Privacy Preserving In Horizontally Partition Data Based on Association Rules Mining: Review Paper

[1]Dharmik Makwana, [2] Prof. Krunal Panchal
[1]Department of Information Technology
[1]L.J.I.E.T, Ahmedabad, India.

***Abstract* - In Data mining is utilized to concentrate intrigued example or learning from extensive measure of data utilizing numerous data mining system. On the other hand it might likewise show touchy data about people trading off the individual right to security When a gathering of data is split among different gatherings. Presently Every last gathering would needs to keep its touchy data private amid the mining procedure. Security safeguarding data mining is to create data mining system without expansions the danger of abuse of data. The primary point of security protecting data mining is to locate the worldwide preserving so as to mine results the individual destinations private information/data.The different routines, for example, randomization, irritation, heuristic and cryptography strategies. To discover security preserving affiliation principle mining in evenly and vertically divided databases. In this paper, the examination of distinctive routines for PPARM is performed and their outcomes are looked at. On a level plane Parceled databases, calculation that joins point of interest of both RSA open key cryptosystem and Homomorphic encryption plan and calculation that uses Paillier cryptosystem to figure worldwide backings are utilized. This paper surveys the wide routines utilized for mining affiliation rules over on horizontally distributed dataset while preserving privacy.**

***Index Terms* - Privacy Preserving, horizontal database, association rule mining, cryptography method,EMHS.**

I. INTRODUCTION

The Data mining procedure utilize a delicate or individual information. It might be of common advantage for two gatherings or different gatherings to share their information for an examination undertaking. Notwithstanding, they might want to guarantee their own particular information stays private. Implies, there is a need to ensure touchy learning during data mining procedure. This issue is called Privacy Preserving Data Mining (PPDM). So keeping up protection is testing issue in data mining.

Many algorithm are proposed for data mining such as decision tree classification, clustering, association rule mining, Neural Networks, Bayesian Networks. while the algorithm are increase helpful information from the entire dataset. Many Privacy Preserving technique are found in data mining such as a randomization, anonymization and encryption method for distributed database. In many cases data is distributed, and bringing the data together in one place for analysis is not possible due these privacy laws or policies. In distributed environment, the database is available across multiple sites and privacy preserved data mining is performed to find the global mining results by preserving the individual sites private data or information. Every site can Compute a one function without knowledge of other parties input and access the global results which are useful for analysis. Distributed data into a two form horizontally partition data and vertically partition data. Horizontally data is each site has a complete information on a distinct set of entity. And vertically partition data is each site has different number of attribute with same number of transaction.

Privacy preserving association rule mining using horizontally partition database using a cryptography technique. In this method use a special encryption protocol is known as a Secure multiparty computation. SCM Provide a sub-protocol such as secure sum secures union, secure compression, secure scalar product..

In this paper, different methodologies are talked about for mining the affiliation rules while satisfying security prerequisites over horizontally partitioned distributed databases and correlation of all strategy.

II. ASSOCIATION RULE MINING

We concentrate on privacy preserving association rules mining on horizontally distributed databases[1]. In this type of database, Each site collect the same attribute of different entities. Each site shares its local itemset to each other to find strong association rules without revealing the sensitive data. As we have known, the strong global association rules are the global rules X Y (where X $\cap$ Y $\neq$ Ø) satisfying both global minimum support (sup%) and global minimum confident (conf%)

$$(X \rightarrow Y).\ \mathrm{sup} = \frac{X.\mathrm{sup}}{|DB|} = \frac{\sum_{i=1}^{n} X.\mathrm{sup}_i}{\sum_{i=1}^{n} |DB_i|} \geq \mathrm{sup}\ \%$$

$$(X \rightarrow Y).conf = \frac{\{X \cup Y\}.\sup}{X.\sup} = \frac{\sum_{i=1}^{n} XY.\sup_i}{\sum_{i=1}^{n} X.\sup_i} \geq conf\,\%$$

Equations (1) and (2) show that site Si does not need to share its local values of X, Y or {X U Y}. This means local data are already protected. However, site si need to share its X.supi,{X U Y}.supi, and |DBi|. In some specific applications, this information may be sensitive.

III. HORIZONTALLY PARTITIONED DATABASE

1. M. Hussein's Scheme

M. Hussein et al. [4] propose a modification to privacy preserving association rule mining on distributed homogenous database algorithm. algorithm is faster than old one which modified with preserving privacy and accurate results. Modified algorithm is based on a semi-honest model with negligible collision probability. The flexibility to extend to any number of sites without any change in implementation can be achieved.

