Privacy-Preserving Data Mining (PPDM) Method for Horizontally Partitioned Data

Mohamed A.Ouda ,Sameh A. Salem, Ihab A. Ali, and *El-Sayed M.Saad*

Department of Communication and Computer, Faculty of Engineering

Helwan University, Cairo - Egypt

Abstract

Due to the increase in sharing sensitive data through networks among businesses, governments and other parties, privacy preserving has become an important issue in data mining and knowledge discovery. Privacy concerns may prevent the parties from directly sharing the data and some types of information about the data. This paper proposes a solution for privately computing data mining classification algorithm for horizontally partitioned data without disclosing any information about the sources or the data. The proposed method (PPDM) combines the advantages of RSA public key cryptosystem and homomorphic encryption scheme. Experimental results show that the PPDM method is robust in terms of privacy, accuracy, and efficiency.

1. Introduction:

Data mining is an important tool to extract patterns or knowledge from data [1]. Data mining technology can be used to mine frequent patterns, find associations, perform classification and prediction, etc. The data required for data mining process may be stored in a single database or in distributed resources. The classical approach for distributed resources is data warehouse. Fig. 1 shows a typical distributed data mining approach for building a data warehouse containing all the data. This requires the warehouse to be trusted and maintains the privacy of all parties. Since the warehouse knows the source of data, it learns site-specific information as well as global results. What if there is no such trusted authority? In a sense, this is a scaled-up version of the individual privacy problem; however it is an area where the Secure Multiparty Computation approach is more likely to be applicable.

In this paper, RSA public key cryptosystem and homomorphic encryption are used to develop a reliable privacy-preserving data mining technique for horizontally partitioned data.

The organization of the paper is as follows: Section 2 briefly describes the related work in the area. Section 3 gives background view about the techniques used as well as description of K nearest neighbor classifier as data mining technique. Section 4 presents the proposed algorithm satisfying privacy requirements. Section 5 presents experiments that are carried out to examine the performance of the proposed PPDM algorithm using three different real-world data sets. Section 6 presents a discussion of the experimental results. Section 7 concludes the paper and gives future directions for this research.

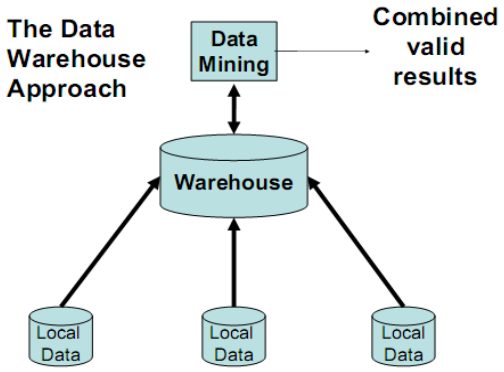

Fig. 1: Data warehouse approach to mining distributed sources

2. 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 information. Many secure protocols have been proposed so far for data mining and machine learning techniques such as [3-6] for decision tree classification, [7-9] for clustering, [10], [11] for association rule mining, [12-14] for Neural Networks, and [15] 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.

3. Background:

This part introduces a brief view about the data mining algorithm used, the form of distributed data as well as the tools and techniques which are used for privacy – preserving during data mining process.

3.1 Data Mining Technique and Distributed data

3.1.1 The k-Nearest Neighbor Classifier:

Standard data mining algorithm K-nearest neighbor classification [16][17] is an instance based learning algorithm that has been shown to be very effective for a variety of problem domains. The objective of k-nearest neighbor classification is to discover k nearest neighbors for a given instance, then assign a class label to the given instance according to the majority class of the k nearest neighbors. The nearest neighbors of an Instance are defined in terms of a distance function such as:
The standard Euclidean distance:

$$D(x_i,x_j) = \sqrt{\sum_{q=1}^{r} \left(a_q(x_i) - a_q(x_j)\right)^2} \qquad (1)$$

Where r is the number of attributes in a record instance x , $a_i(x)$ denote the i^{th} attribute value of record instance x, and $D(x_i, x_j)$ is the distance between two instances x_i, x_j .

