Comprehensive on PPDM And PPDDM Techniques

P.AnnanNaidu[1], Dr. M. Vamsi Krshna[2]
[1, 2] Dept of CSE
[1, 2] Centurion University, India.

Abstract- *Data Mining is the process of extracting the data or knowledge from different data resources like data base and Data warehouse. Data Mining is one of the emerging research area which deals with privacy and security of data. There are many Privacy Preserving Data mining (PPDM) and Privacy Preserving Distributed Data mining (PPDDM) techniques are used to solve privacy issues in Centralized Data mining environment and Distributed data mining Environment respectively. This paper aims to provide the comparative analysis study of PPDM and PPDDM.*

Keywords- Anonymization, Data Mining, Distributed data mining, Privacy preserving, secure communication, , Data distribution.

I. INTRODUCTION

In the recent years there is extreme research concerned about privacy and security of data in the context of data mining and distributed data mining. Data mining provides wide variety of techniques and algorithms for privacy preserving data mining(PPDM) and privacy preserving distributed Data mining(PPDDM).This paper presents the comparative study on PPDM and PPDDM. This part presents about basic Terms used in this paper.

Privacy vs Security: Data security is commonly referred to as the confidentiality, availability, and integrity of data. Data Security ensures that authorized access of data. Data Security ensures data isn't being used or accessed by unauthorized individuals or parties. Data security ensures that the data is accurate and reliable and is available when those with authorized access need it. Data privacy is defined as the appropriate use of data. Data privacy ensures that data should be used upon agreed purposes. Privacy is about informational self-determination--the ability to decide what information about you goes where.

Privacy Violation: Users' privacy can be violated in different ways and with different intentions. Privacy can be violated if personal data are used for other purposes subsequent to the original transaction between an individual and an organization when the information was collected

Privacy Preserving: one major factor contributes to privacy violation in data mining: the misuse of data. Privacy-preserving means preventing privacy violation (i.e data misuse).In general, privacy preservation occurs in two major dimensions: users' personal information and information concerning their collective activity.

Individual privacy preservation: The primary goal of data privacy is the protection of personally available information. In general, information is considered personally identifiable if it can be linked, directly or indirectly, to an individual person. Thus, when personal data are subjected to mining, the attribute values associated with individuals are private and must be protected from disclosure. Miners are then able to learn from global models rather than from the characteristics of a particular individual.

Collective privacy preservation: Protecting personal data may not be enough. Sometimes, we may need to protect against learning sensitive knowledge representing the activities of a group. We refer to the protection of sensitive knowledge as collective privacy preservation. The goal here is quite similar to that one for statistical databases, in which security control mechanisms provide aggregate information about groups (population) and, at the same time, should prevent disclosure of confidential information about individuals. However, unlike as is the case for statistical databases, another objective of collective privacy preservation is to preserve strategic patterns that are paramount for strategic decisions, rather than minimizing the distortion of all statistics (e.g., bias and precision). In other words, the goal here is not only to protect personally identifiable information but also some patterns and trends that are not supposed to be discovered.

Need of Privacy Preserving:

Data mining techniques were developed to extracts knowledge to support various domains like weather forecasting, marketing, medical diagnosis and national security. But it is still challenging to mine specific data without violating data owners 'privacy. For instance, mining patients 'private data is an ongoing problem in health care applications. As data mining become more pervasive, privacy concerns increase. Commercial issues are also linked to the

privacy issue. Most organizations collect information about individuals for specific needs. Frequently different units in an organization may find it necessary to share information. In such cases, each organization/unit must ensure that individual privacy is not violated or sensitive business information revealed. To avoid these types of violations, there is a need of various data mining algorithm for privacy preserving

Some ways in which privacy concerns raised by data mining are as follows (Oliveira et al 2004),

