Privacy-Preserving Distributed Data Mining Techniques: A Survey

V. Baby
Associate Professor
Dept. of Computer Science and Engineering
VNR Vignana Jyothi Institute of Engg. and Technology,
Hyderabad, India

N. Subhash Chandra, PhD
Principal
Holy Mary Institute of Technology,
Hyderabad, India

ABSTRACT
In various distributed data mining settings, leakage of the real data is not adequate because of privacy issues. To overcome this problem, numerous privacy-preserving distributed data mining practices have been suggested such as protect privacy of their data by perturbing it with a randomization algorithm and using cryptographic techniques.

In this paper, we review and provide extensive survey on different privacy preserving data mining methods and analyses the representative techniques for privacy preserving data mining. We majorly discuss the distributed privacy preservation techniques which provide secure solutions using primitive operations of cryptographic protocols such as secure multi-party computation (SMPC), secret sharing schemes (SSS) and homomorphic encryption (HC).

Keywords
Data mining, K-means clustering, Data privacy, Privacy preserving, Multiparty computation, Threshold Cryptography.

1. INTRODUCTION
Data mining is an as of late rising field, interfacing the three universes of databases, Artificial Intelligence and Statistics. The data age has empowered numerous associations to accumulate extensive volumes of information. Data needed for many crucial data mining tasks is distributed among several parties with different security and privacy concerns. Numerous helpful methods have been produced in this area; that includes clustering, classification, regression, and association rule mining and exception discovery. K-Means clustering is a standout amongst the most broadly utilized procedures for statistical data analysis.

Data mining study deals with the mining of potentially helpful information from huge collections of data with a diversity of applications. Knowledge discovery and Data mining in databases are two novel study areas that examine the automatic extraction of earlier unknown patterns from huge volumes of data. Privacy Preserving Data Mining (PPDM) is a new research path in statistical databases, data mining in which the results are scrutinized for the side effects that acquire in privacy of data. The key reflection in preserving privacy in mining the data is twofold. First, private raw data have to be tailored or trimmed out of the real datasets, such that the receiver of the data will not to be capable to compromise privacy. Second, private information which can be extracted from a datasets by applying data mining algorithms must also be disqualified. The main purpose in PPDM is to build up different techniques for changing the real data in some way, such that the sensitive data and information stay private still after the mining process. On the other hand, they would like to guarantee that their personal data leftovers private. In the other way, there is a necessity to guard sensitive knowledge all through a data mining practice. This problem is known as Privacy-Preserving Data Mining (PPDM).

As contrasting to the centralized model, the Distributed Data Mining (DDM) model accepts that the resources of data are spread across different locales. Algorithms developed inside of this field address the issue of proficiently attainment of the mining results from all the data over these distributed resources. Since the main emphasis is on efficiency, the majority of the techniques developed to date do not obtain security consideration into account. Distribution of data could be partitioned into many parts either vertically or horizontally.

1.1 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.

1.2 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.

1.3 Organization of the paper

The rest of the paper is organized as follows: Section 2 presents relationship between data privacy and data security. Related work discussed in section 3. In section 4, outline of data mining techniques are discussed. The overview of previous privacy preserving data mining techniques is given in section 5. In section 6, we presented framework of distributed privacy preserving data mining techniques. Conclusions are given in section 7.

2. DATA PRIVACY AND SECURITY

In spite of the truth that privacy of data and data security are frequently used as equivalent words, they share all the additional supportive category of relationship. Pretty lot as a home security framework secures the safety and integrity of a family unit, data security policy is put in place to ensure data privacy.

2.1 Data Security

Securing sensitive data is usually known as data security and usually referred to as the availability, confidentiality and integrity of data. That means, it is all of the practices and processes that are in place to guarantee data is not being used or accessed by illegal individuals or parties. Data security guarantees that the data is correct , dependable and accessible when those with permitted access require it. A data security agreement incorporates characteristics, for example, gathering just the required data, do the required computations, maintaining privacy of individual data, keeping it safe, and annihilating any data that is no further required.

2.2 Data Privacy

The privacy of data is suitably defined as the appropriate use of data. Every time organizations and traders use information or data that is specified or endowed to them, the information have to be used by concurred purposes. The standard commission authority's punishments against organizations that have invalidated to assure the security of a client's information. At times, organizations have sold, revealed, or leased volumes of the important data that was depended to them to diverse gatherings without attaining earlier approval.

