationally intensive operations and generally consist of hard wired circuits. Secure Multiparty Computation is a technique in which computations are done beforehand on the basis of certain rules in statistical disclosure limitation. Basically there are three broad types of techniques under SMC: homomorphic encryption, circuit evaluation and secret sharing scheme. Both semi-honest and malicious adversaries are addressed by SMC protocols. Directly the information pertaining to personal identification is not removed from the dataset, but a pseudonym is generated and replaced. This information cannot be retrieved without compromising a secret shared previously.

5. Condensation based PPDM

This approach is called condensation approach as it uses condensed statistics of the clusters to generate pseudo data. It constructs group of non-homogenous size from the data, such that it is guaranteed that each record is present in a group whose size is equal to its anonymity level. Eventually, pseudo data is generated from each group, to create a synthetic data set which is same as original data. This approach can be effectively used for the problem of classification.
One the major advantage of this approach is that it provides better privacy protection as compared to other PPDM techniques as it uses pseudo data instead of original data. The use of pseudo data also provides an additional layer of privacy as it becomes

difficult to perform attacks on pseudo data. Because of the use of pseudo data, this approach works without redesigning the data mining algorithms as pseudo data has the same format as that of original data. But along with this advantage, it has a big disadvantage of information loss because of condensation of large number of records into single statistical cluster. The data mining results also get affected because of this information loss.

ADVANTAGES OF PPDM

1. PPDM is very advantageous in development of various data mining techniques.
2. It allows sharing of large amount of privacy sensitive dat for analysis purposes.
3. It can track and collect large amounts of data with the use of current hardware technology.
4. Development of KDD

DISADVANTAGES OF PPDM

1. One of the major problems of privacy preserving data mining is the abundant availability of personal data.
2. Many technologies exist for supporting proper data handling, but much work remains, and some barriers must be overcome in order for them to be deployed.
3. The above discussed privacy preserving data mining techniques are remarkable good, but there is always extent for more enhancements.

IV. CONCLUSION AND FUTURE WORK

Due to the right to privacy in the information, privacy preserving data mining (PPDM) has become one of the newest trends in privacy and security and data mining research.

A single technique does not exceed all the parameters. Rather an algorithm may perform better than other algorithms on certain parameters (Ronica Raj, 2007).

Data mining aims to exact useful information from multiple sources, where as privacy preservation in data mining aims to reserve these data against disclosure or loss. Privacy preservation is one of the most important factors for an individual since he should not embarrass by an adversary.

Privacy constraints need to be developed by consulting various other discipline such as sociology, psychology etc. Efficient algorithms need to be developed that can balance all the parameters. Systems using soft computing techniques can be developed because of their tolerance against impression, uncertainty and partial truth.

Future works in this front can include defining a new privacy measure along with diversity for multiple sensitive attribute and we will focus to generalize attributes without suppression using other techniques which are used to achieve k-anonymity because suppression leads to reduce the precision of publishing table (Abou-el-ela Abdou Hussien, 2013).

REFERENCES

[1] Abou-el-ela Abdou Hussien, N. H. (2013). Attacks on Anonymization-Based Privacy-Preserving: A Survey for Data Mining and Data Publishing. Journal of Information Security.

[2] Alpa Shah, R. G. (2016). Privacy Preserving Data Mining: Techniques, Classification and Implications - A Survey. International Journal of Computer Applications.

[3] Ronica Raj, V. K. (2007). A Study on Privacy Preserving Data Mining:. International Journal of Innovative Research in Computer.

[4] https://ieeexplore.ieee.org/abstract/document/7950921

[5] https://www.researchgate.net/figure/A-PPDM-Framework_fig1_282788523.