I. The implicit patterns involving information about persons that can be derived from data in the data-mining process vs. the Explicit nature of the personal data (in records) extracted in traditional database retrieval techniques.
II. The use of (possibly) a single database (or data warehouse') to extract information about persons vs. the use of multiple databases to exchange and retrieve such information.
III. The use of 'open-ended' queries to discover information on relationships and associations about individuals and groups of individuals vs. (traditional) specific queries to retrieve information about relationships and associations that are already known to exist..
IV. The non-predictive aspect of information about persons gained from data mining vs. the generally predictive aspect of information retrieved from traditional database techniques.
V. The public nature of much of the information about persons that is extracted through the data mining process vs. the private or intimate nature of the information about persons retrieved and exchanged in traditional database-exchange techniques.
VI. The ability to construct new groups or categories of persons based on patterns of information derived from data mining vs. the mere extraction of information about individuals themselves from personal data accessible to traditional techniques of database retrieval.

II. PPDM AND PPDDM

A. PPDM-Privacy-Preserving Data Mining (PPDM):

Privacy-Preserving Data Mining (PPDM) is a data mining and statistical databases innovative field where data mining algorithms are analyzed for side-effects in data privacy. It is also called privacy enhanced/privacy-sensitive data mining dealing with getting valid data mining results without learning underlying data values. This reveals how many different methods and techniques can be used in a PPDM context from a technical perspective.

The goals of a PPDM algorithm include:

i. Prevent the discovery of sensible information.
ii. Being uncompromised to access and to use the non-sensitive data.
iii. Being usable on large amounts of data.
iv. Must have less exponential computational complexity.

To achieve optimized results while preserving the privacy of the data subjects efficiently, five dimensions
need to be considered and listed below:

1) The distribution of the basic data
2) The modification of the basic data
3) Mining method being used
4) If basic data or rules are to be hidden and
5) Additional methods for privacy preservation used.

B. PPDDM Privacy Preserving Distributed Data Mining

Distributed computing plays an important role in the Data Mining process for several reasons. First, Data Mining often requires huge amounts of resources in storage space and computation time. To make systems scalable, it is important to develop mechanisms that distribute the work load among several sites in a flexible way. Second, data is often inherently distributed into several databases, making a centralized processing of this data very inefficient and prone to security risks. Distributed Data Mining explores techniques of how to apply Data Mining in a non-centralized way.

DDM is a branch of the field of data mining that offers a framework to mine distributed data paying careful attention to the distributed data and computing resources. [1] DDM is the process of performing data mining in distributed computing environments, where users, data, hardware and data mining software are geographically distributed. It emerges as an area of research interest to deal with naturally distributed and heterogeneous databases and then to address the scalability bottlenecks of mining very large datasets [3]. A number of distributed algorithms have been developed for different DDM tasks, including distributed classification, clustering and association. DDM refers to the mining of inherently distributed datasets, aiming to generate global patterns from the Union set of locally distributed data **[2]**Let us first take a look at two real-world examples of distributed data mining with different privacy constraints: [3]

A. Scenario 1:

Multiple competing supermarkets, each having an extra-large set of data records of its customer's buying behaviors, want to conduct data mining on their joint data set obtain certain global patterns that will benefit them . These companies are competitors in the market, so they do not want to disclose too much about their customer's information with each other, but definitely they know the results obtained from this collaboration could bring them an advantage over other competitors.

B. Scenario 2:

Success of homeland security aiming to counter terrorism depends on combination of strength across different mission areas, effective international collaboration and information sharing to support coalition in which different organizations and nations have to share some, but not all, information. Thus Information privacy becomes an extremely important; all the parties of the collaboration promise to provide their private data to the collaboration, but neither of them want each other or any other party to learn much about their private data.

Each scenario shows a set of challenges .Scenario 1 is an example of heterogeneous collaboration, while scenario 2 refers to a task in a homogeneous cooperation setting. Technology alone cannot address all of the Privacy Preserving Distributed Data Mining (PPDDM) scenarios [3].