2.3 The Relationship among Data Privacy and Data Security

Organizations want to endorse a policy of data security for the single purpose of guarantying data privacy or the privacy of their consumers 'data, particularly when it is in use. More so, organizations must assure data security on the basis that the data is a source for the organization. A policy of data security is just the way to the sought end, which is protection of data. In any case, no data security strategy can overcome the ready offer or requesting of the buyer data that was trusted to an organization.

2.4 Data Perturbation

Suppose that a government agency would like to publish a set of electronic health records which may facilitate research. One strategy for protecting the privacy of the individual records is to perturb the original data. Data perturbation procedures are statistically based strategies that try to ensure secret data by adding random noise to private, numerical attributes, thereby shielding the original data. Note that these strategies are not encryption methods, where the information is initially altered, then (regularly) transmitted, and afterward got, "decoded" back to the original information.

2.5 Privacy Preserving

Consider a situation in which more than two parties owning confidential data desire to computes an algorithm on the combination of their inputs without revealing any unwanted information. In the perfect situation every party sends their inputs to the confidential party, who next computes the function and sends the correct results to the other parties without losing privacy of individual inputs. In this way we can preserve privacy even in the presence of adversarial participants that attempt to gather information about the inputs of their parties.

3. RELATED WORK

Privacy-preserving data mining was introduced by Agarwal and Srikant [13] and Y.Lindell and Pinkas [21].The amount of data kept in computer records is growing extensively. There are some methods proposed [13,14,15] by previous researchers in regard to synthesizing, classifying existing privacy-preserving literatures in data mining. E. Bertino [1] anticipated five magnitudes to categorize and analyze privacy-preserving algorithms in data mining with a goal of state-of-the-art. Their categorization dimensions are distribution of data, data modification, data mining algorithm, rule or data hiding and preserving the privacy. According to the characteristics of privacy preservation techniques, these algorithms are mainly divided into three categories: heuristic-based, reconstruction based and cryptography-based. In [1] E. Bertino, provided a complete coverage for preserving the privacy in data mining algorithms, it still has two key drawbacks. Firstly, they did not give us with exact cryptographic techniques used. Secondly, in distributed database scenarios; we typically do not pay too much interest to whether raw data or aggregated data is concealed.

The randomization based techniques for clustering using privacy preserving has been tackled in [3]. In this, the data to be clustered is randomly modified and then clustering is carried out on the tailored data. Cryptography based techniques [16, 17,20] offer high rank of privacy but at the cost of high computation and communication [5]. A broad summary of the connection among the fields of cryptography and PPDM [18, 19] may be found in [9]. The Secure Multiparty Computation (SMC) has been useful for clustering in [6,7,8]. The restriction of these methods is that they have high computational cost and hence their extent is restricted to tiny datasets only. Privacy preserving clustering based on homomorphic encryption is proposed in [6,7]. Authors in [6] and [7] address privacy preserving clustering for arbitrarily-partitioned data for partially truthful two party case models. In [8] addresses analysis and design of privacy-preserving clustering by using k-means algorithm for horizontally partitioned data using oblivious polynomial evaluation and homomorphic encryption. Comprehensive and comparative study of encryption-based method and secret sharing is

specified in [2]. According to [2], secret sharing for PPDM achieves greatest of both worlds i.e. privacy at the level of SMC based techniques and efficiency at the level of randomization based technique. Privacy preserving clustering based on secret sharing has been discussed in [10,11]. In [11] proposed technique based on additive secret sharing for vertically partitioned data using two non colluding third parties to calculate cluster means. In this solution, collusion between two specific parties reveals each entity's distance to each cluster mean.

4. DATA MINING TECHNIQUES

Data mining offers promising approaches to reveal concealed hidden patterns within large amounts of data. The accessibility of new data mining algorithms, in any case, should be met with alert. As a matter of first importance, these procedures are just in the same class as the data that has been gathered. Data mining is an extraction method that aims to locate hidden patterns contained in databases. Data mining area uses a combination of machine learning, statistical analysis, modeling techniques and database technology. The objective is finding patterns, hidden relationships and inferring rules that could give more information about the data and might help for improved future planning. Typical applications comprise market basket analysis, customer profiling, detection of fraud, retail promotion evaluation, and credit card risk analysis.

4.1 Privacy Preserving

Classification makes use of known class labels to direct the objects in the data set. Classification approaches in general use a training set in which all objects are previously related with acknowledged class labels. The classification algorithm learns from the training set and constructs a model. Classification is a typical problem in data mining, which is normally solved by means of decision trees.

