Privacy Preserving Distributed Data Mining Techniques

Mayur B Tank
Department of Computer Engineering
L.D. College of Engineering Ahmedabad, India

Tushar A Champaneria
Department of Computer Engineering
L.D. College of Engineering Ahmedabad, India

Abstract

The aim of Privacy Preserving Distributed Data Mining is to extract relevant knowledge from large amount of data while protecting at the same time sensitive information. Due to personal interests, medical databases or business interest privacy is needed. Due to privacy infringement while performing the data mining operation this not often possible to utilize large databases for scientific or financial research. For better decision making we need to perform multi-party computation by combining the database of two or more than two parties, which cannot guarantee security. To address this problem, several privacy preserving data mining techniques are used. In this survey paper, various privacy preserving techniques are reviewed and, also open challenges related to privacy when performing data mining are discussed.

Keywords: Privacy preserving distributed data mining, PPDDM, Secure Data Mining, PPDM, multi-party data mining

I. INTRODUCTION

A. Privacy:

Privacy is the ability to control the view. Being able to control who knows what about us, and when. Threats to privacy due to the Internet, Distributed databases, data mining. Information privacy or data privacy or data protection is the relationship between collection and dissipation of data.

B. Distributed Database:

A distributed database is a database in which storage devices are not all attached to a common processing unit such as the CPU. It may be stored in multiple computers, located in the same physical location or may be dispersed over the network of interconnected computers. A distributed database system consists of loosely-coupled sites that share no physical components. E.g. Cloud Data Distributed databases use client/server architecture to process information request. Two basic schemes for distributed datasets are: homogeneous i.e. Horizontal partition and heterogeneous distribution i.e. vertical partition [1].

C. Data Mining:

Data mining refers to extracting or "mining" knowledge from large amounts of data. The term is actually a wrong name. Usually the Mining of gold from sand or rocks is referred to as gold mining rather than sand or rock mining. So more appropriately the data mining should have been named "knowledge mining from data". There are some other terms which can be used in a similar or slightly different meaning to data mining, like data/pattern analysis, data archaeology, knowledge mining from data, knowledge extraction, and data dredging [2].

The main objective in PPDDM (Privacy Preserving Distributed Data Mining) is to develop algorithms for modifying the original data in some way or to design a new architecture for multi- party data mining, so that the private data and private knowledge remain private even after the mining process.

II. PROBLEM STATEMENT

Different data holders who are located at different places want to undertake a joint data mining task to obtain certain global patterns that will benefit them all while at the same time they each are reluctantly to disclose their private data sets to one another during the execution of the computing. This kind of problem is commonly referred to as privacy-preserving distributed data mining .We consider the following privacy problem: two data owner parties want to collaboratively build a global decision by performing data mining on the union of their database without revealing privacy, and request different degree of privacy protection for sensitive values and non-sensitive values, e.g: Arjun has a private database D1 and Karn has private database D2. Arjun and Karna want to jointly perform the mining to find the association rules on D1∪ D2 without disclosing the contents of their private database to each other.

Let us first take a look at two real-world examples of distributed data mining with different privacy constraints:

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].

III. RELATED WORKS

There are some works by other researchers in regard to synthesizing and classifying existing privacy-preserving data mining literatures. Vassilios S. Verykios, Elisa Bertino and Igor Nai Fovino [4] proposed five dimensions to classify and analyze privacy-preserving data mining algorithms with aims of state-of-the-art. Their classification dimensions were data distribution, data or rule hiding, data modification, data mining algorithm and privacy preservation. Based on their classification dimension, in [4], they proposed classification taxonomy of existing PPDM algorithms. According to the features of privacy preservation solutions, these algorithms were primarily divided into three categories: heuristic-based, reconstruction-based and cryptography-based. The former two categories deal with centralized database and the last one tackled with distributed database. In [5], Xiaodan Wu et al. presented a simplified taxonomy to consolidate the previous one. They analyzed and summarized existing references, thus putting the taxonomy into practical usage

IV. CLASSIFICATION DIMENSIONS OF PPDDM

Here we present a concise classification scheme for Privacy Preserving Distributed Data Mining. In this scheme, four dimensions are identified according to which any privacy preserving distributed data mining problems can be classified and categorized. They are:

- Data Partitioning Model
- Data Mining Algorithms
- Secure Communication Model
- Privacy Preservation Techniques

Here we represent a taxonomy related to PPDDM techniques contained in four levels. It specifically deals with the distributed privacy preserving data mining area in depth. It includes data distribution models of distributed data mining, including vertically partitioned and horizontally partitioned. This Scheme expands cryptographic techniques used in distributed data mining for privacy protecting purpose, like encryption, oblivious transfer, secret sharing etc. Figure 1 shows a general architecture of how these dimensions interrelated to one another.