Some features of a distributed scenario where DDM is applicable are as follows. **[4]**. the system consists of multiple independent sites of data and computation which communicate only through message passing.

2. Communication between the sites is expensive.
3. Sites have resource constraints
4. Sites have privacy concerns

In Data mining, Privacy preserving is addressed by using different PPDM methods. One method is a reconstruction-based approach which reconstructs the distribution probability of the original dataset and creates a new distribution curve. Another method is a heuristics-based approach that protects individual information by using data perturbation methods, such as blocking,generalizing,aggregating and swapping, etc. These two approaches have a major drawback when dealing with privacy-preserving data mining problems. They trade off between the privacy of the individual information and the correctness of the data mining results. That is, privacy is achieved at the cost of accurate outcome. Besides, such kind of solutions can only tackle centralized data mining applications.

The key goal in most distributed methods for privacy-preserving data mining is to allow computation of useful aggregate statistics over the entire data set without compromising the privacy of the individual data sets within the different participants. The participants may wish to collaborate in obtaining aggregate results, but may not fully trust each other in terms of the distribution of their own data sets. This requires secure and cryptographic protocols for sharing the information across the different parties. For this purpose, the data sets may either be horizontally partitioned or be vertically partitioned.

The data may be distributed in two ways across different sites:

Horizontal Partitioning: In this case, the different sites may have different sets of records containing the same attributes.

Vertical Partitioning: In this case, the different sites may have different attributes of the same sets of records.

Both kinds of partitioning pose different challenges to the problem of distributed privacy-preserving data mining.

III. PPDM TECHNIQUES

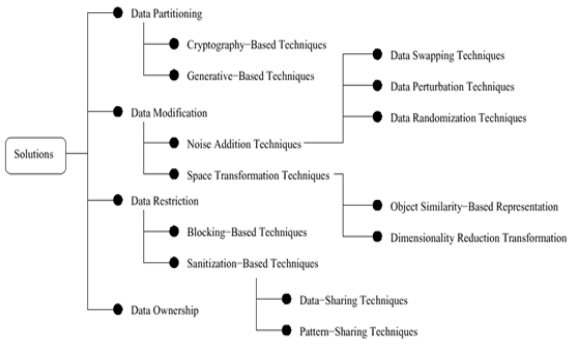

Fig .1 classification of existing PPDM techniques

Data Partitioning Techniques

Data Partitioning Techniques are applied where data can be distributed in different locations. Data can be distributed either horizontally or vertically. In a horizontal partition all data sets are described with same schema in all partitions. while in a vertical partition the attributes of the same entities are split across the partitions. The existing

solutions can be classified into Cryptography-Based Techniques and Generative-Based Techniques.

a. Cryptography-Based Techniques: In context of PPDM over distributed data nature, Cryptography-Based Technique used to solve the following problem, : two or more parties want to conduct a computation based on their private inputs. The issue here is how to conduct such a computation so that no party knows anything except its own input and the results. This problem is referred to as the Secure Multi-Party Computation (SMC) problem. The technique proposed address privacy-preserving classification, while the techniques proposed address privacy-preserving association rule mining, and the technique addresses privacy-preserving clustering.

***b. Generative-Based Techniques*:** These techniques are designed to perform distributed mining tasks. In this approach, each party shares just a small portion of its local model which is used to construct the global model. The existing solutions are built over horizontally partitioned data. The solution presented in addresses privacy-preserving frequent item sets in distributed databases, whereas the solution in addresses privacy-preserving distributed clustering using generative models.