4.2 Association Rules

The data mining also deals with the making of association rules, the alteration in confidence and some property of segments; that is, groups or subsets as elements of some parameter of the segment support of the association rule for protecting sensitive rules is completed. A novel idea named not altering the support is projected to hide an association rule. The support of sensitive data not being altered is the primary feature of planned algorithm. The location of the sensitive data is the only thing which alters. The decrease of the confidence of the sensitive rules without alter in the support of the sensitive data is the approach for altering the transaction of the database. One of the way of promotional business development among the organization is sharing the data. Balancing the data privacy as per the valid requirement of the user is the foremost concern. The original data is altered by the sanitization procedure to mask sensitive knowledge earlier than release so the problem can be addressed. Privacy preservation of sensitive knowledge is addressed by several researchers in the form of association rules by suppressing the frequent item sets.

4.3 Clustering

In an unsupervised learning situation, the framework needs to find its own classes and controlled in which it does this to data clustering in the database. The initial measure is to discover subsets of related items and after that discover metaphors which distinguishes all of these subsets. Clustering and segmentation fundamentally partition or metric. If a quantification of similarity is existing, there are amount of techniques to form clusters. Contribution of groups can be founded on the similarity degree in the middle of individuals and with this the standards of enrolment can be characterized. Another methodology is to develop an arrangement of capacities that measure. Clustering according to optimization of set functions is used in data analysis.

K-Mean Clustering:

K-means clustering is one of the most widely used techniques for statistical data analysis. Analysts use cluster analysis to partition the overall public of buyers into business sector portions and to better comprehend the connections between various groups of purchasers/potential clients.

Clustering by K-means is a technique of vector quantization, formerly from signal processing, that is accepted for clustering in data mining. K-means clustering plans to divide n observations into k clusters where in every observation belongs to the cluster with the nearest mean, helping as a model of the cluster.

Algorithm1: K-means clustering

Input: a dataset X = (x1,…, xn) of n observations with p attributes and the number of clusters k.

Output: a clustering of n observations into k groups Cj (j = 1,…, k).

Initialize the k means v1,…,vk

Repeat

$u_j = v_j$ for j = 1,2,…, k

For each observation xi do

assign xi to cluster j such that $\| x_i - u_j \|^2$ is the minimum over all j = 1,…, k.

End for

Compute the new cluster centers v_j for j = 1,2,…, k.

Until convergence

5. PRIVACY-PRESERVING TECHNIQUES

To preserve client privacy in the data mining process, techniques based on random perturbation of data records are used. In privacy preserving data mining (PPDM), the aim is to carry out data mining operations on datasets without revealing the contents of the private data. Since the outputs of the mining notify us something regarding the data, a few information about the real data is revealed to the mining results. This directs to the loss of privacy. If the data is disturbed on the other side for privacy fears, it directs to data loss, which usually refers to the quantity of essential information preserved about the datasets after the perturbation.

5.1 Secure Multi-Party Computation

In the secure multi-party computation, there are two data mining parties, A and B, in possession of databases a and b, respectively. For some data-mining functionality f, they wish to compute f on their joint database; that is, they wish to compute $f(a \cup b)$. Here the union operator $\cup$ is domain-specific. Consider the case of a micro data database, which is a multidimensional database in which each record contains information about a single individual. The two databases a

and b may represent a vertical partitioning (both databases contain the same records, but have different attributes for each record), horizontal partitioning (both databases have the same attributes, but contain different records), or an arbitrary partitioning. If a and b are weighted graphs, then the union could be a graph where each edge is the minimum weight found in a or b.

In a slightly different secure multi-party computation scenario, party A is in possession of algorithm g, and party B holds database y. Together, they want to compute g(y). One of the contributions of this thesis is to show that this type of private algorithm" multi-party computation can be performed efficiently for certain data-mining tasks by trading off increased communication for decreased computation. Within the context of data mining, the secure multi-party computation problem was first investigated by Lindell and Pinkas. It is the most natural extension of the secure multi-party computation (SMPC) paradigm to the problems encountered in data mining.

5.2 Homomorphism Encryption

Various protocols used in privacy-preserving data mining algorithms for instance secure log(x), secure set operations, and secure dot product protocols uses an additive homomorphic encryption. In order to give a reasonable comparison, we also make use of secure additive homomorphic public key cryptosystem in our explanation aligned with malicious adversaries. Let Epk(.) signify the encryption function with public key pk and Dpr(.) signify the decryption function with private key pr. A secure public key cryptosystem is known as additive homomorphic if it convinces the following requirements:

- known the encryption of m1 and m2, Epk(m1) and Epk(m2), there exists an efficient algorithm to compute the public key encryption of m1 + m2, denoted Epk(m1 + m2) := Epk(m1) + Epk(m2)
- known a constant k and the encryption of m1, Epk(m1), there exists an efficient algorithm to compute the public key encryption of km1, denoted Epk(km1) := k × Epk(m1).

