International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-07, Issue-02, MAY 2021

A Review on Privacy Preserving in Association Rule Mining on Partitioned Database

Rachit V. Adhvaryu, Tejas C. Chauhan

Assistant Professor, Marwadi University, Rajkot, India.

rachit.adhvaryu@marwadieducation.edu.in, tejas.chauhan@marwadieducation.edu.in

Abstract - The progression in data mining plays an vital role in various data applications. In accordance to privacy and security problems, the issues caused by association rule mining technique are examined thoriughly by various researchers. The proof clearly states that the mismanagement of this technique may result into the database owner's private and sensitive information to others. A lot of efforts to have been made in this field. In this paper, we have reviewed about the various techniques and algorithms used for privacy preserving with association rule mining using partitioned database.

Keywords — Association Rule Mining, Data Mining, Database, Partitioned Database, Privacy Preserving Association Rule Mining, Security

I. INTRODUCTION

Knowledge discovery also refered to as Data Mining techniques such as association rule mining, classification, clustering, sequence mining, etc. are the most widely used now a days in the field of research [1]. Prominent applications of these techniques has been demonstrated in many fields like marketing, medical analysis, business, bioinformatics, product control and some other areas that benefit commercial, social and humanitarian activities. Privacy Preserving in Data Mining is a vital factor which every mining system must support. This factor in real sense trie to secure sensitive and private information which the data owners do not want to disclose. 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 randomized value is masked with the original values of the data. The unwanted noise is added in large amount so that the original data value is not recovered in case of theft [3].
   ii. The K-Anonymity Model and L-Diversity: In K-Anonymity, the techniques like generalization and suppression were introduced to regulate data representation. In order to minimize the identification risk, every row in the database must be different. The L-Diversity method was introduced to overcome 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, users are not willing to reveal their private informations to other users. But each user in the loop is interested in achieving the aggregate information from the data set which is divided among all of them. [5]

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

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

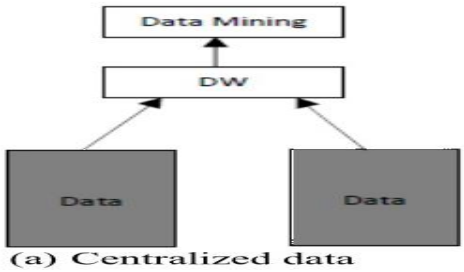
These techniques have been implemented either on Centralized Database or Distributed Database. The detailed information has been described in chapter II.

II. CENTRALIZED AND DISTRIBUTED DATABASE

A. Centralized Database

In the centralized database, all the data are stored in collective datasets and then are merged at one central site which can be called as Data Warehouse. All the mining process are done thereafter to fetch the required data. Fig [a] shows the centralized database.

 DOI : 10.35291/2454-9150.2021.0269



(a) Centralized data

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

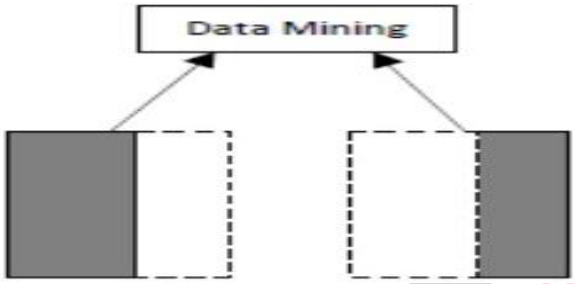
B. Distributed Database

Sometimes, users are not willing to reveal their private informations to other users. But each user in the loop is interested in achieving the aggregate information from the data set which is divided among all of them. [5].

Distributed Database is used now a days. Due to continuous growth in the database, the efficieint data management is not possible in central data. But, each dataset is stored at different places i.e. no site contains common data.

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

i. Vertically Partitioned Database: In this every site has different database schemas. The attribute values may or may not change. Fig [b] indicates Vertical Partitioned Database



(b) Vertical Partioining

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. Ultimately, the final outcome or the required information about the patient can be identified using joining operation on both the datasets. [9].

ii. Horizontally Partitioned Database: In this every site has the same database schema. But the attribute values are not similar [10]. Fig [c] indicates Horizontal Partitioned Database.