Data Modification Techniques

Data modification techniques modify the original values of a database that needs to be shared, and in doing so, privacy preservation is ensured. The transformed database is made available for mining and must meet privacy requirements without losing the benefit of mining. In general, data modification techniques aim at finding an appropriate balance between privacy preservation and knowledge disclosure. Methods for data modification include noise addition techniques and space transformation techniques.

a. Noise Addition Techniques: The idea behind noise addition techniques for PPDM is that some noise (e.g., information not present in a particular tuple or transaction) is added to the original data to prevent the identification of confidential information relating to a particular individual. In other cases, noise is added to confidential attributes by randomly shuffling the attribute values to prevent the discovery of some patterns that are not supposed to be discovered. We categorize noise addition techniques into three groups: (1) data swapping techniques that interchange the values of individual records in a database (2) data distortion techniques that perturb the data to preserve privacy, and the distorted data maintain the general distribution of the original data and (3) data randomization techniques which allow one to perform the discovery of general patterns in a database with error bound, while protecting individual values. Like data swapping and data distortion techniques, randomization techniques are designed to find a good compromise between privacy protection and knowledge discovery

***b. Space Transformation Techniques*:** These techniques are specifically designed to address privacy-preserving clustering. These techniques are designed to protect the underlying data values subjected to clustering without jeopardizing the similarity between objects under analysis. Thus, a space transformation technique must not only meet privacy requirements but also guarantee valid clustering results. We categorize space transformation techniques into two major groups: (1) object similarity-based representation relies on the idea behind the similarity between objects, i.e., a data owner could share some data for clustering analysis by simply computing the dissimilarity matrix (matrix of distances) between the objects and then sharing such a matrix with a third party. Many clustering algorithms in the literature operate on a dissimilarity matrix). This solution is simple to be implemented and is secure, but requires a high communication cost); (2) dimensionality reduction-based transformation can be used to address privacy-preserving clustering when the attributes of objects are available either in a central repository or vertically partitioned across many sites. By reducing the dimensionality of a dataset to a sufficiently small value, one can find a trade-off between privacy, communication cost, and accuracy. Once the dimensionality of a database is reduced, the released database preserves (or slightly modifies) the distances between data points. In tandem with the benefit of preserving the similarity between data points, this solution protects individuals' privacy since the attribute values of the objects in the transformed data are completely different from those in the original data.

Data Restriction Techniques

Data restriction techniques focus on limiting the access to mining results through either generalization or suppression of information (e.g., items in transactions, attributes in relations), or even by blocking the access to some patterns that are not supposed to be discovered. Such techniques can be divided into two groups: Blocking-based techniques and Sanitization-based techniques.

***a. Blocking-Based Techniques*:** These techniques aim at hiding some sensitive information when data are shared for mining. The private information includes sensitive association rules and classification rules that must remain private. Before releasing the data for mining, data owners must consider how much information can be inferred or calculated from large databases, and must look for ways to minimize the leakage of

such information. In general, blocking-based techniques are feasible to recover patterns less frequent than originally since sensitive information is either suppressed or replaced with unknowns to preserve privacy. The techniques in address privacy preservation in classification, while the techniques in address privacy-preserving association rule mining.

b. ***Sanitization-Based Techniques:*** Unlike blocking-based techniques that hide sensitive information by replacing some items or attribute values with unknowns, sanitization-based techniques hide sensitive information by strategically suppressing some items in transactional databases, or even by generalizing information to preserve privacy in classification. These techniques can be categorized into two major groups: (1) data-sharing techniques in which the sanitization process acts on the data to remove or hide the group of sensitive association rules that contain sensitive knowledge. To do so, a small number of transactions that contain the sensitive rules have to be modified by deleting one or more items from them or even adding some noise, i.e., new items not originally present in such transactions and (2) pattern-sharing techniques in which the sanitizing algorithm acts on the rules mined from a database, instead of the data itself. The existing solution removes all sensitive rules before the sharing process and blocks some inference channels. In the context of predictive modeling, a framework was proposed for preserving the anonymity of individuals or entities when data are shared or made publicly.