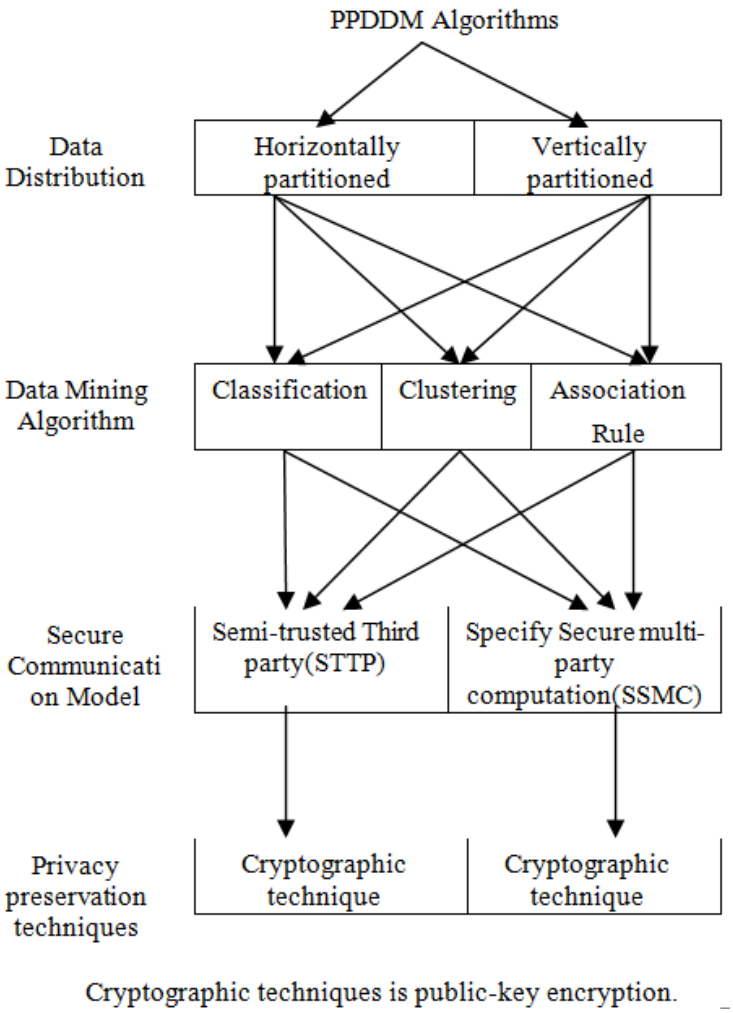

Fig. 1: Classification of PPDDM

A. Data Partitioning Model:

In the scenario of distributed data mining, datasets can be distributed and scattered in different locations in different models. Two basic ways of distribution of datasets are: homogeneous distribution i.e. horizontal partitioning and heterogeneous distribution i.e. vertical partitioning. These two models are formally defined as follows [6]

B. Data Mining Tasks / Algorithms:

The second dimension is data mining algorithms on which privacy preserving techniques are imposed. Generally, the data mining algorithms include association rule mining, clustering and classification .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. There are several different ways to classify a new instance like naïve Bayes classifier [11], decision tree classifier [12] and k-nearest neighbor classifier. 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. For clustering the most commonly used algorithms are k-means clustering and EM clustering.

C. Secure Communication Model:

Here, we present our third classification dimension, secure communication model, which generally refers to the interactive relation of the participants joining in the cooperative computation and the roles they play in the whole process of privacy-preserving distributed data mining tasks. Another similar term is “coopetative” model [7]. This term stems from the word “coopetation”, which was originally employed in social-economics to describe the situation that competing entities producing the same line of products and services have to cooperate with each other to improve the overall value of their market by means of making decisions based on the joint analysis of their private data. Similarly, distributed data mining tasks commonly feature a scenario where all the data holders participating in the joint computation on their individually private data sets naturally have the desire and interest to obtain the final result of the application. As the proprietary owner of their individual data set, it is

understandable that each data holder is reluctant to share private information with other data holders. However, in order to reach the final result of the distributed data mining, they are ready and motivated to provide inputs to the computation, as long as the privacy requirements are met.

Generally, most practical approaches to solve this scenario is to conduct the secure computation at one or more of the participants or at one or more third parties with the assumption that all of participants are semi-honest [8] and the third parties are semi-trusted participants [8]. Herein, let us give the informal definition of both semi-honest and semi-trusted. In [9], a semi-honest party (i.e. honest but curious) follows the rules of the protocol using its correct input, but is free to learn from what it sees during the execution of the protocol to compromise security. In [10], a third party is semi-trusted if it fulfills the following condition: the third party is trusted to provide some commodities or compute intermediate outcome of the computation based on encrypted input it receives; it follows the execution of the protocol correctly, just like all the other users as well, although it tries to learn and deduce some information from its own input and output.

Under such scenario and assumption, 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

D. Privacy Preservation Techniques:

The fourth dimension to classify PPDDM algorithms are the 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 [13][14], etc. All of them serve as the building blocks for other more advanced and high-level protocols, such as privacy-preserving frequency mining, privacy-preserving summation, etc

We present some efficient methods to conduct secure computation in distributed data mining settings. This is by no means an exhaustive list of efficient methods and techniques to achieve privacy preserving data mining protocols. They are, however, sufficient to allow us to present several privacy preserving solutions for distributed data mining problems.

