Journal of Theoretical and Applied Information Technology
31st March 2012. Vol. 37 No.2

ISSN: 1992-8645www.jatit.org E-ISSN: 1817-3195

CLASSIFICATIONANDEVALUATION THE PRIVACY PRESERVING DISTRIBUTED DATA MININGTECHNIQUES

[1] SOMAYYEH SEIFI MORADI, [2] MOHAMMAD REZA KEYVANPOUR

[1]Department of Computer Engineering, Qazvin University, Qazvin, Iran

[2]Department of Computer Engineering, Al-Zahra University, Tehran, Iran

E-mail: [1]s_seifi_moradi@yahoo.com, [2]Keyvanpour@alzahra.ac.ir

ABSTRACT

In recent years, the data mining techniques invarious areas have met serious challenges increasingconcernsaboutprivacy. Different techniques and algorithms have been already presented for Privacy preserving data mining (PPDM), which could be classified in two scenarios: centralized data scenario and distributed data scenario. This paper presents a Framework for classification and evaluation of the privacy preserving data mining techniques for distributed data scenario. Based on our framework the techniques are divided intothree major groups, namely Secure Multiparty Computation based techniques, Secret Sharing based techniques and Perturbation based techniques.Also in proposed framework, seven functional criteria will be used to analyze and analogically evaluation of the techniques in these three major groups. The proposed framework provides a good basis for more accuratecomparison of the given techniques to privacy preserving distributed data mining. In addition, this framework allows recognizing the overlapping amount for different approaches and identifying modern approaches in this field.

Keywords: *Privacy Preserving Distributed Data Mining (PPDDM), Secure Multiparty Computation (SMC), Perturbation, Secret Sharing*

1. INTRODUCTION

Rapidandsignificance progressinnetwork, storage and processortechnologies leads tocreation ofultralargedatabases that storean unprecedentedamountofinformation. An issuethatpeoplearefacingisnotsufficientinformation, but how toextractinformationfromthe massive collection of data. Dataminingtechnologytriestorespondtotheseneedsa nd extract unknown patternsandrules. Butinrecent years, increasing concernsaboutprivacyhas ledthe data owners are not willing to sharetheirdataand create a shared data warehouse.Incidence ofsuchproblemsindata collectioncanaffectthe success ofdatamining, thusprotectingdataprivacyis animportant issueindatamining development.

Giventheseproblems,a newtechnologycalledPrivacy Preserving Data mining (PPDM)was introducedwhichaims toachieve validresults of data mining andprovideprivacyrequirementssimultaneously [1]. So far,severalmethodsforprivacy preserving data mining have been raised. Someoftheseapproachesfocus on centralized datascenariowhere the owners or data providers are publishing or sharing their data to acquire data mining results and /or joining the data mining process.

Also, someotherPPDMapproaches have are presented for distributed scenarios. Inthisscenario, there areseveral site, each one owns a part of the private data and wantto computea data mining algorithm on the union of their databases without revealing the data at their individual sites and the results of data mining will only be revealed.Inthissituation, these sitesto achieve thedata miningresultsare workingtogetherif the guarantee is given that their private information will not be disclosed during the mining process.

Depending on how the data is distributed across the sites, distributed data mining algorithms can be divided into two categories: Vertical partitioning and horizontal partitioning. In horizontally partitioned scenario, different sites collect the same set of information but about different entities. In

ISSN: 1992-8645www.jatit.org E-ISSN: 1817-3195

vertically partitioned scenario, although different sites gather information about the same set of entities, they collect different feature sets.

The simplestsolutiontosolvesuchproblems is to use aTrustedThird Party(TTP)which performallcommoncomputationsandalsomaintainse curity, but if nobody can be trusted enoughtoknowallthe inputs, privacywillbecomesaconcern. Privacy Preserving DistributedDataMining(PPDDM)providesmethodst hataim toachieve data mining reliableresultsonadistributeddatasetsand limitation ondatasharingbetweensites.

Thispaperattemptstoprovidea frameworkforclassificationandevaluation of privacy preserving distributeddata mining. The rest of the paper is organized as follows. In section 2, the recommended classification framework for PPDDM will be presentedand then we introduce these techniques. In section 3 we propose the evaluation framework and analyze these techniques under this framework and finally, in section 4, the paper will be finalized by conclusion as well.