(c) Horizontal Partioining

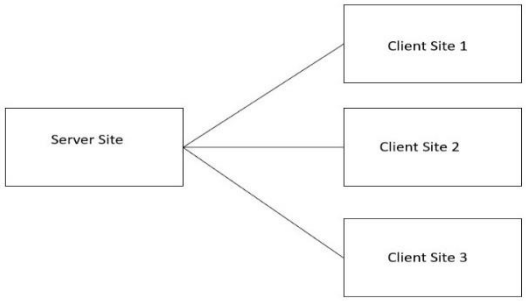
It means the information will be different at each site, but the schema to store the information at each site is same. For example, Credit card systems of two different banks have the same schema of storing the information of the users holding the credit card. But, the information of the users like user name, contact details, credit limit, etc. may be different with both the banks. [10]

In the later part of paper, we have mainly focused on the various techniques and algorithms used for Privacy Preserving on Horizontally Partitioned Database.

III. PRIVACY PRESERVING APPROACHES FOR HORIZONTALLY PARTITIONED DATABASE

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

It is a technique which uses Client-Server Architecture where one site acts as a server and other sites act as clients. All the client sites believe the server site to be trusted and honest one such that the server site won't disclose their sensitive and private data to another site [11]. Fig [d] shows SMC with trusted third party.



(d) SMC with trusted third party

In this, each site finds the frequent itemsets and its local support count and sends it the third party. On receiving these information, the third party finds global frequent itemset and global support count. The outcome of this process is returned back to all the client sites for further calculations and implementations [11].

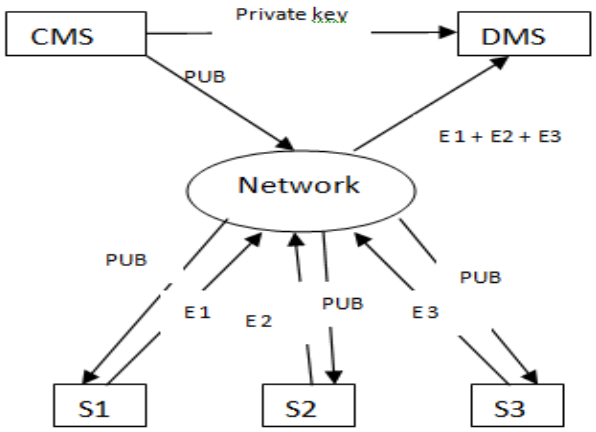
The limitation of this method was what if the third party fails or any mutual agreement occurs between thids party and some other site, then [11], huge data loss may incur and sesinsitve information of the user would be available with unwanted site. There were more chances of data loss in this technique. Algorithm describing this technique is as below

i. PDM-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 CMS generates and provides public key and private key for encryption and

 DOI : 10.35291/2454-9150.2021.0269

the DMS server decrypts the information and generates the Final Result. Fig [e] shows PPDM-ARBSM.



(e) PPDM-ARBSM ALGORITHM

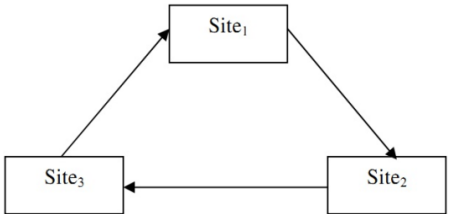
The working of the algorithm was as follows:

1. Identity of CMS is authenticated.
2. 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 limitation of this algorithm was the use of communication network. If the network is not string or if network fails, there would be loss of huge information as well as no global information will be generated.

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

This technique is quite similar to SMC with Trusted Third Party Model. A partially-honest site is one who follows the standard rules. But feels free to move around all the sites as well as in between the sites transferring data to gain more information and satisfy an independent agenda of interests [13]. In other words, a partial-honest site follows the rules step by step and exactly evaluates the required outputs based on the input from the other sites and it can analysis other site's data. However, it is sure that it will not inject any false value which results into the failure. The mututal agreement to share the information among the sites does not occur in this model and thus the privacy of information is not violated [13]. Fig [f] shows SMC with the semi honest model.