Oblivious transfer: refers to a protocol by which a sender sends some information to the receiver, but remains oblivious as to what is received.

Public-key encryption: an encryption holds the property of additively homomorphic if the functionality of the encrypted values can be obtained by means of the encryption of the addition of the values, i.e. $E(a) * E(b) = E(a + b)$.

Secret sharing: refers to any the method for distributing a secret amongst a group of participants, each of which is allocated a share of the secret. The secret can only be reconstructed when the shares are combined together; individual shares are of no use on their own, i.e. Shamir secret sharing scheme.

Randomization: refers to adding a noise to the original data to hide its real value, thus protecting privacy of the data sets.

V. CONCLUSIONS AND FUTURE WORK

In this paper we present a clear view of current scenario in privacy preserving data mining area. The dimensions we have identified include data partitioning model, data mining algorithms, secure communication model and privacy preservation techniques. We have also defined and demonstrated the meaning of these dimensions in the context of real-world applications. Currently, most cryptographic solutions to PPDDM problems are constructed and analyzed with the assumption of semi-honest model. However, in real world applications, the case of pure semi-honest scenario is rare. Most parties should be regarded as malicious users, that is, they can deliberately provide false information or corrupt the execution of the algorithm. Research work into this area has gained great momentum and requires further efforts to clarify. Cryptography-based SMC has the highest accuracy in data mining and good privacy preservation capability. The Solutions using conventional cryptography methods have failed to defeat the scalability parameter in their performance evaluation.

Theoretical approach is another emerging field that aims to tackle privacy-preserving distributed data mining problems. This area is a very proposing one, the framework of which has been proposed, yet the solution and evaluation work is still open for further investigation [15] , Standardization issues in privacy-preserving distributed data mining cover a wide range of topics, including a common framework of PPDDM with respect to privacy definitions, principles, policies and requirements as well as more effective and precise evaluation metrics regarding efficiency, privacy and complexity of PPDDM algorithms.

REFERENCES

[1] V.S. Verykios, E. Bertino, I.N. Fovino, L.P. Provenza, Y. Saygin and Y. Theodoridis. State-of-the-art in privacy preserving data mining. ACM SIGMOD Record, 33 (1), 2004. (Partitioning DB)
[2] Data mining Concepts & Techniques, Second Edition By: Jiawei Han and Micheline Kanber.
[3] W. Ouyang and Q. Huang. Privacy Preserving Association Rules Mining Based on Secure Two-Party Computation. LECTURE NOTES IN CONTROL AND INFORMATION SCIENCES, pp. 344-969, 2006
[4] 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
[5] V.S. Verykios, E. Bertino, I.N. Fovino, L.P. Provenza, Y. Saygin and Y. Theodoridis. State-of-the-art in privacy preserving data mining. ACM SIGMOD Record, 33 (1), 2004
[6] X. Yi and Y. Zhang. Privacy-Preserving Distributed Association Rule Mining via Semi-Trusted Mixer. In: Data and Knowledge Engineering, vol. 63, no. 2, pp. 550-567, 2007
[7] T. B. Pedersen, Y. Saygm and E. Savas. Secret sharing vs Encryption Based Techniques For Privacy Preserving Data Mining. Joint UNECE/Eurostat work session on statistical data confidentiality,Manchester, United Kingdom, December 17-19, 2007
[8] C. Su and K. Sakurai. Secure Computation Over Distributed Databases IPSJ Journal, Vol. 0 No. 0
[9] J. Vaidya, Y. Michael Zhu and C.W. Clifton. Privacy Preserving Data Mining. Boston, MA: Springer Science + Business Media, Inc., 2006
[10] A.C. Yao. Protocols for Secure Computations. In Proceedings: 23rdIEEE Symposium, on Foundations of Computer Science, pp. 160-164,Chicago, 1982
[11] M. Kantarcioglu and J.Vaidya. Privacy preserving naïve Bayes classifier. for horizontally partitioned data. In IEEE ICDM Workshop on Privacy Preserving Data Mining, Melbourne, FL, pp. 3-9, November 2003.
[12] F. Emekci, O. D. Sahin, D. Agrawal and A.E. Abbadi. Privacy preserving decision tree learning over multiple parties. Data & Knowledge Engineering, 63: pp. 348-361, 2007
[13] M. Kantarcioglu and C. Clifton. Privacy-preserving distributed mining. of association rules on horizontally partitioned data. IEEE Transactions on Knowledge and Data Engineering, 16(9): pp. 1026-1037, 2004.
[14] W. Ouyang and Q. Huang. Privacy Preserving Sequential Pattern Mining Based on Secure Multi-party Computation. Information Acquisition, 2006 IEEE International Conference on, pp. 149-154, 2006
[15] Zhuojia Xu, Xun YiClassification of Privacy- preserving Distributed Data Mining Protocols, 978-1-4577-1539-6/11 IEEE – 2011.