5.3 Secret Sharing Scheme

We discuss the most popular development of a (k, n)-threshold scheme, called the Shamir Threshold Scheme [12], designed in 1979. A secret-sharing proposal involves a dealer (owner) who has secret information, a group of n parties (clients), and a group A of subsets of parties (clients) known as the access structure. A secret-sharing method for approved set A is a method where the dealer allocates shares to the parties so that: Any subset in approved set A can restructure the secret from its shares, and any subset which is not in approved set A cannot disclose any incomplete information on the secret. To begin with let P = {P1, P2,...,Pn} be the set of participants, k be the threshold and S b secret. In the Shamir Threshold Scheme Zp is the field, where $p \geq n + 1$ is a prime.

Distribution:

1. Choose t and n where $0 < t \leq n < P$ and $S < P$;
2. Choose randomly t-1 positive integers $(a_1, a_2, ..., a_{t-1}) < p.$
3. Let secret be $a_0 = s.$
4. Construct the polynomial of degree (t – 1),

$$f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + ... + a_{t-1} x^{t-1} \bmod P.$$

5. Let us construct any n points out of it. For $i \in [1, n],$ calculate $(i, f(i))$ and distribute to each player.

Secret Reconstruction:

Given any subset of t points from the selected t participants, the coefficients of the polynomial can be constructed using Lagrange interpolation method in the following:

1. Select t share points like $(x_0, y_0), (x_1, y_1), ..., (x_i, y_i), ..., (x_t, y_t).$
2. Calculate $l_j(x) = \prod_{0 \leq m \leq t, m \neq j} \frac{x - x_m}{x_j - x_m}$, where $l_j(x)$ called Lagrange's coefficients are.
3. Calculate $f(x) = \sum_{j=0}^{t} y_i . l_j(x) \bmod P.$

Hence the secret is the constant term a_0.

Correctness: secret S is reconstructed uniquely by any k or more than k shares from {s1. . ., sn} using Lagrange interpolation.

Privacy: having to any k-1 shares or less than that from {s1. . ., sn} gives no information about the secret S, i.e., the probability of having k-1 shares is independent of secret S.

5.4 Randomization Techniques

The clients can look after the privacy of their data by disturbing it with a randomization algorithm and then submitting the randomized data. This method is called randomization. The randomization algorithm is selected so that aggregate characteristics of the data can be improved with adequate precision, while individual entries are considerably distorted. For the notion of using value distortion to look after privacy to be useful, we require to be able to rebuild the original data distribution.

6. PRIVACY PRESERVING DISTRIBUTED DATA MINING

Lindell and Pinkas [18] first applied the concept of secure computation in the field of data mining and developed a provably secure two-party decision tree over horizontally partitioned data. Since then, privacy preserving distributed data mining has attracted much attention and many secure protocols have been proposed for specific data mining algorithms, including support vector machines, Bayesian network, k-nearest neighbour, k- means clustering, EM-clustering, Regression, association rules mining.

The concepts and techniques in cryptography such as secure multi party computation, homomorphic encryption, secret sharing schemes and oblivious transfer protocol can be applied to the area of privacy preserving distributed data mining.

Framework:
In below figure1, we present a overview of for Privacy Preserving Distributed Data Mining protocols (PPDDM). In this framework, we divided the classification scheme according to which any privacy preserving distributed data mining problems can be categorized and classified

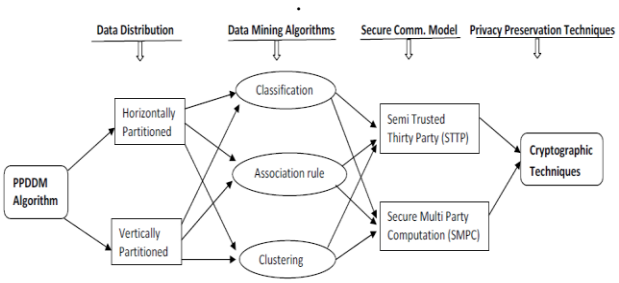

Figure1: PPDDM Protocol Overview

7. CONCLUSION

Researchers use group clustering to partition the common population of consumers into business sectors and to better understand the connections between various groups of purchasers or potential clients. In this paper, we present an efficient survey on privacy preserving distributed data mining. We presented privacy-preserving K-means clustering algorithm, data mining techniques and privacy preserving data mining techniques. Previous research results have revealed diverse approaches to reduce the complexity of privacy-preserving K-means clustering by using semi-trusted third parties.

