# Encryption-Based Techniques For Privacy Preserving Data Mining

V.Sathya, and Dr.V. Gayathiri

**ABSTRACT:**

In present years, advances in hardware proficiency have lead to an increase in the capability to store and record personal data about consumers and individuals. This has lead to concerns that the personal data may be misused for a variety of purposes. In order to lighten these concerns, a number of techniques have newly been proposed in order to perform the data mining tasks in a privacy-preserving way. 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 paper is to analyze and assess techniques of privacy preserving, introducing a framework based on methods of cryptography in data mining with respect to the privacy preserving. Considering the prevailing application of data mining methods in distributed databases, the suggested classification can possibly be influential in opting for a proper approach.

**Keywords –** Cryptography, Distributed Data Mining, Privacy Preserving.

## I. INTRODUCTION

Explosive development in networking, storage, and processor technology has led to the creation of very large databases that record unprecedented amount of transactional information. The main problem is that with the availability of non-sensitive information or unclassified data, one is able to infer sensitive information that is not supposed to be disclosed. Despite its benefits in various areas such as marketing, business, medical analysis, bioinformatics and others, data mining can also pose a threat to privacy in database security if not done or used properly. Privacy preserving data mining, is a novel research direction in data mining and statistical databases, where data mining algorithms are analyzed for the side-effects they incur in data privacy[1]

Privacy preserving data mining (PPDM) has emerged to address this issue. Most of the techniques for PPDM uses modified version of standard data mining algorithms, where the modifications usually using well known cryptographic techniques ensure the required privacy for the application for which the technique was designed. In most cases, the constraints for PPDM are preserving accuracy of the data and the generated models and the performance of the mining process while maintaining the privacy constraints. The several approaches used by PPDM can be summarized as below:

1. The data is altered before delivering it to the data miner.
2. The data is distributed between two or more sites, which cooperate using a semi-honest protocol to learn global data mining results without revealing any information about the data at their individual sites.
3. While using a model to classify data, the classification results are only revealed to the designated party, who does not learn anything else other that the classification results, but can check for presence of certain rules without revealing the rules.

Many approaches emerged for privacy preserving data mining. The first approach involved perturbing the input before mining. Though it has the benefit of simplicity it does not provide a formal framework for proving how much privacy is guaranteed. Secure Computation technique [2] has the advantage of providing a well defined model for privacy using cryptographic techniques and is also accurate. However

V. Sathya Guest Lecturer, Department of Computer Science, Govt. Arts college for Women, Krishnagiri - 635 001, Tamil Nadu,India. E-mail:sathyashri2009@yahoo.co.in
Dr.V. Gayathiri, Assistant Professor, Department of ComputerScience, Pee Gee College of Arts And Science
Dharmapuri – 636803, Tamil Nadu,India,
E-mail: gayhar11@gmail.com

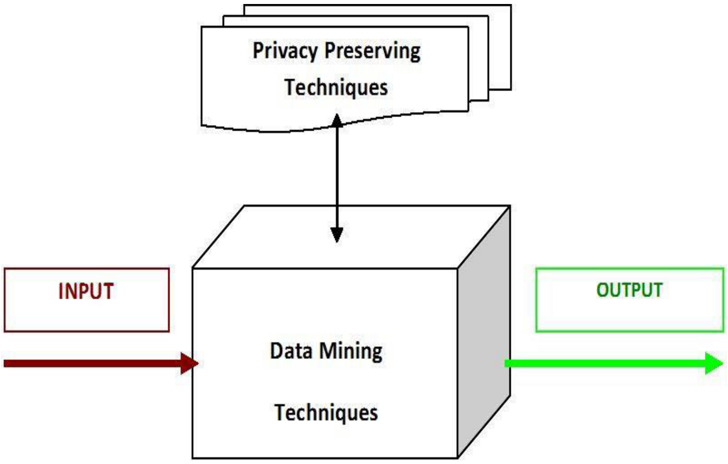
it is a slower method. In PRBAC technique, access to sensitive objects (SOBS) is based on roles [1].But it has the drawback of space complexity. i.e., as all data are stored in the Database Server, it leads to a large memory requirement. Also, risk of illegal access of data is not completely ruled out as the entire data is stored at one site. In our paper, we address these problems by applying vertical fragmentation and cryptographic techniques for data storage.

Here, we propose a new approach to privacy preserving data mining based on cryptographic role based access control approach (PCRBAC) where we have 2 sets of object: Sensitive objects (SOBS) and Non sensitive objects (NSOBS). Using the data mining technique, users are allowed to mine different sets of data based on their roles. The data is first classified as sensitive objects and non sensitive objects. Sensitive objects are encrypted and stored. The permitted user can access the sensitive objects only after decryption ensuring privacy.