3.1.2 Vertically and Horizontally Data Partition

When the input to a function is distributed among different sources, the privacy of each data source comes into question. The way in which the data is distributed also plays an important role in defining the problem because data can be partitioned into many parts either vertically or horizontally [18] .Vertical partitioning of data implies that different sites or organizations gather different information about the same set of entities or people, e.g hospitals and insurance companies collecting data about the set of people which can be jointly linked. So the data to be mined is the join of data at the sites. In horizontal partitioning, the organizations collect the same information about different entities or people. As an example supermarkets collecting transaction information of their clients. As a result, the data to be mined is the union of the data at the sites.

In this paper it is supposed that all organizations or departments that to be mined have the same information (homogenous) but different entities (records or tuples), so horizontal approach is conducted.

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 an example consider the scenario where a number of distinct, yet connected, computing devices (or parties) wish to carry out a joint computation of some function [21]. Let n parties with private inputs $x_1,\ldots,x_n$ wish to jointly compute a function f of their inputs. This joint computation should have the property that the parties learn the correct output $y=f(x_1,\ldots,x_n)$ 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 Digital Envelope

A digital envelope [22] is a random number (or a set of random numbers) 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.

3.2.3 RSA Public-Key Cryptographic Algorithm

RSA public-key cryptosystem was named after its inventor, R. Rivest, A. Shamir and L. Adleman [23]. 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.

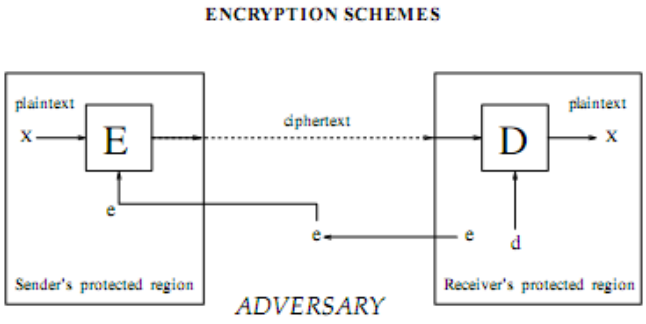

Fig. 2 Public-key encryption schemes: an illustration

3.2.4 Homomorphic Encryption and Decryption Scheme

A cryptosystem is homomorphic [24] with respect to some operation $*$ on the message space if there is a corresponding operation $*'$ on the ciphertext space such that $e(m) *' e(m') = e(m * m')$.

In this part, it is proposed an additively homomorphic encryption and decryption scheme, which is as follows [25]:

Encryption Algorithm

1) The algorithm uses a large number N, such that $N = P \times Q$, where P and Q are large security prime numbers.

2) Given X, which is a plaintext message, the encrypted value is computed:

$$Y = E_p(X) = mod\big((X + P \times R), N\big) \qquad (2)$$

Where mod ( ) is a common modulo N – operation, R is a random number within the uniform distribution $(1, Q)$.

Decryption Algorithm

Given y, which is a cipher text message, we use the security key p to recover plaintext

$$X = E^{-1}(Y) = D_p\,(Y) = mod(Y, P) \qquad (3)$$

$$,Y = mod\big((X + P \times R), N\big)$$

Note that: for any X although $E_1(X) \neq E_2(X)$, $D\big(E_1(X)\big) = D\big(E_2(X)\big)$ which means there is one to many relationship between plaintext X and ciphertext $E(X)$.

3.2.5 Permutation Mapping Table

For a sequence d_1,d_2,…, d_n, 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 d_1-d_2>= 0 the value in the mapping table is +1 otherwise is -1. So the permutation mapping table of the sequence d_1,d_2,…, d_4 will be as follows:

Table 1: An example of permutation mapping table

	d_1	d_2	d_3	d_4	weight
d_1	$+1$	$+1$	-1	-1	0
d_2	-1	$+1$	-1	-1	-2
d_3	$+1$	$+1$	$+1$	$+1$	$+4$
d_4	$+1$	$+1$	-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.

4. Proposed Algorithm

1- In this paper, a semi-honest model for adversary is used, where each party follows correctly the protocol of secure computing function but curiously try to infer data about other parties. A key result which is also used in this work is the composition theorem. We state it for the semi-honest model.

