ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

A Study on Privacy Preserving Data Mining: Techniques, Challenges and Future Prospects

Ronica Raj, Veena Kulkarni

M.E. Student, Dept. of Computer Science, Thakur College of Engineering and Technology, Mumbai, India

Assistant Professor, Dept. of Computer Science, Thakur College of Engineering and Technology, Mumbai, India

ABSTRACT: Data mining produces a large amount of data that needs to be analysed 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. Hence 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 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. Rather an algorithm may perform better than other algorithms on certain parameters.

KEYWORDS: Data mining, sensitive data, privacy preservation data mining techniques.

I. INTRODUCTION

Data mining is extracting or mining knowledge from large amounts of data. Data mining can also be referred to as KDD, i.e. Knowledge data discovery. Knowledge discovery is a process where the data goes through the following processes [1]:

1. Data cleaning: Removes noise and inconsistent data.
2. Data integration: Multiple data sources may be combined.
3. Data selection: Data relevant to analysis task are retrieved from database.
4. Data transformation: Data is transformed into forms appropriate for mining.
5. Data mining: Intelligent Methods are applied in order to extract data patterns.
6. Pattern evaluation: To identify and evaluate patterns representing knowledge.
7. Knowledge presentation: Knowledge representation techniques are used to present mined knowledge to the user.

When the data goes through all these processes the main concern is to preserve the privacy of the data so that the sensitive information does not get leaked to un-trusted parties. Thus privacy plays an important role in data mining. In the context of data mining, privacy is state of guarding against misuse or unauthorized access of individual data. For preserving the privacy of data there have been many algorithms and techniques introduced in the field of data mining.

PPDM i.e. privacy preserving data mining is one of the technique that is used to preserve the privacy of sensitive information. The area of privacy-preserving data mining has been growing over the last decade. 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 number of 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.

This study discusses about various techniques of PPDM, their advantages and their limitations and comes up with a conclusion and proposes a future prospect. This paper is organized as follows. Section 2 explains various keywords and concepts related to PPDM. Section 3 discusses about all five techniques of PPDM. Section 4 explains the evaluation of PPDM algorithms. Finally, section 5 gives a conclusion and some future prospects.

DOI: 10.15680/IJIRCCE.2015. 0311069

II. KEYWORDS AND CONCEPTS OF PPDM

PPDM consists of various techniques which will be explained further. But before moving on, there is a need to get some knowledge about some of the keywords and concepts which will be used in these techniques. This section gives some information about the keywords and concepts related to PPDM or used in the PPDM techniques.

A. Secure Multiparty Computation

Secure Multiparty Computation is a subfield of cryptography with the goal to create methods for parties to jointly compute a function over their inputs, keeping these inputs private. The most basic properties that a secure multiparty computation protocol aims to ensure are [2]:

1. Input privacy: No party should learn anything more than its prescribed output. The only information that should be learned about other parties' inputs is what can be derived from output itself.
2. Correctness: Each party is guaranteed that the output that it receives is correct.
3. Independence of Inputs: Corrupted parties must choose their inputs independently of the honest parties' inputs.
4. Guaranteed Output Delivery: Corrupted parties should not be able to prevent honest parties from receiving their outputs.
5. Fairness: Corrupted parties should receive their outputs if and only if honest also receive their outputs.

Secure Multiparty Computation methodologies have been proposed to enable a number of data holders to collectively mine their data without having to reveal their datasets to each other [3].

B. PPDM in Distributed Data Mining Model

Data mining refers to mining a large amount of data with an aim to extract useful knowledge from it. DDM i.e. distributed data mining refers to mining the same data using distributed resources.

Distributed data mining model assumes that the data sources are distributed across multiple sites. Algorithms developed in this field face the problem of efficiently getting the mining results from all data across these distributed sources. A simple approach to avoid this problem and perform data mining over multiple sources that will not share data is to run existing data mining tools independently and combine the results. But, this technique will often fail to give globally valid results because of the following reasons [3]:

1. Values of a single entity may be split across sources. Data mining at individual sites will be unable to detect cross-site correlations.
2. The same item may be duplicated at different sites and will be over-weighted in the results.
3. Important geographic and demographic distinction between that population and others cannot be seen on a single site.

