Research Trends in Privacy Preserving in Association Rule Mining (PPARM) On Horizontally Partitioned Database

[1] Rachit Adhvaryu, [2] Nikunj Domadiya

[1] PG Student, [2] Professor

[1] Computer Science & Engineering, B. H. Gardi College of Engineering & Technology, Rajkot, Gujarat, India.

[2] Computer Science & Engineering, B. H. Gardi College of Engineering & Technology, Rajkot, Gujarat, India.

[1] rachit.adhvaryu@yahoo.com, [2] nhdomadiya@gardividyapith.ac.in

***Abstract* - The advancement in data mining techniques plays an important role in many applications. In context of privacy and security issues, the problems caused by association rule mining technique are investigated by many research scholars. It is proved that the misuse of this technique may reveal the database owner's sensitive and private information to others. Many researchers have put their effort to preserve privacy in Association Rule Mining. In this paper, we have presented the survey about the techniques and algorithms used for preserving privacy in association rule mining with horizontally partitioned database.**

***Keywords* - Data Mining, Horizontally Partitioned Database, Privacy Preserving Association Rule Mining**

I. INTRODUCTION

Data mining or knowledge discovery techniques such as association rule mining, classification, clustering, sequence mining, etc. have been most widely used in today's information world [1]. Successful application of these techniques has been demonstrated in many areas like marketing, medical analysis, business, Bioinformatics, product control and some other areas that benefit commercial, social and humanitarian activities. Privacy Preserving Data Mining is an important feature which every mining system must support. This feature actually secures the private and sensitive information which the database owners do not want to reveal. The sensitive data can be anything like Identification Number, Name, Address, Disease etc. [2]

The works required in the privacy preserving data mining area are as follows:

a) *Privacy Preserving Data Publishing*: These techniques try to study different techniques associated with privacy. These techniques consist of:
   i. The Randomization Method: In this technique, any random value is added to the original value of the data to mask the values of the data. The noise is added in large amount so that the original data value is not recovered [3].
   ii. The K-Anonymity Model and L-Diversity: In K-Anonymity, the techniques like generalization and suppression were introduced to normalize data representation. In order to reduce the identification risk, every tuple in the database must be indistinguishable. The L-Diversity method was introduced to overcome some weaknesses of K-Anonymity. The new concept of intra group diversity of sensitive and private values within anonymization scheme was discovered [4].
   iii. Distributed Privacy Preserving: Sometimes, some users do not wish to disclose their information to other users. But the individual users are interested in achieving the aggregate results from the data set which are divided among the users. [5]

b) *Modifying the record values to preserve privacy*: In this technique, Association Rule Hiding methods were used to preserve privacy. Using these methods, the association rules are encrypted in order to secure the data.

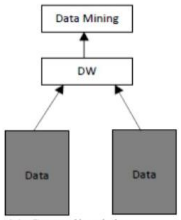
c) *Query Auditing*: In this technique, either the result of the query is modified or the result of the query is restricted. Many Perturbation methods are used to achieve this. [6]

These techniques have been implemented either on Centralized Database or Distributed Database. The detailed description has been described in chapter II.

II. CENTRALIZED AND DISTRIBUTED DATABASE

A. Centralized Database

In the centralized database, all the datasets are collected at one central site which can be called as Data Warehouse and then all the mining operations are performed. Fig [a] shows the centralized database.



(a) Centralized data

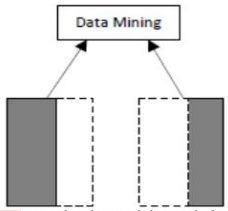
The various techniques used in centralized database are Data Perturbation, Data Blocking and Reconstruction Based Techniques [7].

B. Distributed Database

Sometimes, some users do not wish to disclose their information to other users. But the individual users are interested in achieving the aggregate results from the data set which are divided among the users. Distributed Database is used now a days. Due to large and fast growing database, the data are not managed centrally. But they are stored at different places i.e. different data stored in different places.

The Distributed Database can be further classified into: [8]

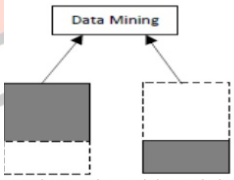
a. *Vertically Partitioned Database*: In this every site has different schemas. The attribute values may or may not be same. Fig [b] indicates Vertical Partitioned Database



(b) Vertical partitioned data

For example, Hospital may have records of the patient like name, contact details, disease, attending doctor, bill amount, etc. Also same name and contact details with any other information like mediclaim amount, insurance id etc. may be found with the insurance company. Thus, at the end one final result can be obtained by combining both the records [9].

b. *Horizontally Partitioned Database*: In this every site has the same schema. But the attribute values are different [10]. Fig [c] indicates Horizontal Partitioned Database.