Theorem (1): "Suppose that g is privately reducible to f and that there exists a protocol for privately computing f. Then there exists a protocol for privately computing g". Loosely speaking the composition theorem states if a protocol consists of several sub-protocols, and can be shown to be secure other than the invocations of the sub-protocols, if the sub-protocols are themselves secure, then the protocol itself is also secure. A detailed discussion of this theorem, as well as the proof, can be found in [26].

2- The proposed algorithm presents a method for privately computing data mining process from distributed sources without disclosing any information about the sources or their data except that revealed by final classification result. The proposed algorithm develops a solution for privacy–preserving k-nearest neighbor classification which is one of the commonly used data mining tasks.

The proposed algorithm determines which of the local results are the closest by identifying the class of minimum weight using K nearest neighbors. We assume that attributes of the instance needed for classification are not private (the privacy of the query instance is not protected). Therefore, it is necessary to protect the privacy of the data sources i.e. a site / party S_i is not allowed to learn anything about any of the data of the other parties, but is trusted not to collude with other parties to reveal information about the data.

3- The idea of the proposed algorithm is based on finding K-nearest neighbors of each site, then scramble and encrypts the local $d_{i\,min}$ with homomorphic encryption and its class y_i with the public key e_i sent from Encryption Decryption Management Server (*EDMS*). The results from all sites are combined to produce the permutation table at *EDMS* and instance with minimum weight with its class is determined as the class

of querying instance which is transferred to querying site. Each site learns nothing about other sites. Since the KNN algorithm executed locally for every site S_i.

The standard data mining algorithm is K nearest neighbor for each site / party S_i will be as follows:

1- Determine the parameter K=number of nearest neighbors beforehand.
2- Calculate the distance between the query instance and all the training samples using Euclidean distance algorithm.
3- Sort the distances for all the training samples and determine the nearest neighbor based on the K[th] minimum distances.
4- Since this supervised learning, get all the classes of training data for the sorted value which falls under K.
5- Use the majority of nearest neighbor as the prediction value.

Notations:

$E(x)$ means to encrypt data x using a special encryption algorithm E. $E_k(x)$; refers to encrypting data x using a special algorithm E with a specified key k.

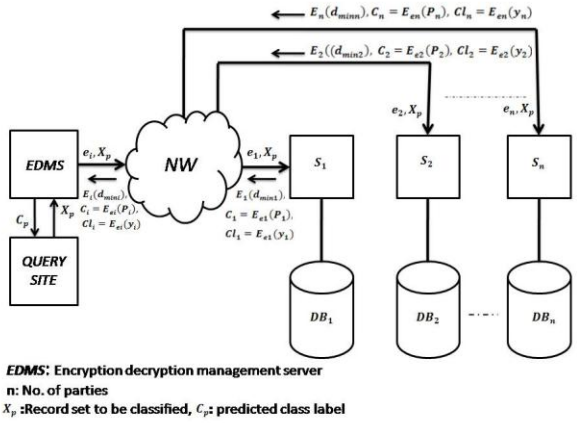

Fig. 3 PPDM K Nearest Algorithm :an illustration

The Integrated PPDM Algorithm of K Nearest Classifier is as follows:

1- **Require**: m parties, y_i class values, l attribute values, X_p query instance $\{x_1, x_2, \ldots x_l\}$
2- P_i and Q_i are large security prime numbers. $N_i = P_i \times Q_i$
3- (e_i, d_i) represent the encryption and decryption keys of RSA algorithm are generated at Encryption Decryption Management Server (*EDMS*).
4- $d_{i\,min}$ represents minimum neighbor distance with majority class relative to query instance X_p, and class label y_i is the corresponding class of $d_{i\,min}$
5- ***For i =1 … m do*** // generating encryption-decryption keys
6- *EDMS* generates (e_i, d_i) using RSA Algorithm ;
7- Transport e_i to Party S_i;
8- ***End For*** // generating encryption-decryption keys
9- ***For i =1 … m do*** // scan m parties, computing $d_{i\,min}$ and encryption process
10- Party S_i locally computes $d_{i\,min}$ and its class value Cl_i according to K nearest algorithm relative to query instance X_p .
11- Encrypt $d_{i\,min}$ as in Eq. (2) to get homormophic encryption $E_i(d_{i\,min})$.
12- RSA encrypts P_i to $C_i = E_{ei}(P_i)$ & class label y_i to $Cl_i = E_{ei}(y_i)$;
13- Transport $E_i(d_{i\,min})$, C_i, and Cl_i to *EDMS*;
14- ***End For*** //computing $d_{i\,min}$ and encryption process
15- ***For i =1 … m do*** // Decryption process at *EDMS*
16- Decrypt $E_i(d_{i\,min})$ as per Eq. (3) to get $d_{i\,min}$ and Cl_i to get y_i
17- ***End For*** //decryption process