2. CLASSIFICATION PPDM TECHNIQUES

Studying and analyzing privacy preserving distributeddata mining techniques indicate that these methods can be classified based on the conditions of privacy protection into three principle groups of Secure Multiparty Computation based techniques, Secret Sharing based techniques and Perturbation based techniques. The proposed classification framework is shown in figure (1).

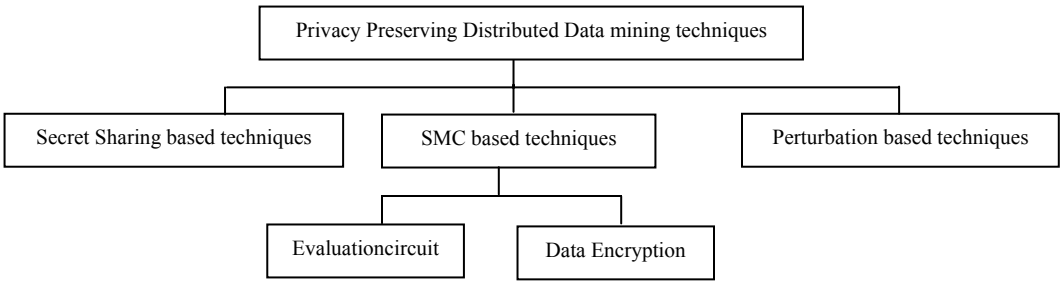

Figure1. PPDDM Techniques classification framework

2.1. Secure Multiparty Computation based techniques

Privacy preserving distributed data mining has similarities with multiparty computation of Cryptography field. The idea of secure two party computations was suggested by Yao [2] in 1986 for the first time. In this idea, every function that it's inputs is distributed between two sites can be calculated securely so that nothing is revealed except that computation results for sites and also no site is informed of the another site's inputs. Later this approach was extended to multiparty computation and become proved that every computation which could form a Boolean circuit size polynomial can be described as safe to be solved [3].

Security Model

There are two security models in SMC approach: semi-honest model and malicious model. In the first model will assume all sites are rules to follow protocol, but the sites are also curious and can use from data that are achieved during the implementation process. In the malicious model, no assumption is made about sites' behavior and in it every site can have thoughts and intentions are malicious. Solutions based on malicious model to model semi-honest harder and more expensive.

SMC Solutions

All the SMC proposed solutions are based one of the tow following models:

1. **Trust model**: The sites without requiring a trusted third party, themselves implement SMC protocols (Figure 2.a).

2. **Ideal model**: The sites for their calculations rely on a trusted third party [5]. In fact, in this model, it is tried to improve performance with creating acceptable compromise in the trust model, on the basis, it is used from semi-trusted non-participation third parties (Figure 2.b).

As in Figure (2) are seen, a series of communications between sites are need to run secure multiparty computation. In SMC in order to

ISSN: 1992-8645www.jatit.org E-ISSN: 1817-3195

preserving the privacy of values that are sent in middle communication, is used from randomization or cryptography approaches.