(c) Horizontal partitioned data

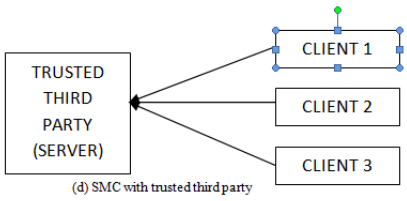
It means the records may be different, but the way the records are stored is same. For example, Credit card systems of two different banks have the same schema of storing the information of the users holding the credit card. But, the information of the users like user name, contact details, credit limit, etc. may be different with both the banks. [10]

In the later part of paper, we describe broad description about the techniques and algorithms used for Privacy Preserving on Horizontally Partitioned Database.

III. PRIVACY PRESERVING APPROACHES FOR HORIZONTALLY PARTITIONED DATABASE

1. SECURE MULTIPARTY COMPUTATION (SMC) WITH TRUSTED THIRD PARTY

It was Client – Server Architecture bases technique where one party is a client and other parties are clients. All the client parties believe the server party to be trusted and honest such that the server party won't reveal their sensitive and private data to another party [11]. Fig [d] shows SMC with trusted third party.



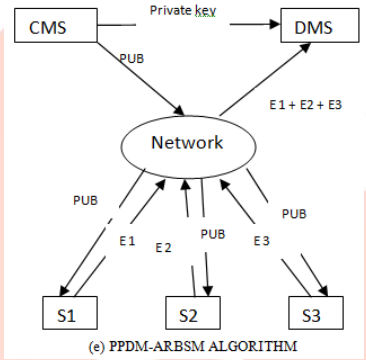
(d) SMC with trusted third party

In this, each party finds the frequent itemset and its local support count and sends it the third party. On receiving these data, the third party evaluates this data to find global frequent itemset and global support count. The result found is returned back to all the client sites for further manipulations [11].

The limitation of this technique was what if the third party fails or collusion between third party and any client [11]. There were more chances of data loss in this technique. Algorithm describing this technique is as below

A. PPDM-ARBSM ALGORITHM

In Privacy Preserving Distributed Mining algorithm of Association Rules (PPDM-ARBSM), the advantages of RSA Public Key encryption was used [12]. The algorithm used mainly 2 types of servers. CMS (Cryptosystem Management Server) and DMS (Data Mining Server). The task of the CMS was to provide public key and private key for encryption and the task of DMS server was decryption and final result generation. Fig [e] shows PPDM-ARBSM.



(e) PPDM-ARBSM ALGORITHM

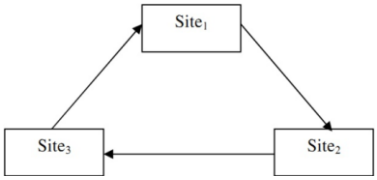
The working of the algorithm was as follows:

1. Firstly, Identity of CMS is authenticated.
2. Secondly, CMS generates Public Key (pub) for RSA and sends to each site (S1, S2, S3). Also CMS generates and sends a Private Key to DMS [12].
3. At each site in a communication network, for a transaction D, frequent itemset (Fi) is generated. Thereafter, Fi is encrypted using public key and sent to DMS.
4. Once DMS receives all the data, it decrypts it using the private key, evaluates the data and generates the final result. This result is sent back to all sites [12].

The drawback of this algorithm was the use of communication network. If this network fails, whole data may be lost and no global result would be generated.

2. SECURE MULTIPARTY COMPUTATION (SMC) WITH SEMI-HONEST MODEL

This technique is quite different than SMC with Trusted Third Party Model. A partially-honest party is one who follows the standard rules. But feels free to migrate in between the steps to gain more information and satisfy an independent agenda of interests [13]. In other words, a partial-honest party follows the rules step by step and computes exactly required values based on the input from the other parties and it can analysis other party's data. But it is sure that it will not insert any false value which results in failure and try to use all the information secured to know sensitive information of other parties. The collusion among the parties does not occur in this model and thus private data is not revealed [13]. Fig [f] shows SMC with the semi honest model.



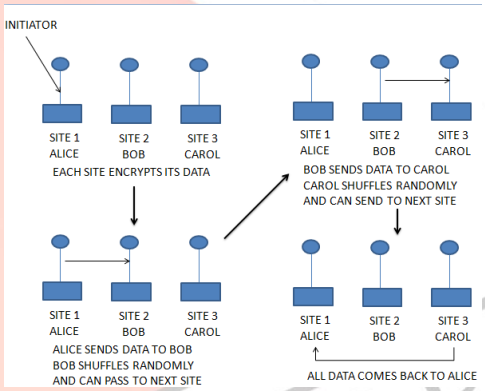
(f) SMC with semi honest model

- Each site assumes the other party to be honest.
- Each site follows the protocol
- Each site computes only the required data.
- Sites do not collude with each other. But try to find some information about other site [13].