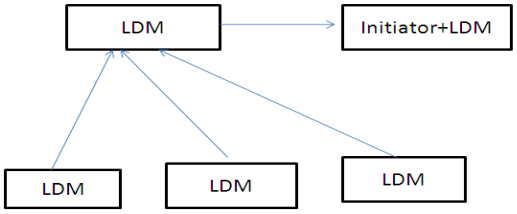

Fig.1 General Structure of Scheme

Step1: All local data mining (LDM) compute the mining results using fast distributed mining of association rules (FDM) as locally large k-item sets (LLi (k)) and local support for each item set in LLi (k) then Encrypt frequent item sets and support (LLei (k)) then send it to the data mining combiner.
Step 2: The combiner merge all received frequent items and supports with the data mining combiner frequent items and support in encrypted form then send LLe (k) to algorithminitiator to compute the global association rules.
Step3: The algorithm initiator receives the frequent items with support encrypted. The initiator first decrypts it, and then merges it with his local data mining result to obtain global mining results L(k), then compute global association rules and distribute it to all protocol parties. This algorithm is more flexible to extend it to any number of sites without any change in implementation.

2. Enhance M.Hussein'S Scheme

This algorithm is more flexible to extend it to any number of site implementation [4]. This method for privacy preserving association rules mining on horizontally distributed databases. It improves privacy and performance when number of sites gets increased. This algorithm uses two servers one is Initiator and other is Combiner and homomorphic Paillier cryptosystem to compute global supports.
There are three phases of this scheme.

Init Phase:
Initiator sends RSA's and Paillier's public key to all other sites.

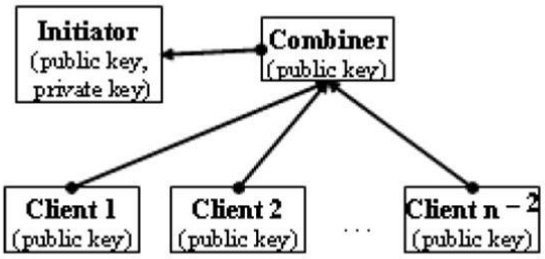

Fig.2 EMHS Model

First phase: Candidate set generation phase
Step 1: Each site independently and parallel finds its local Frequent itemset , and encrypts its local itemset by using its RSA's public key. Then send to their encrypted data to Combiner.

Step 2: Combiner merges the data received from Clients with its encrypted data and then sends the union data to Initiator.
Step 3: Initiator decrypts the data received from Combiner and combines the decrypted data to find global frequent itemset
Then Initiator sends the global MFI to all other sites. Each site generates candidate set, where each candidate is subset generated from each maximal frequent itemsets in global MFI. The candidates are different with each other and are sorted in the same order at all sites.

Second Phase: Global support computation phase
Step 1: Each site computes its local support count of each candidate and encrypts its support counts by using its Paillier's public key. Then send their encrypted data to Combiner. The encrypted of local support count of candidate X at site si is denoted as E (X. supi).
Step 2: With each candidate X, Combiner computes: E (X.supCombiner) = E (X.supCombiner) * E (X.supk) After that, decrypted data are sent to Initiator.
Step 3: Initiator decrypts the data received from Combiner and computes global support count of each candidate X as follows: X.sup = D (E (X.supCombiner)) + X.sup Initiator

Final Phase:
Each Site together computes Then Initiator finds strong global association rules and sends the result to all other sites. In EMHS, applying MFI approach in the first phase will reduce the size union data in this phase; and in the second phase, the union data is fixed when increasing the number of sites.

3. Improved EMHS

An algorithm to improve privacy and performance of EMHS when increasing the number of sites. They maintain the model of EMHS and apply ElGamal Cryptography in the first phase and Paillier cryptosystem in the second phase[7].

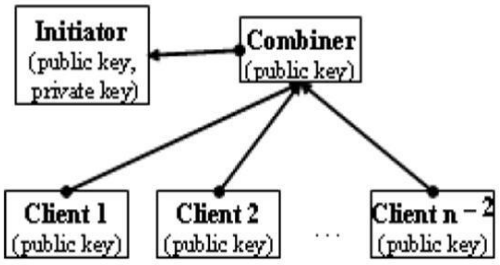

Fig.3 EMHS Model

Phase 1:
1. The initiator shares ElGamal public key *Epu* and Paillier public key *Ppu* with all the sites. It also generates Elgamal private key *Epk* and Paillier private key *Ppk*. The keys are generated using ElGamal and Paillier Cryptography.
2. Each site computes its local MFI. Then, all the sites except Initiator and Combiner encrypts it local MFI using ElGaml public key(*Epu*) and sends it to the Combiner.
3. The Combiner merges the received data with its own data and sends the union of all data to the Initiator.
4. Initiator decrypts the received data using ElGamal private key(*Epk*). Then it adds its own data and computes the Global MFI. Then, final Global MFI is shared to all other sites.