18- Construct the mapping table that maps the relative difference between $d_{i\ min}$ with all $d_{j\ min}\{\ i \neq j\ \&\ i,j \in (1,m)\}\ to\ +1,-1$
19- Calculate the weight for each row in the mapping table by adding the row elements and get the sum.
20- Determine the global min distance which corresponds to min weight in the mapping table.
21- Get the predicted class that match global min distance (min weight in the mapping table).

5. Experimental Results

Three real-world datasets are used to examine the reliability of the proposed algorithm. Table 2 shows details of the datasets [27]. The proposed algorithm is developed using C# standard Edition 2010 on Intel® Core2 Duo, 2.0 GHz, 3 GB RAM system. The accuracy of the K nearest neighbor classifier for the three different data sets is shown in Fig.4 .

Table2: Data Sets

Data Set Name	Attribute Characteristics:	Number of Instances:	Number of Attributes:	Area:
Adult	Categorical, Integer	48842	14	Social
Breast Cancer	Real	699	10	Life
Heart Spect	Integer	267	44	Life

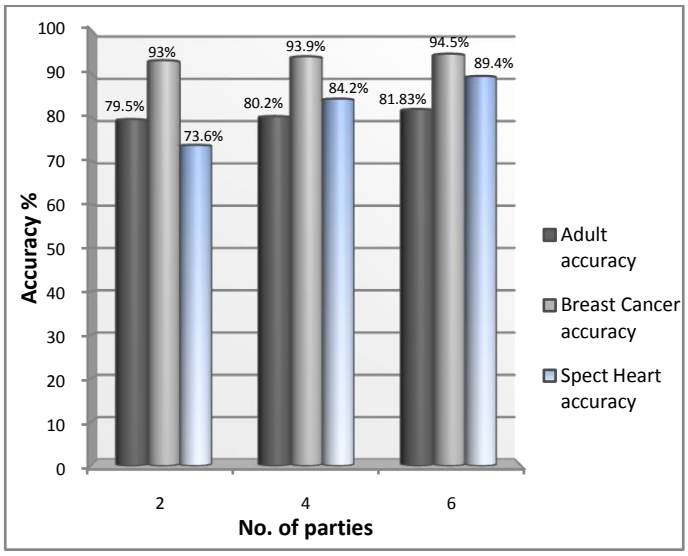

Fig. 4 The accuracy of distributed KNN classifier for three data sets

Table 3: Adult data set (size is order of thousands) (K=3)

No of Parties	2	4	6
Accuracy %	79.5	80.20	81.83
Performance time(msec)	960	1114	1222
Records size(training set)	2000	4000	6000

Table 4: Breast Cancer data set (size is order of hundreds) (K=3)

No of Parties	2	4	6
Accuracy %	93.0	93.9	94.5
Performance time(msec)	477	488	601
Records size(training set)	200	400	600

Table 5: Spect. Heart data set (size is order of tens)(K=3)

No of Parties	2	4	6
Accuracy %	73.6	84.2	89.4
Performance time(msec)	462	469	619
Record size(training set)	40	80	120

6. Discussion

The purpose of privacy-preserving data mining is to discover accurate, useful and potential patterns and rules and predict classification without precise

access to the original data. Therefore, evaluating a privacy-preserving data mining algorithm often requires three key indicators, such as privacy (security), accuracy and efficiency.

Privacy: In the proposed PPDM algorithm, cryptogram management at different levels was adopted.