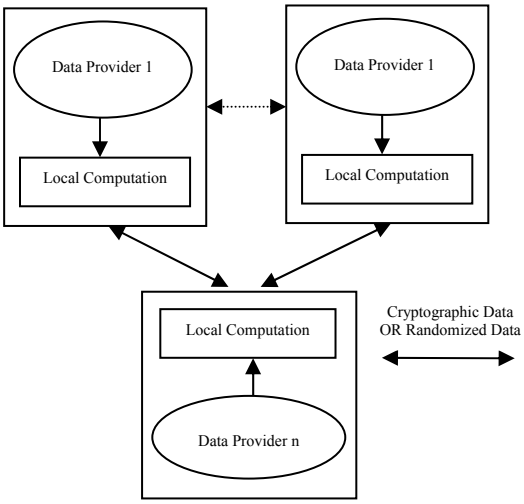

Figure2.a. Trust model based SMC framework

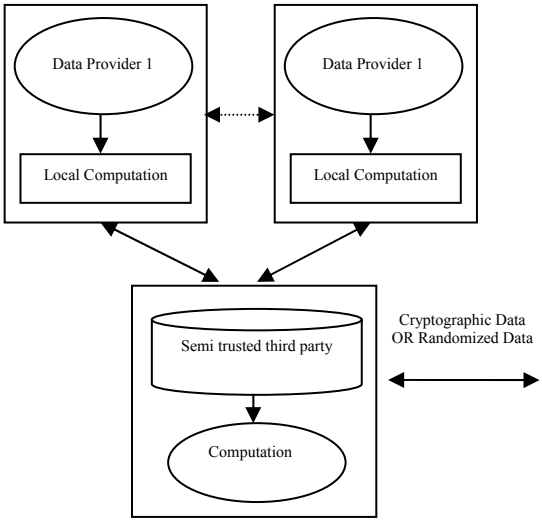

Figure2.b. Ideal model based SMC framework

Cryptography approaches provide solutions with high accuracy and assurance of preserving privacy, but in large scale distributed systems have low efficiency. In contrast, the randomization approaches are more efficient, but suffer from the aspect of preserving balance between privacy and accuracy. In fact, randomization approach is much more efficient but less accurate, while the cryptography approach is less efficient but more accurate [6].

In Figure (3), the comparison of these two approaches in terms of Inefficiency, privacy loss and Inaccuracy in data mining results is shown.

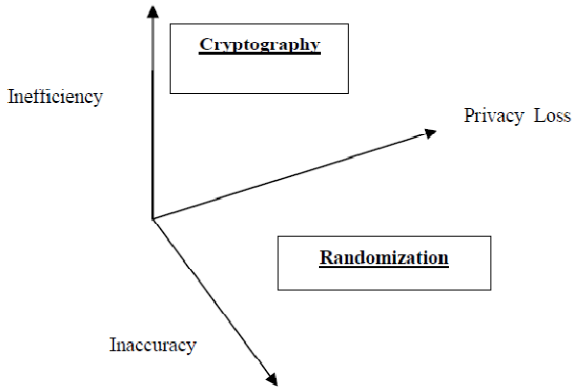

Figure3. Kinds of SMC Solutions [6]

The methods based on secure multiparty computation for PPDDM can be classified into two general classes: techniques based on circuit evaluation and techniques based on data encryption (homomorphic or Commutative encryption). Additively some other tools like Secure Sum, Secure Set, Union, Secure Size of Set intersection, Scalar Product, EM clustering etc can also be employed along with above mentioned approaches to find the SMC solutions [4].

2.1.1 Circuit evaluation

General solution of Yao for SMC is circuit evaluation method.Many of the protocols based on encryption use the idea introduced by Yao. In Yao's protocol one of the parties compute a scrambled version of a Boolean circuit for evaluating the desired function. The scrambled circuit consists of encryptions of all possible bit values on all possible wires in the circuit. The scrambled circuit is sent to the other party, which can then evaluate the circuit to get the final result.

Although Yao' circuit evaluation method is secure, but it poses significant computational problems since the computational complexity of this method depends on input size and then it is expensive, since they require complicated encryptions for each individual bit. Then computational cost of the approach for data mining

ISSN: 1992-8645www.jatit.org E-ISSN: 1817-3195

tasks is very high, so that preclude using this method. Then some PPDDM methods use the idea only as sub-protocols to compute certain simple functions [1, 13].

2.1.2. Data Encryption

Another method for privacy preserving in SMC is processing encrypted data and using homomorphic and commutative properties of encryption systems. As a example, in [10-13, 20] based on homomorphic encryption, solutions for scalar product computation and in [4, 15, 14] based on commutative encryption, solutions for secure sum computation and secure size of set intersection is offered.

A public encryption system $P(G,E,D)$ is a collection of three probabilistic polynomial time algorithms for key production, encryption and decryption. The algorithm of key production $G(r)=(pk,sk)$ based on random argument *r* produce a couple of keys which sk is private key and pk is public key. Everyone can encrypt a massage with public key pk, but just the holder of private key can decrypt the massage. Encryption algorithm E based on plaintext m, random value r and public key as input, produce encrypted text $E_{pk}(m,r)$. The decryption algorithm D based on encrypted text c and private key sk (corresponding public key pk) produce plaintext $D_{sk}(c)$,so that $D_{sk}\left(E_{pk}\left(m,r\right)\right)=m$.