The drawback of the model was that each site is assumed to be honest. But there is no surety about sites not colluding with another site. Few algorithms describing this technique are explained below.

A. FAST PRIVATE ASSOCIATION RULES MINING FOR SECURELY SHARING

In this model, it was assumed that each site follows semi-honest protocol. The sites can share the union of data without the use of any trusted third party. The information which is hidden is the records and the position of the records in the data set. The whole model is governed by one site which is either called Model Driver or Model Initiator [14]. The Fig [g] shows the working of this model and is described as:

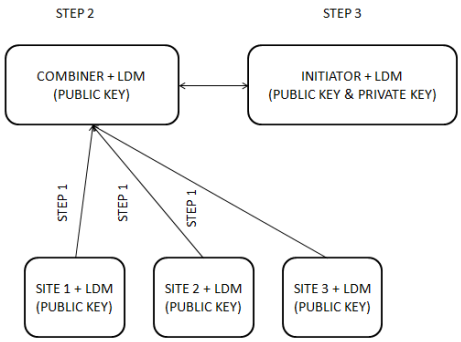


(g) Scenario of Algorithm

1. Predefined one party acts as Model Initiator. For e.g. Alice.
2. Alice generates RSA Public Key and Private Key. It sends the public key to each site.
3. Each site encrypts its data using Alice's public key [14].
4. Alice sends it to next site Bob. As Bob does not know decryption key, it won't be able to know the Alice's data. So it merges its own data, shuffles it and sends it next site Carol.
5. Carol again merges its own data as it cannot decrypt the data, shuffles it and sends it to the next site.
6. This process continues and each site merges its own data and sends it to the next site till the data reaches to the last site.
7. The last site merges its own data and sends it Initiator [14].
8. Initiator decrypts the complete data using the private key and removes the duplicate data. The initiator is unable to know the other site's data as data were shuffled by each site.
9. Finally Initiator (Alice) publishes the union of all the data sites [14].

B. MHS ALGORITHM FOR HORIZONTALLY PARTITION DATABASE

M. Hussein et al.'s Scheme (MHS) was introduced to improve privacy and try to reduce communication cost on increasing number of sites. The main idea was to use effective cryptosystem and rearrange the communication path. For this, two sites were discovered. This algorithm works with minimum 3 sites. One site acts as Data Mining Initiator and other site as a Data Mining Combiner. Rests of other sites were called client sites [2]. This scenario was able to decrease communication time. To improve the algorithm, Apriori-Tid data mining algorithm was used instead of standard Apriori algorithms. Fig [h] shows MHS algorithm.



(h) MHS Algorithm

The working of the algorithm is as follows:

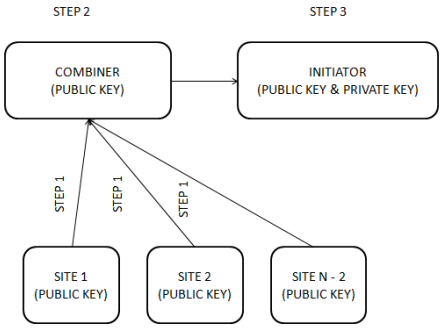
1. The initiator generates RSA public key and a private key. It sends the public key to combiner and all other client sites.
2. Each site, except initiator computes frequent itemset and local support for each frequent itemset using Local Data Mining (LDM) [2].
3. All Client sites encrypt their computed data using public key and send it to the combiner.
4. The combiner merges the received data with its own encrypted data, encrypts it again and sends it to initiator to find global association rules.
5. Initiator decrypts the received data using the private key. Then it merges its own LDM data and computes to find global results.
6. Finally, it finds global association rules and sends it to all other sites [2].

C. EMHS ALGORITHM FOR HORIZONTALLY PARTITION DATABASE

Enhanced M. Hussein et al.'s Scheme (EMHS) was introduced to improve privacy and reduce communication cost on increasing number of sites. This algorithm also works with minimum 3 sites. One site acts as Data Mining Initiator and other site as a Data Mining Combiner. Rests of other sites were called client sites [15]. But this algorithm works on the concept of MFI (Maximal Frequent Itemset) instead of Frequent Itemset. Also the algorithm uses two different cryptosystem as mentioned in II and III.

I. *MFI (Maximal Frequent Itemset):* A Frequent Itemset which is not a subset of any other frequent itemset is called MFI. By using MFI, communication cost is reduced [15].
II. *RSA (Rivest, Shamir, Adleman) Algorithm:* one of the widely used public key cryptosystem. It is based on keeping factoring product of two large prime numbers secret. Breaking RSA encryption is tough [15].
III. *Homomorphic Paillier Cryptosystem:* Paillier cryptosystem is an additive homomorphic cryptosystem, meaning that one can compute cipher texts into a new cipher text that is encryption of sum of the messages of the original cipher texts. For E.g. Let m1, m2 be the two messages. Then Encryption= E (m1+m2) =E (m1) *E (m2) and Decryption= D (E (m1) *E (m2)) =m1+m2 i.e. the sum of m1 and m2. Also, if the size of the public key is to (bit) then the size of cipher text c is 2*t (byte) [15].