C. PPDM using Soft Computing Techniques

Soft computing techniques can be used to handle inexact solutions to computationally hard tasks such as the solution of NP-complete problems, for which there is no known algorithm that can compute an exact solution in polynomial time. Soft computing differs from conventional (hard) computing in that, unlike hard computing, it is tolerant of imprecision, uncertainty and partial truth. The role model for soft computing is the human mind [4].

The use of soft computing leads to development of systems which has high Machine Intelligence Quotient (MIQ). The high MIQ of soft computing based systems helps for rapid growth in a good number and variety of applications. The soft computing techniques include [3]:

1. Fuzzy logic: Provides a natural framework for the process in dealing with uncertainty.
2. Neural networks: Highly interconnected network of a large number of processing elements called neurons in an architecture inspired by brain.
3. Genetic algorithms: Are adaptive, robust, efficient and global search methods, suitable in situations where the search space is large.
4. Rough sets: A mathematical tool for managing uncertainty that arises from indiscernibility between objects in a set.

Each of the soft computing techniques presents a different methodology to deal with the problems in its domain which results in an intelligent and robust system providing a human-interpretable, low cost and approximate solution as compared to traditional techniques. All these characteristics denote the possible and advantageous use of soft computing in PPDM.

D. Dimensions for Classification of PPDM Techniques

PPDM tends to transform the original data so that the results of data mining tasks should not defy the privacy constraints. Lists of dimensions on which PPDM techniques are based are [3]:

1. Data distribution: This dimension is related to distribution of data. Data can be Centralized or Distributed. Distributed data can be of two types:
   - (i) Horizontal distribution: Refers to cases where different records reside in different places.
   - (ii) Vertical distribution: Refers to cases where all values of different attributes reside in different places.
2. Data Modification: This dimension refers to modification of original values of data that are to be released for data mining task. Modification is carried out by using the following techniques:
   - (i) Perturbation: Perturbation of data is an easy and effective technique for protecting sensitive electronic data from unauthorized use.
   - (ii) Blocking: Blocking-based technique aims at hiding some sensitive information when data is shared for mining [5].
   - (iii) Aggregation: Data aggregation is a process in which information is gathered and expressed in a summary form for the purpose of statistical analysis [6].
   - (iv) Merging: Data merging refers to combination of several values of data.
   - (v) Swapping: Data swapping refers to interchanging values of various data records [7].
   - (vi) Sampling: Data sampling refers to releasing data for only a sample of population [7].
3. Data Mining Algorithms: Data mining algorithms are applied on transformed data to get useful nuggets of information that where hidden previously.
4. Data hiding: This dimension refers to whether the raw data or aggregated data should be hidden.
5. Privacy Preservation: This dimension refers to techniques that are used for protecting privacy.

III. PPDM TECHNIQUES

Based on the dimensions mentioned in the previous section and, PPDM techniques are of five types:

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

Let’s look at each technique in detail.

A. Anonymization based PPDM

Before proceeding with the explanation of anonymization based PPDM, it is necessary to know about the basic form of data in a data table and its attributes.

Table 1 gives four types of attributes a data table can have [3]:

TABLE I
TYPES OF ATTRIBUTES IN A DATA TABLE

Attributes	Description
Explicit Identifiers	Set of attributes containing information that identifies a record owner explicitly such as name, SS number etc.
Quasi Identifiers	Set of attributes that could potentially identify a record owner when combined with publicly available data.
Sensitive Identifiers	Set of attributes that contains sensitive person specific information such as disease, salary etc.
Non-sensitive Identifiers	Set of attributes that creates no problem if revealed even to untrustworthy parties.

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 publically available data. Such an attack can be termed as linking attack.

Latanya Sweeney [8] proposed *k*-anonymity model using generalization and suppression to achieve *k*-anonymity. *K*-anonymity model states that any individual is distinguishable from at least *k*-1 other ones with respect to quasi-identifier attribute in the anonymized dataset.

Releasing such data for mining reduces the risk of identification of record owner when combined with public data. But *k*-anonymity model has a lot of disadvantages such as [3]:

1. Accuracy of application is reduced on transformed data.
2. Anonymization method suffers heavy information loss.
3. It becomes very hard for the owner of the database to determine which of the attributes of data are available in the external table and which are not.
4. *K*-anonymity model assumes a certain method of attack whereas the attacker can also try different methods.
5. 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.