Homomorphic encryption

An encryption system is homomorphic when one can perform a certain algebraic operation on plaintext through an efficient operation on encryptedtext. This characteristic allows a site without having public key, add or multiply on the plaintext with performing the simple computations on encrypted text. Homomorphic encryption systems are a special type of public key encryption systems. As an example, in public key encryption, paillier [18] that is additively homomorphic, the equation 1 holds:

$$\forall m_1,m_2,r_1,r_2 \in Z_\mu: D_{sk}\left(E_{pk}\left(m_1,r_1\right)E_{pk}\left(m_2,r_2\right) \mod \mu^2\right) = m_1+m_2 \mod \mu^2 \quad (1)$$

Commutative encryption

An encryption system is commutative when encryption of a plaintext based on two different keys and excluding of encryption order, produce a similar output, also the encryption function be such that the encrypted text of two different plaintexts is never the same. Also decryption of the encrypted text for retrieving the plaintext takes polynomial time. In other words, the encryption algorithm E is commutative, if for different encryption keys $k_1,\ldots,k_n \in k$, for any *m* in domain M and for any permutation *i,j*, the following two equations hold:

$$E_{k_{i_1}}(\ldots E_{k_{i_n}}(M)\ldots.) = E_{k_{j_1}}(\ldots E_{j_n}(M)\ldots.) \quad (2)$$

$$\forall M_1, M_2 \in M \text{ such that } M_1 \neq M_2 \text{ and for given } k, \grave{o}\frac{1}{2^{\varepsilon}}$$

$$pr(E_{k_{i_1}}(\ldots E_{k_{i_n}}(M_1)\ldots) = E_{k_{j_1}}(\ldots E_{k_{j_n}}(M_2)\ldots.)) < \varepsilon \quad (3)$$

This feature of encryption allows that without revealing the two items, we evaluate whether the two items are equal or not?

2.2. Secret Sharing based techniques

Although Secure multiparty computation techniques provide superior privacy guarantees, but often are very inefficient for using in practical applications, because this techniques are based on primitives based on cryptography which needs considerable cost and since in data mining tasks, this primitives should be implemented many times, in practice, the produced algorithms are not applicable on large data sets.

As is indicated in figure 4, in secret sharing based techniques, data providers, send their data for a collection of third party. It is assumed the third parties would be semi-honest and don't collude with each other. Finally, data miners who can be indistinct entity or one of data provider, produce the desired output. Also in this approach, data providers can participate in distributed data mining with the role of third party.

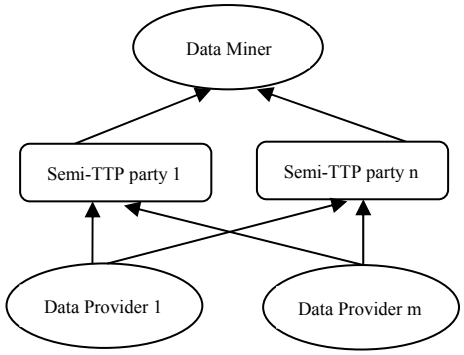

Figure4: PPDDM framework based on the secret sharing model

ISSN: 1992-8645www.jatit.org E-ISSN: 1817-3195

The idea of secret sharing approach [9] is such that every site that has a secret, distribute it between the n-sites, such that no one of the n-site can recover the secret is shared such that the information of at least t-site of n-site is needed for it is recovering.

The scheme secret sharing A (t, n) is a set of two functions of S and R. The function S is a sharing function which takes a secret as input and produces n secret shares in the form of: $S(s)=(s_1,..,s_n)$ .The two functions are selected in the manner that for any collection $I \subseteq \{1,\ldots,n\}$ of t indices, would hold the relation $R\left(I,\ s_{I_1},\ldots,s_{I_t}\right)=S$. In addition, it is necessary that recovering s from a set of t-1 secret shares would be an impossible.