- **First**, Party S_i encrypts $d_{i\,min}$ with homomorphic encryption , and R_i is a random number within $(1, Q_i)$,used as digital envelope for $d_{i\,min}$.
  $E_i(d_{i\,min}) = mod\big((d_{i\,min} + P_i \times R_i), N\big)$;
- **Second**, The corresponding class y_i of $d_{i\,min}$ is encrypted with RSA public key encryption as well as the prime number P_i
  $Cl_i = E_{ei}(y_i)$; // Cl_i cipher encryption of class label
  C_i=$E_{ei}(P_i)$; // C_i cipher encryption of prime number P_i, e_i public key encryption

Since RSA public key encryption is semantically secure; hence, each party is semantically secure where no party can learn about private data of other parties except its input and the final result. As privacy is preserved for each party, applying the composition theorem (theorem 1), then the total proposed PPDM algorithm is secured.

Accuracy: *EDMS*, which decrypts, $E_i(d_{i\,min})$ and its class label cipher $Cl_i = E_{ei}(y_i)$, and produce accurate results with RAS and homomorphic cryptosystem. As shown in tables 3, 4, and 5, the accuracy of the classifier for parties between 2 to 6 is 73.6 – 94.5 % which is comparable to accuracy of classical approach. As in Fig. 4 the accuracy is varied according to data set size and number of parties but accuracy range is still accepted and as long as the number of parties increases the accuracy is better.

Efficiency: Raising efficiency of the algorithm is mainly shown the decreases in time complexity. PPDM-KNN algorithm reduces the time complexity mainly in two aspects.

- ***First***: global K-distances are quickly generated, since the KNN algorithm executed locally for every site S_i, this enables solutions where the communication cost is independent of the size of the database and greatly cut down communication costs comparing with centralized data mining which needs to transfer all data into warehouse data to perform data mining algorithm.
- ***Second***: Site S_i only have to encrypt encryption parameter P_i of homomorphic encryption system and class label y_i of $d_{i\,min}$ with public key e_i of RSA. So, the algorithm avoids numerous exponent operations and improves the speed of operation greatly. Tables 3, 4, and 5 show that the maximum performance time is 1222 msec for training set of size 6000 records.

These results show that privacy of the data sources is preserved while there is no information loss with accurate results.

Conclusion

In this paper, a privacy-preserving distributed KNN mining algorithm has been presented. As demonstrated, the proposed algorithm is based on the technology homomorphism and RSA encryption which is semantically secured. Moreover, no global computations at the centralized site are conducted but the KNN algorithm is computed locally for each site and local

results are transferred to the centralized site to be compared. Experimental results show that PPDM has good capability of privacy preserving, accuracy and efficiency, and relatively comparable to classical approach.

References

[1] Ming-Syan Chen, Jiawei Han, and Philip S. Yu, "Data mining: an over view from the database perspective", in IEEE Transactions on Knowledge and Data Engineering, vol 8, NO 6, december 1996 .

[2] Jian Wang, Yongcheng Luo, Yan Zhao and Jiajin Le, "A Survey on Privacy Preserving Data Mining" ,in IEEE, 2009 First International Workshop on Database Technology and Applications.

[3] Rakesh Agrawal and Rama krishnan Srikant. "Privacy-Preserving Data Mining". In Proceedings of the ACM Special Interest Group on Management of Data Conference (SIGMOD), pages 439–450, Dallas, TX, USA, 2000.

[4]Yehuda Lindell and Benny Pinkas. "Privacy Preserving Data Mining". In Proceedings of the 20th Annual International Cryptology Conference (CRYPTO), pages 36–54, Santa Barbara, CA, USA, 2000.

[5] JaideepVaidya and Chris Clifton. "Privacy-Preserving Decision Trees over Vertically Partitioned Data". In Proceedings of the 19th Annual IFIP WG 11.3 Working Conference on Data And Applications Security (DBSec), pages 139–152, Storrs, CT, USA, 2005.

[6] Ming-Jun Xiao, Liu-Sheng Huang, Yong-Long Luo, and Hong Shen. "Privacy Preserving ID3 Algorithm over Horizontally Partitioned Data". In Proceedings of the Sixth International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT), pages 239–243, Dalian, China, 2005.