Eventually many models such as *p*-sensitive *k*-anonymity model [9] which protect against both identity and attributes disclosure, *t*-closeness [10] which works on the principal that the distance between two distributions of sensitive attributes in a class should not be more than a threshold *t* etc. where proposed to handle the limitations of *k*-anonymity model.

B. Perturbation based PPDM

Perturbation of data is a very easy and effective method for protecting the sensitive electronic information of the data from unauthorised users or hackers.

There are two types of data perturbation for protecting electronic data namely [11]:

1. Probability distribution approach: This approach takes the data and replaces it from the same distribution sample or from the distribution itself.
2. Value distortion approach: This approach perturbs data by adding noise, or other randomized processes.

In perturbation based PPDM, the original values of the data are replaced with some artificial values so that the result computed from the perturbed data does not differ from the result computed from the original data to a larger extent. But the individual records of the perturbed data are of no use as only statistical properties of records are preserved.

As the perturbed data records does not match with the original records, the attacker cannot recover the sensitive information from the perturbed data. Perturbation of data can be done by adding noise to the data, swapping the data or replacing values of the original data with artificial values.

Perturbation based PPDM only reconstructs the distribution of data not the original values of the data, hence new algorithms needs to be developed for mining the data for each individual problems like classification, clustering or association rule mining.

One disadvantage of perturbation based PPDM is that each dimension is reconstructed independently. Hence there is a loss of implicit data in multidimensional records as any distribution based data mining algorithm treats different attributes independently.

C. Randomized Response based PPDM

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 behaviour while maintaining confidentiality [12].

In the randomised 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 [6].

Randomized response model consist of two steps for the process of collecting data [3]:

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.

These two steps of randomized response model are shown in Fig 1.

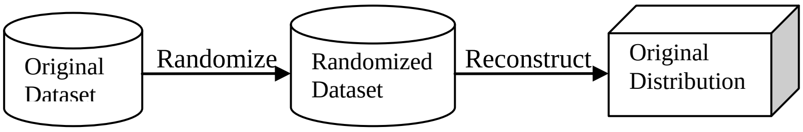

Fig 1 Steps in Randomized Response Model

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 such as Agarwal and Srikant [13] introduced a randomization scheme in which a random number is added to a value of an attribute whose value is a sensitive one, Kargupta et al. [13] proposed a random matrix based spectral filtering technique through which the original data will be recovered from the perturbed data and so on.

D. *Condensation based PPDM*

Condensation approach was first introduced by Charu C. Aggarwal and Philip S. Yu [14] which condenses the data into multiple group of pre-defined size and for each group a certain level of statistical information about different records is maintained.

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 [3].

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.

E. *Cryptography based PPDM*

Cryptography is a technique of hiding the data and transmitting it in such a way that only those for whom it is intended can read and process it. This technique is mainly used for secure communication between two parties in the presence of the third party.

Cryptographic techniques find their usage in such scenarios where multiple parties collaborate to compute results or share non-sensitive mining results, thereby avoiding disclosure of sensitive information [3].

It finds its utility in such scenarios for two reasons:

1. It offers a well defined model for privacy.
2. A vast set of cryptographic algorithms to implement PPDM algorithms are available.

There have been many privacy preserving data mining techniques introduced under cryptography based PPDM which reveal nothing other than the final results. [3]

These techniques include: Murat Kantarcioglu and Chris Clifton [15] proposed a method incorporating cryptography techniques to preserve privacy in association rule mining over horizontally partitioned data to minimize the information shared and overheads while sharing this information, Yehuda Lindell and Benny Pinkas [16] has proposed a protocol which shows how to generate ID3 decision trees on horizontally partitioned data, Zhiqiang Yang, Sheng Zhong and Rebecca N. Wright [17] propose a simple cryptographic approach which is based on horizontally partitioned data and provides strong privacy for each customer without losing any accuracy as cost of privacy and so on.

Although cryptographic techniques ensure that the transformed data is secure but this approach fails to deliver when more than a few parties are involved and also the data mining results may breach the privacy of an individual [3].

IV. Evaluation of PPDM Algorithms

After studying each privacy preserving data mining algorithm in detail, it is important to evaluate each technique against certain parameters.
Table 2 gives these parameters along with the description of each parameter [3].