Fig [I] shows the EMHS algorithm. The working of the algorithm was divided in two phases as follows:



(i) EHMS Algorithm

Phase-I:

a) The initiator generates RSA & Paillier public key and private key. It sends public keys to combiner and all other client sites [15].
b) Each site, except initiator computes its MFI, encrypts it using RSA public key and sends it to the combiner.
c) The combiner merges the received data with its own data and sends it to the initiator.
d) Initiator decrypts the received data using the private key. Then it adds its own data with the decrypted data and computes to find global MFI. Then the result is sent to all other sites [15].

Phase-II:

a) Each site finds frequent itemset and its local support count on the basis of MFI [15].
b) Each site, except initiator encrypts the data using Paillier's public key and sends it to the combiner.
c) The combiner merges its own data with the received data and sends it to the initiator.
d) Initiator decrypts the received data using the private key. Then it adds its own data with the decrypted data and computes to find global Association Rules. Then the result is sent to all other sites [15].

IX. CONCLUSION

In this paper, we discussed about data mining. We also discussed about various types of database. We presented a broad survey on privacy preserving in association rule mining on horizontally partitioned database. We discussed a variety of techniques and algorithms used for maintaining privacy or securing private and sensitive information. Still, there can be improvements in the defined algorithms. Better and improved algorithms must be defined which provides more security.

REFERENCES

[1] M. Atallah, A. Elmagarmid, M. Ibrahim, E. Bertino, and V. Verykios, Disclosure limitation of sensitive rules, in Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange, ser. KDEX 99. Washington, DC, USA: IEEE Computer Society, pp. 45-52 1999
[2] Mahmoud Hussein, Ashraf El-Sisi, Nabil Ismail Fast Cryptographic Privacy Preserving Association Rules Mining on Distributed Homogenous Data Base, Knowledge-Based Intelligent Information and Engineering Systems, Lecture Notes in Computer Science, Volume 5178/2008, pp. 607-616 2008.
[3] Agrawal D. Aggarwal C. C. On the Design and Quantification of Privacy-Preserving Data Mining Algorithms.ACM PODS Conference, 2002.
[4] Machanavajjhala A., Gehrke J., Kifer D., and Venkita subramaniam M.: l-Diversity: Privacy Beyond k-Anonymity. ICDE, 2006.
[5] Pinkas B.: Cryptographic Techniques for Privacy-Preserving Data Mining.ACM SIGKDD Explorations, 4(2), 2002.
[6] Blum A., Dwork C., McSherry F., Nissim K.: Practical Privacy: The SuLQ Framework.ACM PODS Conference, 2005.
[7] LiWu Chang and Ira S. Moskowitz, Parsimonious downgrading and decision trees applied to the inference problem, In Proceedings of the 1998 New Security Paradigms Workshop, 82-89. 1998
[8] D.W.Cheung,etal.,Ecient Mining of Association Rules in Distributed Databases, "IEEE Trans. Knowledge and Data Eng., vol. 8, no. 6, 1996,pp.911-922;
[9] YangZ., ZhongS.,Wright R.: Privacy-Preserving Classification of Customer Data without Loss of Accuracy. SDM Conference, 2006.
[10] Yi, X., Zhang, Y. Privacy-preserving distributed association rule mining via semi trusted mixer. Data Knowledge. Eng. 63(2), 550-567. 2007
[11] N V Muthu Lakshmi and Dr. K Sandhya Rani, PRIVACY PRESERVING ASSOCIATION RULE MINING WITHOUT TRUSTED PARTY FOR HORIZONTALLY PARTITIONED DATABASES, International Journal of Data Mining AND Knowledge Management Process (IJDKP) Vol.2, No.2 March 2012
[12] GUI Qiong, CHENG Xiao-hui, A Privacy-Preserving Distributed Method for Mining Association Rules", 2009 International Conference on Artificial Intelligence and Computational Intelligence, pp 294-297 2009.
[13] J. Han and M. Kamber, Data Mining: Concepts and Techniques. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 227-245.
[14] Estivill-Castro, V., Hajyasien, A Fast Private Association Rule Mining by a Protocol Se-curely Sharing Distributed Data, 2007 IEEE Intelligence and Security Informatics (ISI 2007), New Brunswick, New Jersey, USA, May 23-24, pp. 324-330 2007
[15] Xuan C. N., Hoai B. L., Tung A. C., An enhanced scheme for privacy preserving association rules mining on horizontally distributed databases, 2012 IEEE RIVF International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), pp 1 - 4 2012.