## II. RELATED WORK

Privacy-Preserving Data Mining (PPDM) [2] is a new research area that investigates how the privacy of data can be maintained either before or after applying Data Mining (DM) techniques on the data. Privacy-preservation of sensitive information in data mining methods is an important subject in data communication and knowledge discovery systems. As a simple example, suppose some hospitals want to get useful aggregated knowledge about a specific diagnosis from their patients' records while each hospital is not allowed, due to the privacy laws, to disclose individuals' private data. Therefore, they need to run a joint and secure protocol on their distributed database to reach the desired Many secure protocols have been proposed so far for data mining and machine learning techniques such as [5] for decision tree classification, for clustering, for association rule mining, [6] for Neural Networks, and for Bayesian Networks. The main concern of these algorithms is to preserve the privacy of parties' sensitive data, while they gain useful knowledge from the whole
dataset.

information.



## III. APPLICATION

Privacy issues arise when distributed data computing applications become popular in private and public sectors. Let us first investigate two genuine illustrations of distributed data mining with various privacy limitations:

- Numerous competing general stores, each having an additional huge arrangement of data records of its client's purchasing practices, require directing data mining on their shared datasets for common advantage. Since these organizations are rivals in the business sector, they would prefer not to uncover a lot about their client's data to each other, however they know the outcomes got from this cooperation could present to them leeway over different competitors.
- Accomplishment of homeland security aiming to counter terrorism relies on upon mix of strength over various mission ranges, viable worldwide joint effort and information sharing to bolster coalition in which distinctive associations and countries must share a few, yet not all, data. Data security in this way turns out to be critical; every one of the gatherings of the joint effort guarantee to give their private information to the cooperation, however neither of them needs each other or some other gathering to learn much about their private data.

### 3.1 Data distribution

**Vertically Partitioning:**

The data for a particular entity are divided across several locations, and every location has information for every single entity for a precise subset of the attributes. We believe that the reality of an entity in a particular location's database may be exposed; it is the values related with an entity that are sensitive. The aim is to cluster the recognized set of common entities without disclosing any of the values that the clustering is based on. e.g insurance companies, hospitals collecting data about the set of people which can be mutually linked. So the data to be extracted is the unit of data at the locations.

**Horizontally Partitioning:**

A situation connecting two parties, both of them owning a database of diverse transactions, in which all the transactions have the equal set of attributes, this situation is also known as a "horizontally partitioned" database. For example supermarkets collecting transaction data of their clients. As a result, the data to be extracted is the unification of the data at the locations.

### 3.2 Privacy - Preserving Tools and Techniques

### 3.2.1 Secure Multi -Party Computation (SMC)

SMC concept was introduced by Yao [19] where he gave a solution to two millionaire's problem. Each of the millionaires wants to know who is richer without disclosing individual wealth. This idea was further extended by Goldreich et al. [20] to multi party computation problem. The aim of a secure multiparty computation task is for the participating parties to securely compute some function of their distributed and private inputs. Each party learns nothing about other parties except its input and the final result of data mining algorithm. As examples consider the scenario where a number of distinct, yet connected, computing devices (or parties) wish to carry out a joint computation of some function. Let *n* parties with private inputs *x*1,…,*x*n wish to jointly compute a function *f* of their inputs. This joint computation should have the property that the parties learn the correct output *y*=*f*(*x*1,…,*x*n) and nothing else, and this should hold even if some of the parties maliciously attempt to obtain more information. The function *f* represents a data mining algorithm that is run on the union of all of the *x*i's.

### 3.2.2 Secure Sum

Distributed data mining algorithms often calculate the sum of values from individual sites. Assuming three or more parties and no collusion, the following method securely computes such a sum.

Let $v = \sum_{i=1}^{s} v_i$ is to be computed for s sites and v is known to lie in the range [0..N]. Site 1, designated as the master site generates a random number R and sends $(R + v_1) \bmod N$ to site 2. For every other site l = 2, 3, 4 … s, the site receives:

$$V = (R + \sum_{j=1}^{l-1} v_j) \bmod N .$$

Site l computes:

$$(V + v_l) \bmod N = (R + \sum_{j=1}^{l} v_j) \bmod N$$

This is passed to site (l+1). At the end, site 1 gets:

$$V = (R + \sum_{j=1}^{s} v_j) \bmod N$$

And knowing R, it can compute the sum v. The method faces an obvious problem if sites collude. Sites (l-1) and (l+1) can compare their inputs and outputs to determine $v_l$ . The method can be extended to work for an honest majority. Each site divides $v_l$ into shares. The sum of each share is computed individually. The path used is permuted for each share such that no site has the same neighbors twice.