Further, we propose mechanism where the confidential data of the users, private intermediate values and the ending clustering assignments are confined by means of sharing data in distributed manner. We address this issue and aims to design secure protocols which allow multiple parties to conduct collaborative data mining while protecting the privacy of their data.

8. REFERENCES

[1] 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.

[2] Pedersen, T.B., Saygin, Y., Savas, E.: Secret sharing vs. encryption-based techniques for privacy preserving data mining. In: UNECE/Eurostat Work Session on SDC (2007).

[3] Oliveira, S.R.M.: Privacy preserving clustering by data transformation. In: 18th Brazilian Symposium on Databases, pp. 304–318 (2003)

[4] Vaidya, J., Clifton, C.: Privacy-preserving k-means clustering over vertically partitioned data. In: 9th ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining.ACM Press (2003).

[5] Inan, A., Kaya, S.V., Saygin, Y., Savas, E., Hintoglu, A.A., Levi, A.: Privacy preserving clustering on horizontally partitioned data. Data Knowl. Eng., 646–666 (2007).

[6] Jagannathan, G., Wright, R.N.: Privacy-preserving distributed k-means clustering over arbitrarily partitioned data. In: KDD, pp. 593–599 (2005).

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

[8] Jha, S., Kruger, L., McDaniel, P.: Privacy Preserving Clustering. In: di Vimercati, S.d.C., Syverson, P.F., Gollmann, D. (eds.) ESORICS 2005. LNCS, vol. 3679, pp. 397–417.Springer, Heidelberg (2005).

[9] Pinkas, B.: Cryptographic techniques for privacy-preserving data mining. SIGKDD Explor. Newslett. 4(2), 12–19 (2002).

[10] Upmanyu, M., Namboodiri, A.M., Srinathan, K., Jawahar, C.V.: Efficient Privacy Preserving K-Means Clustering. PAISI 2010. LNCS, vol. 6122, pp. 154–166. Springer, Heidelberg (2010).

[11] Doganay, M.C., Pedersen, T.B., Saygin, Y., Savas, E., Levi, A.: Distributed privacy preserving k-means clustering with additive secret sharing. In: 2008 International Workshop on Privacy and Anonymity in Information Society, Nantes, France, pp. 3–11 (2008).

[12] Shamir, A.: How to share a secret. Communications of the ACM 22(11), 612–613 (1979).

[13] Agrawal, R., and Srikant, R. (2000). Privacy Preserving Data Mining. ACM SIGMOD International Conference on Management of Data, SIGMOD'00, Dallas, USA. 439-450.

[14] G. Aggarwal, N. Mishra, and B. Pinkas. Secure computation of the k-th ranked element. In Proc. Advances in Cryptology – EUROCRYPT 2004, volume 3027 of LNCS, pages 40–55. Springer, 2004.

[15] Kantarcioglu, M., Clifton, C.: Privacy-preserving Distributed Mining of Association Rules on Horizontally Partitioned Data. In: ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD), pp. 639–644 (2002)

[16] Verykios, S., Bertino, E., Fovino, I., Provenza, L., Saygin, Y., Theodoridis, Y.: State of the-art in Privacy Preserving Data Mining. ACM SIGMOD Record 33(1), 50–57 (2004)

[17] Bertino, E., Fovino, I., Provenza, L.: A Framework for Evaluating Privacy Preserving DataMining Algorithms. Data Mining and Knowledge Discovery 11(2), 121–154 (2005).

[18] Lindell, Y. & Pinkas, B. (2002). Privacy Preserving Data Mining. Journal of Cryptology, 15 (3), 177-206. (An extended abstract appeared in Advances in Cryptology, CRYPTO'00. 36-54.)

[19] J. Brickell and V. Shmatikov. Privacy-preserving graph algorithms in the semi-honest model. In Proc. Advances in Cryptology – ASIACRYPT 2005, volume 3778 of LNCS, pages 236–252, 2005.

[20] J.Brickell and V.Shmatikov. Privacy-preserving classifier learning. In Proc. 13th International Conference on Financial Cryptography and Data Security, 2009.

[21] Y Lindell, B Pinkas, Privacy preserving data mining, in CRYPTO '00: Proceedings of the 20th Annual International Cryptology Conference on Advances in Cryptology (Springer, London, 2000), pp. 36–54,2000.

[22] L.Kissner and D. Song. Privacy-preserving set operations. In Proc. Advances in Cryptology – CRYPTO 2005, volume 3621 of LNCS, pages 241–257. Springer, 2005.