Phase 2:
1. Based on globalMFI, Each site finds Frequent Itemsets and its Local Support Count, encrypts the data using Paillier public key(*Ppu*) and sends it to the Combiner. The encryption of the local support count of candidate X at site *Si* is denoted by E(*X.supi*)
2. With each X, combiner computes:
E(*X.supCombiner*) = E(*X.supCombiner*) $*\_n-2$
k=1 $E(X.supk)$
After this, encrypted data is sent to Initiator.

3. Initiator decrypts the received data using Paillier private key(*Ppk*). It generates a global support count of each candidate X as:
X.sup = D(E(*X.supCombiner*)) + *X.supInitiator*

Phase 3:
1. Each site together computes | *DB* |=*n*
i=1
| *DBi* | in the same way used in phase 2.
2. Finally, Initiator generates the global association rules and sends the result to all other sites.

IV. COMPARETIVE STUDY

The methods proposed in three papers based on PPARM in horizontal partitioning of databases. EMHS follows MFI approach and does not modify the original data in both two phases. Thus, Initiator will find global frequent item sets accurately means the final results are accurate. Commutative encryption is used for algorithm proposed in [7] which didn't violate privacy constraints.
Both MHS [3] and EMHS [4] scheme satisfies semi-honest model. EMHS uses Paillier cryptosystem in the second phase and MHS uses RSA cryptosystem. For this reason, Combiner is much more difficult to attack in EMHS. Means EMHS has higher privacy than MHS. Both EMHS and MHS are two phase schemes, the communication cost (or cost) of each scheme is the sum of the one in each phase. EMHS has better performance than MHS in sparse datasets when increasing the number of sites.

V. CONCLUSION

In Privacy Preserving Association rule mining over horizontally partition database to find a global association rule from the local frequent itemset. In horizontally partition database utilize an alternate method to give a privacy. Like a M. hussein's schema, Enhance M. hussein's Schema, Secure CK sum. And also use a cryptographic method for encrypt the message is homomorphic encryption schema and RSA public cryptography provide the high security. Also give the point of preference and burden of this technique. Improve Computation and communication cost in multi-party horizontally partition data over malicious model

REFERENCES

[1] R. Agrawal, and R. Srikant. Fast algorithms for mining association rules. In: Proceeding of the 20th International Conference on Very Large Data Bases, 1994:487-499.
[2] D. Cheung, J. Han, V. Ng, et al. A fast distributed algorithm for mining association rules. In: Proceedings of 1996 International Conference of Parallel and Distributed Information Systems 1996:31-42.
[3] Mahmoud Hussein, Ashraf El-Sisi, Nabil Ismail: Fast CryptographicPrivacy Preserving Association Rules Mining on DistributedHomogenous Data Base. Knowledge-Based Intelligent Information andEngineering Systems, Lecture Notes in Computer Science, Volume5178/2008, pp. 607 -- 616 (2008)
[4] Xuan Canh Nguyen, Hoai Bac Le, Tung Anh Cao, "An Enhanced Scheme For Privacy-Preserving Association Rules Mining On Horizontally Distributed Databases," In 2012 IEEE
[5] R. Agrawal, A. Evfimievski, and R. Srikant. Information sharing across private databases. In: Proceedings of The 2003 ACM SIGMOD International Conference on Management of Data. 2003:86-97.
[6] M. Kantarcioglu, and C. Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned Data. IEEE Transaction on Knowledge and Data Engineering. 2004, 16(9):1026-1037.
[7] Rachit Adhvaryu and Nikunj Domadiya," An Improved EMHS Algorithm for Privacy Preserving in Association Rule Mining on Horizontally Partitioned Database", J. Lloret Mauri et al. (Eds.): SSCC 2014, CCIS 467, pp. 272–280, 2014._c Springer-Verlag Berlin Heidelberg 2014
[8] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy Preserving Mining of Association Rules," Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 217-228, 2002, available: http://doi.acm.org/10.1145/775047.775080.
[9] Murat Kantarcioglu and Chris Clifton: Privacy-preserving Distributed Mining of Association Rules on Horizontally Partitioned Data. IEEE transactions on knowledge and data engineering - Volume 16 Issue 9,September 2004 (2004)
[10] B. Pinkas, " Cryptographic techniques for Privacy-preserving data mining," ACM SIGKDD Explorations Newslettle,vol. 4,no.2,pp 12-19,2002.