[7] Geetha Jagannathan and Rebecca N. Wright. "Privacy-Preserving Distributed k-Means Clustering over Arbitrarily Partitioned Data". In Proceeding of the eleventh ACM SIGKDD International conference on Knowledge discovery in data Mining (KDD), pages 593–599, Chicago, IL, USA, 2005.

[8] Somesh Jha, Louis Kruger, and Patrick Mc Daniel. "Privacy Preserving Clustering". In Proceedings of the 10th European Symposium On Research In Computer Security (ESORICS), Pages 397–417, Milan, Italy, 2005.

[9] Jaideep Vaidya and Chris Clifton. "Privacy-Preserving k-Means Clustering over Vertically Partitioned Data". In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (KDD), pages 206–215, Washington, DC, USA, 2003.

[10] C.Clifton, M. Kantarcioglu, and J. Vaidya. "Defining Privacy For Data Mining". In Proceedings of the National Science Foundation Work shop on Next Generation Data Mining (NGDM), pages 126–133, Baltimore, MD, USA, 2002.

[11] Murat Kantarcioglu and Chris Clifton. "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data". IEEE Transactions on Knowledge and Data Engineering, 16 (9) : 1026–1037, 2004.

[12] 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, pp. 146–151, Geneva, Switzerland, 2006.

[13] 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), pages 10–65–10–70, Varna, Bulgaria, 2008.

[14] Jimmy Secretan, Michael Georgiopoulos, and Jose Castro. "A Privacy Preserving Probabilistic Neural Network for Horizontally Partitioned Databases". In Proceedings of the International Joint Conference on Neural Networks (IJCNN), pp. 1554–1559, Orlando, FL, USA, 2007.

[15] Zhiqiang Yang and Rebecca N. Wright. "Privacy-Preserving Computation of Bayesian Networks on Vertically Partitioned Data". IEEE Transactions on Knowledge and Data Engineering, 18 (9): 1253–1264, 2006.

[16] Nitin Bhatia and Vandana , "Survey of Nearest Neighbor Techniques", in (IJCSIS) International Journal of Computer Science and Information Security, Vol. 8, No. 2, 2010

[17] T.Cover and P. Hart. "Nearest neighbor pattern classification". In IEEE Transaction of Information Theory, Vol. 13, pp.21-27, January, 1968.

[18] J.Vaidya, C.Clifton "Privacy-Preserving Data Mining: Why, How, and When". IEEE Computer Society, pp. 19-27, November/December 2004.

[19] A.C.Yao, "protocol for secure computations", in proceedings of the 23rd annual IEEE symposium on foundation of computer science, pp. 160-164, Nov.1982.

[20] O. Goldreich, S. Micali and A. Wigderson, "How to play any mental game", In proceedings of the 19th annual ACM Symposium on Theory of Computation, pp. 218-229, May 1987.

[21]Yehuda Linedell and Benny Pinkas, "Secure Multiparty Computation for Privacy-Preserving Data Mining", The Journal of Privacy and Confidentiality , Number 1, pp. 59-98, 2009.

[22] R. Rivest, A. Shamir and L. Adleman. "A Method for Obtaining Digital Signatures and Public-Key Cryptosystems". Communications of the ACM, 21 (2), pp. 120-126, February 1978.

[23] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes", Advances in Cryptography - EUROCRYPT '99, pp. 223-238, Prague, Czech Republic, 1999.

[24] R.Rivest, L. Adleman, and M. Dertouzos. "On data banks And privacy homomorphisms". In Foundations of Secure Computation, eds. R. A. DeMilloetal., Academic Press, pp. 169-179.,1978.

[25] GuiQiong, Cheng Xiao-hui. "A privacy – preserving distributed for mining association rules". International Conference on Artificial Intelligence and Computational Intelligence, pp. 294-297, 2009.

[26] O. Goldreich., "Secure multi-party computation" ,Sept.1998.(working draft).

[27] Ronny Kohavi and Barry Becker "UCI Repository of Machine Learning Databases", Available at http://archive.ics.uci.edu/ml/datasets.html , Data Mining and Visualization , Silicon Graphics., 1996