(f) SMC with semi honest model

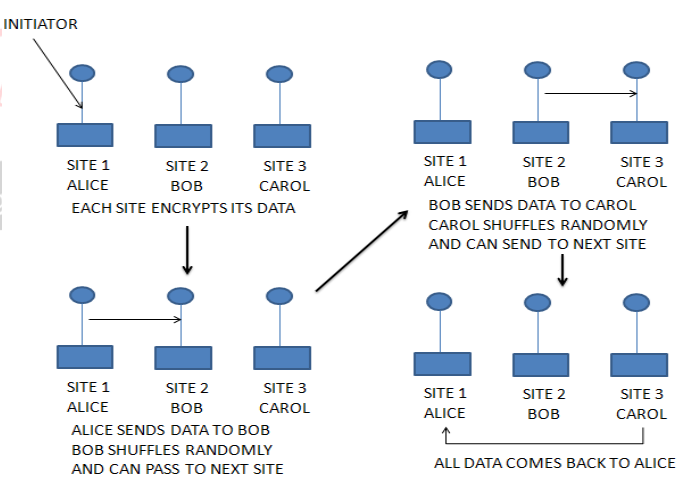
i. Each site assumes the other site to be honest.
ii. Each site follows the protocol
iii. Each site computes only the required data.
iv. Sites do not mutually agree with each other. But try to find some information about other site [13].

The limitation of the model was that each site is assumed to be honest. However, there is no assurance on mutual agreement of the sites .

Few algorithms describing this technique are explained below:

i. FAST PRIVATE ASSOCIATION RULES MINING FOR SECURELY SHARING

In this model, it has been assumed that each site follows semi-honest protocol. The sites can share the merged data without the using the data available with any trusted third party. The information like the records and the position of the records in the data set is made hidden. The whole scenario is governed by one initiator site which is either known as 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 site acts as Model Initiator. For e.g. Alice.
2. Alice generates RSA Public Key and Private Key. It sends the public key to every other 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

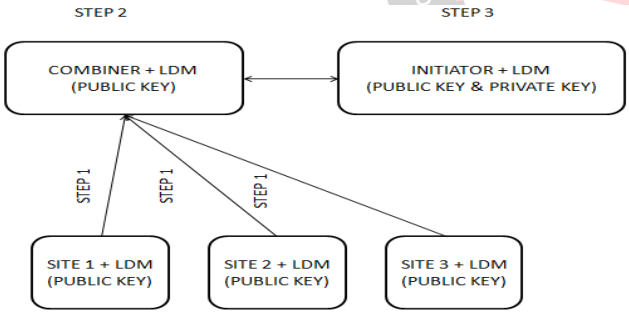
 DOI : 10.35291/2454-9150.2021.0269

the Alice's data. Here, Bob merges its own data, randomize the data and forwards it to next site Carol.

5. Carol again merges its own data as it cannot decrypt the data, shuffles it and forwards it to the next site.
6. This process continues and each site merges its own data and forwards it to the next site till the data reaches to the last site.
7. The last site aggregates its own data and lastly forwards it to Initiator [14].
8. Initiator decrypts the complete data using the private key and filters 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].

ii. MHS ALGORITHM FOR HORIZONTALLY PARTITION DATABASE

M. Hussein et al.'s Scheme (MHS) was introduced to improve security of data and minimize the communication cost on increasing number of communicating sites. The main idea was to use effective cryptographic system and rearrange the communication path. To implement this, two sites were used. One site acts as Data Mining Initiator and other site as a Data Mining Combiner. This algorithm works with minimum 3 sites. Rest of the other sites are known as client sites [2]. the communication time was decreased on implementing this structure. To improve the efficiency of 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 forwards 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 pass it to the combiner.
4. The combiner merges the received data with its own encrypted data, encrypts it again and forwards 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 forwards it to all other sites [2].

iii. EMHS ALGORITHM FOR HORIZONTALLY PARTITION DATABASE

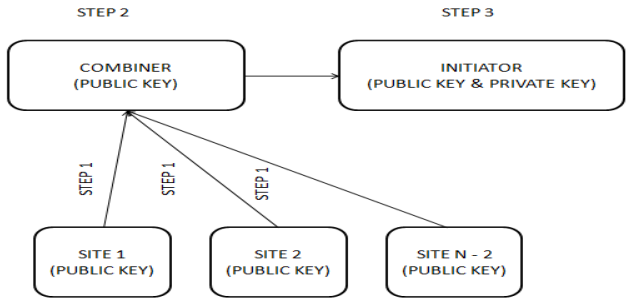
Enhanced M. Hussein et al.'s Scheme (EMHS) was introduced to improve efficiency in terms of privacy and reduce communication cost on increasing number of sites. This algorithm also works with minimum 3 sites. One site acts as Data Mining Initiator and other site as a Data Mining Combiner. Rest of the other sites were called client sites [15]. However, this algorithm works on the concept of MFI (Maximal Frequent Itemset) rather than using Frequent Itemset. Also the algorithm uses two different cryptographic system as discussed in b) and c.)

a) 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].

b) 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].

c) 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) EMHS Algorithm

Phase-I:

a) The initiator evaluates and generates RSA & Paillier public key and private key. It passes public keys to combiner and all other client sites [15].