In order that third party can work on secret, they should use from secret sharing schemes which would be homomorphic for the special operations like adding and multiplying, a secret sharing scheme is additively homomorphic when the following relation hold:

$$R\left(I,\ s_{I_1}+s'_{I_1},\ldots,s_{I_t}+s'_{I_t}\right)=S+S' \qquad (4)$$

In [16, 17], solutions are presented based on secret sharing for the purpose of distributed privacy preserving clustering and classification.

Since secret sharing based techniques don't use data encryption, comparing to the secure multiparty computation method are more efficient. One of the disadvantages of Shamir secret sharing scheme is that is not multiplying homomorphic and for performing multiply operation needs a more information exchange between sites.

2.3. Perturbation based techniques

Since perturbation based techniques in comparison to the SMC-based techniques are efficient from the computational aspect, in this approach, it is tried that using combining the advantages of perturbation strategy to achieve a better strategy. The main issue in using perturbation strategy in distributed scenario is that we can unify perturbations that are used in different parties securely. In [7, 8], for producing similar random perturbation amount, it is used from cryptography and SMC-based techniques.

Perturbation-based methods in comparison to the secure multiparty computation-based techniques have low computational cost and from the aspect of participant's site number have high scalability. But in this strategy, the cost of communications and information transitions for data collection is much, because large amounts of perturbed data should be shared.

3. EVALUATING PPDDM TECHNIQUES

At present, the privacy Preserving DistributedDataMining study is in development stage. Then most current PPDDM techniques are on the theory level and are developed for specific applications andagainst some certain aspects. Therefore so far, there is no a technique to effectively achieve the PPDDM goals. So the evaluation framework recommended for assessing and evaluating PPDDM techniques, is in accordance with the following criteria:

- *Efficiency:* is defined based on techniques running time (computational cost) and cost of information exchange between sites (communication cost).
- *Privacy level*:in PPDDM a computation is called secure if the information obtained by any party can be obtained through only its own input and output.
- *Mining accuracy*: is defined based on amount of data mining result accuracy that achieved in PPDDM techniques.
- *Scalability:*scalability of the technique refers to the ability to efficient handle many participant sites, when the number of participant site increases.
- *Security model:*is defined based on assumptions of sites' behavior (semi- honest and malicious model) that is considered in techniques.
- *Applicable areas*: is defined based on appropriate distributed areas that these techniques are applicable.

The framework allows identifying the overlapping amount of different approaches in this field and recognizing the new approaches in the mentioned area. The result of evaluating the PPDDM methods based on the framework is indicated in table 1.

As is indicated in table 1, the main challenge of SMC-based techniques is their non scalability against participant sites number, because in this strategy, the cost of computational and

ISSN: 1992-8645www.jatit.org E-ISSN: 1817-3195

communications is considerably high. Therefore secret sharing and perturbation based approaches are proposed for the purpose of removing these problems and scalable up preserving privacy distributed data mining.

Also, since most the offered techniques in the area are based on the assumption that most sites are semi-honest, so one of the other disadvantages of PPDDM methods is their ineffectiveness against malicious security model.

4. CONCLUSIONS

In this paper, it was tried to offer a framework for classify and evaluating PPDDM techniques. At first, these techniques divided into three approaches of secure multiparty computation, secret sharing and perturbation and then every approach was being investigated. Accordance proposed evaluation framework, the premise of ensuring the privacy of how to plan an effective technique against malicious model and independent from the assumptions, how to further improve the technique efficiency, mining accuracy and scalability against the large distributed environment are directions of the future studies.

	Privacy Preserving Distributed Data Mining Techniques			
Comparison Criteria	based on SMC		Based on secret sharing	Based on perturbation
	Data encryption	Circuit evaluation		
computational cost	High	Very high	Average	Average
Communication cost	High	Very high	Average	High
Privacy level	High	Very high	High	Average
Mining Accuracy	High	High	High	Average
Scalability	Low	Very Low	Average	Average
Security model	Semi-honest	Malicious	Semi-honest	Semi-honest
Applicable area	Small distributed environment	as sub-protocols	Middle distributed environment	Middle distributed environment

TABLE I. *EVALUATION FRAMEWORK OF THE PPDDM TECHNIQUES*

REFERENCES

[1] Stanley R. M. Oliveira1,2 and Osmar R. Zaïane1, "Towards Standardization in Privacy-Preserving Data Mining, In ACM SIGKDD 3rd Workshop on Data Mining Standards, pp. 7–17, 2004.