### 3.2.3 Digital Envelope

A digital envelope is a random number only known by the owner of private data used to hide the private data.
A set of mathematical operations are conducted between a random number (or a
set of random numbers) and the private data. The mathematical operations could be addition, subtraction, multiplication, etc. For example, assume the private data value is À. There is a random number R which is only known by the owner of À. The owner can hide À by adding this random number, e.g., À + R.

### 3.2.4 RSA Public-Key Cryptographic Algorithm

RSA public-key cryptosystem was named after its inventor, R. Rivest, A. Shamir and L. Adleman. So far, RSA is the most widely used in public-key cryptosystem. Its security depends on the fact of number theory in which the factorization of big integer is very difficult.

In RSA algorithm, key-pair (*e*, *d*) is generated by the receiver, who posts the encryption-key *e* on a public media, while keeping the decryption-key *d* secret.

### 3.2.5 Permutation Mapping Table

For a sequence $d1, d2, \ldots, dn$ , every value is relatively compared with other values of the sequence and if the result is equal or greater than zero the result will be +1 otherwise will be -1 as shown in Table 1, e.g if $d1$-$d2$>= 0 the value in the mapping table is +1 otherwise is -1. So the permutation mapping table of the sequence $d1$,$d2$,…, $d4$ will be as follows:

Table 1: An example of permutation mapping table

| | $d1$ | $d2$ | $d3$ | $d4$ | weight |
|---|---|---|---|---|---|
| $d1$ | +1 | +1 | −1 | −1 | 0 |
| $d2$ | −1 | +1 | −1 | −1 | −2 |
| $d3$ | +1 | +1 | +1 | +1 | +4 |
| $d4$ | +1 | +1 | −1 | +1 | +2 |

The weight for any element in the sequence relative to the others is the algebraic sum of the row Corresponding to that element.

## IV. CONCLUSION

Privacy preserving data mining is an ongoing research area and there are a lot of issues that needs to be addressed. PPDM emerged in response to two equally important needs data analysis in order to deliver better services and ensuring the privacy rights of the data owners. First of all, the databases that are collected for mining are huge, and scalable techniques for privacy preserving data mining are needed to handle these data sources. Secret sharing based methods can be considered one step forward in scalable privacy preserving data mining. This paper is a review of the popular approaches for doing Privacy Preserving Data Mining was presented, namely: distribution, cryptography and summarization. Techniques which minimize the amount of computation and data transfer are needed in highly distributed environments. association rule mining, In Proceedings of the 12th International Workshopon Research Issues in Data Engineering.

## REFERENCES

[1]. Anor F.A. Dafa-Alla, Eun Hee Kim, Keun Ho Ryu, *Yong Jun Heo "PRBAC: An Extended Role Based Access Control for Privacy preserving Data mining" In Proceedings of the Fourth Annual ACIS International Conference on Computer and Information Science (ICIS'05) of IEEE, 2005 .

[2]. Alex Gurevich, Ehud Gudes "Privacy preserving Data Mining Algorithms without the use of Secure Computation or Perturbation" In the proceedings of the 10th international Database Engineering and Applications Symposium (IDEAS'06) IEEE , 2006.

[3]. M. Barni, C. Orlandi, and A. Piva. "A Privacy-Preserving Protocol for Neural-Network-Based Computation". In Proceeding of the 8th Workshop on Multi media and Security, Switzerland, 2006.

[4]. Bunn, P., Ostrovsky, R.: Secure two-party k-means clustering. In: ACM Conference On Computer and Communications Security, pp. 486–497 (2007) .

[5].E. Bertino, I.N. Fovino, L.P. Provenza. A Framework for Evaluating Privacy Preserving Data Mining Algorithms. Data Mining and Knowledge Discovery, 11 (2): pp. 121- 154, 2005.

[6]. Chai Wah Wu IBM T. J. Watson Research Center "Privacy preserving data mining with unidirectional interaction" In the proceedings of the international conference of IEEE, 2005.

[7]. Jha, S., Kruger, L., McDaniel, P.: Privacy Preserving Clustering. In: di Vimercati, S.d.C., Syverson, P.F.,

[8]. Rakesh Agrawal and Rama krishnan Srikant. "Privacy- Preserving Data Mining". In Proceedings of the ACM Special Interest Group on Management of Data Conference (SIGMOD), Dallas, TX, USA, 2000.

[9]. Saeed Samet and Ali Miri. "Privacy-Preserving Protocols for Perception Learning Algorithm in Neural Networks". In Proceeding of The 4th IEEE International Conference on Intelligent Systems (IS), Varna, Bulgaria, 2008.

[10].Yucel Saygin, Vassilios S.Verykios and Ahmed Elmagarmid K. Privacy preserving