Data Ownership Techniques

Data ownership techniques can be applied to two different scenarios: (1) to protect the ownership of data by people about whom the data were collected. The idea behind this approach is that a data owner may prevent the data from being used for some purposes and allow them to be used for other purposes. To accomplish that, this solution is based on encoding permissions on the use of data as theorems about programs that process and mine the data. Theorem proving techniques are then used to guarantee that these programs comply with the permissions; and (2) to identify the entity that receives confidential data when such data are shared or exchanged. When sharing or exchanging confidential data, this approach ensures that no one can read confidential data except the receiver(s). It can be used in different scenarios, such as statistical or research purposes, data mining, and on-line business-to-business (B2B) interactions.

Table-1: Merits and Demerits of different techniques of PPDM

Techniques of PPDM	**Merits**	**Demerits**
ANONYMIZATION	This method is used to protect Respondents' identities while releasing truthful information. While k-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure.	There are two attacks: the homogeneity attack and the background knowledge attack. Because the limitations of the k-anonymity model stem from the two assumptions. First, it may be very hard for the owner of a database to determine which of the attributes are or are not available in external tables. The second limitation is that the k-anonymity model assumes a certain method of attack, while in real scenarios there is no reason why the attacker should not try other methods.
PERTURBATION	Independent treatment of the different attributes by the perturbation approach	The method does not reconstruct the original data values, but only distribution, new algorithms have been developed which uses these reconstructed distributions to carry out mining of the data available.
RANDOMIZED RESPONSE	The randomization method is a simple technique which can be easily implemented at data collection time. It has been shown to be a useful technique for hiding	Randomized Response technique is not for multiple attribute databases.

	individual data in privacy preserving data mining. The randomization method is more efficient. However, it results in high information loss.	
CONDENSATION	This approach works with pseudo-data rather than with modifications of original data, this helps in better preservation of privacy than techniques which simply use modifications of the original data.	The use of pseudo-data no longer necessitates the redesign of data mining algorithms, since they have the same format as the original data.
CRYPTOGRAPHIC	Cryptography offers a well-defined model for privacy, which includes methodologies for proving and quantifying it. There exists a vast toolset of cryptographic algorithms and constructs to implement privacy- Preserving data mining algorithms.	This approach is especially difficult to scale when more than a few parties are involved. Also, it does not address the question of whether the disclosure of the final data mining result may breach the privacy of individual records.

IV. PPDDM TECHNIQUES

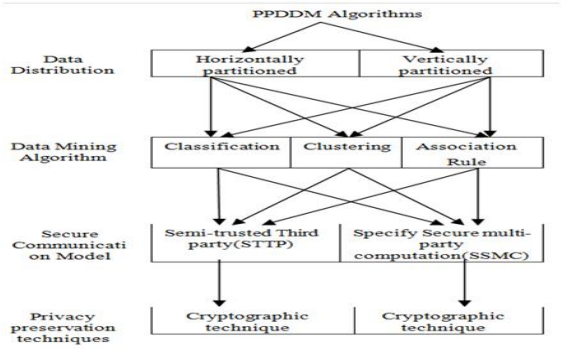

Fig.2 classification related to PPDDM techniques

Data Partitioning Model: In this scenario, Data sets can be spread in different sites. There are two ways of data distribution such as homogeneous distribution (horizontal partitioning) and heterogeneous distribution (vertical partitioning).In Horizontal Partitioning, the different sites or locations may have different data sets of records containing with the same attributes. In Vertical Partitioning, the different sites or locations may have different attributes of the same datasets of records.

Data Mining Tasks / Algorithms:

Classification concern the problem of finding a set of models that describe and distinguish the data classes .We use this models to predict the class of records whose class label is unknown. Association rule mining is the process of discovering association rules and showing attribute value and conditions that occur frequently in a given set of data.