TABLE 2
Evaluation Table

Parameters	Description
Performance	Measured in terms of the time required to achieve the privacy criteria.
Data Utility	Measure of information loss or loss in the functionality of data in providing the results, which could be generated in the absence of PPDM algorithms
Uncertainty Level	Measure of uncertainty with which the sensitive information that has been hidden can still be predicted.
Resistance	Measure of tolerance shown by PPDM algorithm against various data mining algorithms and models.
Complexity	Measure of the amount of time and/or space required by an algorithm for an input of a given size

V. Conclusion and Future Prospect

Privacy and accuracy in case of data mining is a pair of contradiction.After a review of a good number of existing PPDM techniques, this study concludes that:
A single technique does not exceed all the parameters. Rather an algorithm may perform better than other algorithms on certain parameters.
Table 3 shows the comparison of various PPDM techniques measured in terms of its performance, data utility, level of uncertainty, resistance to data mining algorithms and complexity.

TABLE 3
Conclusion Table

Parameters Techniques	Performance	Data Utility	Level of uncertainty	Resistance to data mining algorithms	Complexity
Anonymization based PPDM	Accuracy is reduced	Suffers heavy data loss	Level of uncertainty exists	Various algorithms have been proposed to implement k-anonymity model	Complex
Perturbation based PPDM	Efficient and the ability to preserve statistical information	Loss of implicit information in multidimensional records	Low level of uncertainty exists	New algorithms to be developed for mining data. Or Restricts the range of algorithmic techniques	Simple
Randomized Response based PPDM	Accurate	Reduces the utility of data	Level of uncertainty is negligible	Can be useful for decision tree classification	Simple
Condensation approach based	This approach helps in better	Large amount of information loss	Low level of uncertainty	Can be useful for a variety of data	Simple

DOI: 10.15680/IJIRCCE.2015. 0311069

PPDM	privacy preservation			mining problems	
Cryptography based PPDM	Offers a well defined model for privacy	It is a good technique to preserve the data	Uncertainty increases when more than few parties are involved	A vast set of cryptographic algorithms are available in this domain	Difficult to apply for huge databases

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.

REFERENCES

[1] Jiawei Han, and Micheline Kamber, "Data Mining: Concepts and Techniques", pp. 5-7

[2] Yehuda Lindell and Benny Pinkas, "Secure Multiparty Computation for Privacy-Preserving Data Mining", *The Journal of Privacy and Confidentiality*, 2009, pp 62-63.

[3] Majid Bahir Mailk, M. Asger Gahzi, Rashid Ali, "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects", *IEEE Third International Conference on Computer and Communication Technology*, 2012.

[4] https://en.wikipedia.org/wiki/Soft_computing.

[5] Stanley R. M. Oliveira and Osmar R. Zaiane, "Privacy-Preserving Data Mining on the Web: Foundations and Techniques".

[6] http://searchsqlserver.techtarget.com/definition/data-aggregation.

[7] Lai Xu, Katalin Tarney and Sandor Imre, "Research and development in E-business through Service Oriented Solutions", 2013, chapter 4, pp 74.

[8] Latanya Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression", volume 10, issue 5, 2002.

[9] Traian Marius Truta and Bindu Vinay, "Privacy Protection: *p*-Sensitive *k*-Anonymity Property", Department of Computer Science, Northern Kentucky University.

[10] Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian, "*t*-Closeness: Privacy beyond *k*-Anonymity and *l*-Diversity".

[11] https://www.techopedia.com/definition/25013/data-perturbation.

[12] https://en.wikipedia.org/wiki/Randomized_response.

[13] A.S. Shanthi and Dr. Karthikeyan, "A Review on Privacy Preserving Data Mining".

[14] Charu C. Aggarwal and Philip S. Yu "A Condensation Approach to Privacy Preserving Data Mining", *Springer-Verlag Berlin Heidelberg*, 2004.

[15] Murat Kantarcioglu and Chris Clifton, "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data", *IEEE Transaction on Knowledge and Data Engineering*, volume. 16, issue 9, 2004.

[16] Yehuda Lindell and Benny Pinkas, "Privacy Preserving Data Mining".

[17] Zhiqiang Yang, Sheng Zhong and Rebecca N. Wright, "Privacy-Preserving Classification of Customer Data without Loss of Accuracy", *Fifth SIAM International Conference on Data Mining*, 2005.