 DOI : 10.35291/2454-9150.2021.0269

b) Each site, except initiator computes its MFI, encrypts it using RSA public key and then passes it to the combiner.
c) The combiner merges the received data with its own data and forwards 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. This result is then sent to all other sites for further computation [15].

Phase-II:

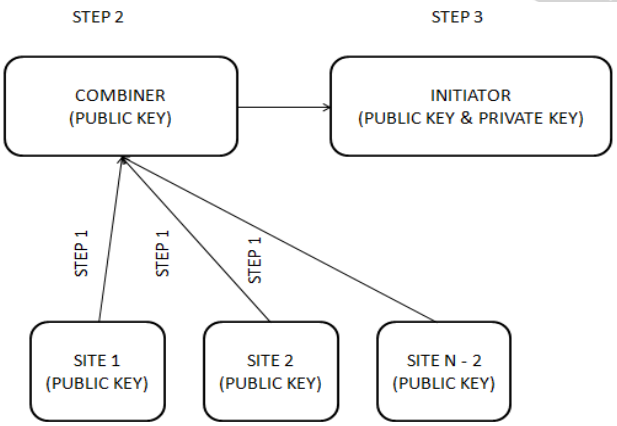
a) Each site finds frequent itemset and its local support count on the basis of MFI [15].
b) Each site, except initiator encrypts the data using Paillier's public key and passes it to the combiner.
c) The combiner merges its own data with the received data and forwards it to the initiator.
d) Initiator 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. Lastly, these results are shared with each site in the communication network [15].

iv. An Improved EMHS ALGORITHM FOR HORIZONTALLY PARTITION DATABASE

Here, improved version of EMHS algorithm has more efficiency in terms of Security of data and the overall communication cost. New cryptographic system in accordance with existing cryptographic system was introduced. The cryptographic system used was Elliptic Curve Cryptosystem.

The main focus was on to reduce communication cost and this ElGamal Algorothm was implemented to improve algorithm efficiency.

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



(j) Improved EMHS Algorithm

Phase-I:

a) The initiator shares ElGamal public key and Paillier public key with all the sites. It also generates Elgamal private key and Paillier private key. The keys are generated using ElGamal and Paillier Cryptography.
b) Each site, except initiator computes its MFI, encrypts it using ElGamal public key and then passes it to the combiner.
c) The combiner merges the received data with its own data and forwards union of all the data to the initiator.
d) Initiator decrypts the received data using the ELGamal private key. Then it adds its own data with the decrypted data and computes to find global MFI. This result is then sent to all other sites for further computation [16].

Phase-II:

a) Based on global MFI, each site finds frequent itemset and its local support count [16].
b) Each site, except initiator encrypts the data using Paillier's public key and passes it to the combiner.
c) The combiner computes all the collected data generates an intermediate output, merges its own data and forwards it to the initiator.
d) The initiator decrypts the received data using Paillier Private Key, evaluates received data and generates global support count for all sites.

Phase-II:

a) Each site follows phase II and sends all its data to combiner and then to the initiator [16].
b) Finally, initiator generates global association rules based on all data received from all the sites. These results are passed over to each site for furthere computation.

IV. CONCLUSION

In this paper, we discussed about data mining techniques. Moreover, We described various types of database. We presented a detailed review on privacy preserving in association rule mining on partitioned database and mainly focused on Horizontally Partitioned Database. We discoursed a variety of techniques and algorithms used for preserving the secrecy or make safe private and sensitive information. Still, there can be further improvements in the described algorithms. Better and improved algorithms can be implemented to get more efficient outcomes and better results without compromising 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*

 DOI : 10.35291/2454-9150.2021.0269

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.

[16] Adhvaryu R. V., Domadiya N. J., *An Improved EMHS Algorithm for Privacy Preserving in Association Rule Mining on Horizontally Partitioned Database*, 2014, Conference: Security in Computing and Communications Communications in Computer and Information Science Volume 467, 2014, pp 272-28At: Greater Noida, IndiaVolume

 DOI : 10.35291/2454-9150.2021.0269