Clustering analysis involves the process of decomposing or partitioning a data set into group so that the points in one group are similar to each other and are as different as possible from the points in other groups.[3]

Secure Communication Model:

privacy-preserving distributed data mining problems can be solved mainly based on two types of secure computation model: One is based on Semi-trusted Third Party (STTP) model. Theoretically, the general secure multi-party computation protocols can be used to deal with any collaborative data mining problems, yet this kind of solutions are too inefficient when the database is huge in amount and the number of participants is large, due to its intricate and complicated design. On the other hand of the spectrum, the trusted third party (TTP) model is too navie and Straightforward, so the privacy is compromised to a larger extent at the point of the TTP. Therefore, more practical solutions have been put forward in the past few years with respect to how to solve the privacy issues of distributed data mining more efficiently and accurately. Among them, two broad streams of ideas are manifesting themselves: one is to introduce a semi-trusted third party, as compared to the trusted third party (TTP). In real world, it is much more feasible to find such a semi-trusted third party than to find a trusted third party. This semi-trusted third party can be implemented by means of a miner, a mixer, or a commodity server, that all act in a semi-trusted manner.

The other stream is based on Specific Secure Multi-party Computation under Semi-honest assumption (SSMC). It aims at accomplishing efficient and accurate solution for the PPDDM problems. Under the semi-honest assumption, specific secure multi-party computation protocols are employed to deal with functions commonly used in data mining applications rather than the general secure multi-party computation protocol. These techniques include secure sum, secure set union, secure intersection, secure scalar product, etc. The advantage of such kind of protocols and tools lies in that they are designed to specially fit in with the data mining tasks, instead of any general functions. As the function for secure computation can be identified, the computing complexity is reduced greatly and a linear proportional cost can be obtained. **[3]**

Privacy Preservation Techniques: privacy preserving techniques used to protect the private information communicated among sites and central miner or mixer. These techniques include homomorphic encryption scheme, public key cryptosystem

A homomorphic encryption scheme is an encryption scheme which allows certain algebraic operations to be carried out on the encrypted plaintext, by applying an efficient operation to the corresponding cipher text.

V. CONCLUSION

This paper presents overall brief discussion on techniques of PPDM and PPDDM with its applications and its comparison. Future scope this paper is developing different algorithms for PPDM and PPDDM techniques.

REFERENCES

[1] Distributed data mining and agents,J osenildo C. da Silva a , Chris Giannella,Engineering Applications of Artificial Intelligence 18 (2005) 791–807,2005 Elsevier Ltd.
[2] Distributed data mining in grid computing environments, Ping Luo, Kevin L¨uc, Future Generation Computer Systems 23 (2007) 84–91, Elsevier.
[3] Privacy Preserving Distributed Data MiningTechniques,Mayur B Tank,Tushar A Champaneria,IJIRST,Volume 1 | Issue 9 | February 2015.
[4] Distributed data mining and agents,J osenildo C. da Silva a , Chris Giannella,Engineering Applications of Artificial Intelligence 18 (2005) 791–807,2005 Elsevier Ltd.
[5] Privacy-Preserving Data Mining on the Web: Foundations and Techniques, Stanley R. M. Oliveira, Osmar R.
[6] Ning Zhang, Ming Li, Wenjing Lou Distributed Data Mining with Differential Privacy, IEEE Communications Society subject matter experts for publication in the IEEE ICC 2011 proceedings.
[7] V. Baby, Privacy-Preserving Distributed Data Mining Techniques: A Survey,IJCA, Volume 143 – No.10, June 2016
[8] Chris Clifton, Tools for Privacy Preserving Distributed Data Mining, SIGKDD Explorations, Volume 4, Issue 2.
[9] V.Thavavel , S.Sivakumar, A generalized Framework of Privacy Preservation in Distributed Data mining for Unstructured Data Environment, IJCSI, Vol. 9, Issue 1, No 2, January 2012.
[10] Masooda Modaka ,Rizwana Shaikh, Privacy Preserving Distributed Association Rule Hiding Using Concept Hierarchy, 7th International Conference on Communication, Computing and Virtualization 2016.