[2] A. C. Yao, "How to generate and exchange secrets". In Proceedings of the twenty-seventh annual IEEE Symposium on Foundations of Computer Science, IEEE Computer Society, pages 162–167, 1986.

[3]O. Goldreich, S. Micali, and A. Wigderson, "How to Play any Mental Game: A Completeness Theorem for Protocols with Honest Majority," Proc. 19th ACM Symp. Theory of Computing, ACM Press, 1987, pp. 218–229; http://doi.acm.org/10.1145/28395.28420.

[4] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y. Zhu." Tools For Privacy Preserving

ISSN: 1992-8645www.jatit.org E-ISSN: 1817-3195

Distributed Data Mining" . SIGKDD Explorations, 4(2):28-34, 2002

[5] Rebecca Wright, "Progress on the PORTIA Project in Privacy Preserving Data Mining," A data surveillance and privacy protection workshop held on 3rd June 2008.

[6] Dr. D.K.Mishra, P.Trivedi, S.Shukla, "A Glance at Secure Multiparty Computation for Privacy Preserving Data Mining", International Journal on Computer Science and Engineering Vol.1(3), 2009, 171-175

[7]Feng LI, Jin MA, Jian-hua LI," Distributed anonymous data perturbation method for privacy-preserving data mining" , Journal of Zhejiang University SCIENCE A ISSN 1673-565X (Print); ISSN 1862-1775 (Online)

[8] Keke Chen, Ling Liu,"Privacy-preserving Multiparty Collaborative Mining with Geometric Data Perturbation" , IEEE Transactions on Parallel and Distributed Systems, Volume 20 , Issue 12 (December 2009), Pages: 1764-1776 ,ISSN:1045-9219

[9] Adi Shamir." How to Share a Secret". Communications of the ACM, 22(11):512–613, 1979.

[10] Xun Yi, YanchunZhang ,"Privacy-preserving distributed association rule mining via semi-trusted mixer", Data & Knowledge Engineering 63 (2007) 550–567

[11]J.Zhan, S.Matwin, L.Chang , "Privacy-preserving collaborative association rule mining" , 19th Annual IFIP WG 11.3 Working Conference on Data and Applications Security, Nathan Hale Inn, University of Connecticut, Storrs, CT, U.S.A.,August 7-10, 2005.

[12]Sheng Zhong, "Privacy-preserving algorithms for distributed mining of frequent itemsets", Information Sciences 177 490–503(2007).

[13] JaideepVaidya and Chris Clifton." Privacy-preserving k-means clustering over vertically partitioned data". In KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 206–215, New York, NY, USA, 2003. ACM Press.

[14] Murat Kantarcioglou and Chris Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data", In Proceedings of the ACM SIGMOD Workshop on Research Isuues in Data Mining and Knowledge Discovery, 24–31(2002).

[15] J. Vaidya, C. Clifton: "Secure set intersection cardinality with application to association rule mining". Journal of Computer Security 13(4): 593-622 (2005).

[16] Mahir Can Doganay , Thomas B. Pedersen , "Distributed Privacy Preserving k-Means Clustering with Additive Secret Sharing" , PAIS'08, March 29, 2008, Nantes, France.

[17] O.D. Sahin, D. Agrawal, A. El Abbadi , "Privacy preserving decision tree learning over multiple parties ",Data & Knowledge Engineering 63 (2007) 348–361

[18] P. Paillier. "Public-key cryptosystems based on composite degree residuosity classes". In J. Stern, editor, Advances in Cryptology RUROCRYPT'99, volume 1592 of Lecture Notes in Computer Science, pages 223–238, 1999.

[19] J.Zhan, L.Chang and S.Matwin1, "Privacy Preserving K-nearest Neighbor Classification", International Journal of Network Security, Vol.1, No.1, PP.46–51, July 2005 (http://isrc.nchu.edu.tw/ijns/)

[21] J.Vaidya ,M. Kantarcıo ˘glu ,C.Clifton , "*Privacy-preserving Naïve Bayes classification"*, The VLDB Journal 17:879–898 DOI 10.1007/s00778-006-0041-y (2008)