Privacy Preserving Association Rule Mining using Horizontally Partition Data: Review Paper

Arpita B. Modh
Department of Computer Science and Engineering
L.J.Institute of Technology
Ahmedabad-382210, Gujarat

ABSTRACT
In Data mining is used to extract interested pattern or knowledge from large amount of data using many data mining technique. However it may also display sensitive information about individuals compromising the individual right to privacy When a collection of data is split among various parties. Now Each and Every party would wants to keep its sensitive information private during the mining process. Privacy preserving data mining is to develop data mining method without increases the risk of misuse of data. The main aim of privacy preserving data mining is to find the global mining results by preserving the individual sites private data/information. The various methods such as randomization, perturbation, heuristic and cryptography techniques. To Find privacy pre serving association rule mining in horizontally and vertically partitioned databases. In this paper, the analysis of different methods for PPARM is performed and their results are compared. Horizontally Partitioned databases, algorithm that combines advantage of both RSA public key cryptosystem and Homomorphic encryption scheme and algorithm that uses Paillier cryptosystem to compute global supports are used. This paper reviews the wide methods used for mining association rules over horizontally distributed dataset while preserving privacy.

Keywords
Privacy Preserving association rule mining, cryptography method , Secure multiparty computation.

1. INTRODUCTION
During the data mining process use a sensitive or personal data. It may be of mutual benefit for two parties or multiple parties to share their data for an analysis task. However, they would like to ensure their own data remains private. Means, there is a need to protect sensitive knowledge during a data mining process. This problem is called Privacy-Preserving Data Mining (PPDM). So maintaining privacy is challenging 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 gain useful knowledge from the whole dataset. Many Privacy Preserving technique are found in data mining such as a randomization, anonymization and encryption method for distributed database. One of the most studied problems in data mining is 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 approaches are discussed for mining the association rules while fulfilling privacy requirements over horizontally partitioned distributed databases and comparison of all method.

2. HORIZONTALLY PARTITIONED DATABASE

2.1 Association Rule mining
We focus 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 ∩ Y ≠ Ø) satisfying both global minimum support (sup%) and global minimum confident (conf%)

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

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

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

2.2 Distributed Association Rule
This paper investigates the mining of association rules in a distributed environment[2]. Let DB be a database with D transactions. Assume that there are n sites, S1, S2… Sn, in a distributed system, and the database DB is partitioned over the n sites into {DB1,DB2 …. DBn}, respectively.

Let the size of the partitions DBi be Di, for i = 1…. n.

Let X.sup be the global support count, and X.supi be the local support count of X at site Si. For a given minimum support threshold s, X is globally frequent if X.sup ≥ s × D; correspondingly, X is locally frequent at site Si, if X.supi ≥ s×Di. In the following, L denotes the globally frequent

itemsets in DB, and L(k) the globally frequent k-itemsets in L. The essential task of a distributed association rule mining algorithm is to find the globally frequent itemsets L.

A fast algorithm for distributed association rule mining is given in [1]. The procedure for the fast distributed mining of association rules (FDM) can be summarized below:

1.Candidate Sets Generation: Using the classic Apriori candidate generation algorithm. Each site generates candidates item set

2. Local Pruning: scan the database DBi at Si to compute X.supi. If X is not locally frequent at site Si, it is Remove from candidate sets LLi(k).

3. Support Count Exchange: Broadcast the candidate set in LLi(k) to other sites in order to collect support count. To Compute the global frequent itemset.

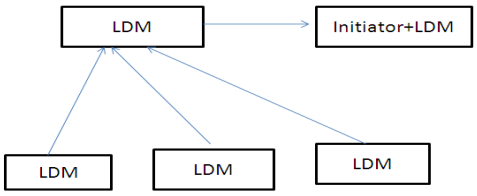


Fig- 1. General Structure of Scheme

Step1: All site (LDM) compute the mining results using fast distributed mining of association rules (FDM) found the locally large item sets (LLi (k)) after that 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 algorithm initiator to compute the global association rules.

Step3: The 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 efficient to extend it to any number of sites without any change in implementation.

2.3 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

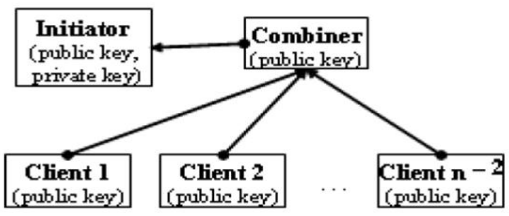


Fig-2 EMHS Model

There are three phases of this scheme.

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

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.

2.4 Commutative Encryption

Commutative Encryption System[5]. An encryption system F = (M,K, f, g) (where M is the domain of plain and encrypted messages, K = (E,D), E, D are sets of encryption keys and decryption keys, respectively) is commutative if both the encryption function f : E ×M → M and the decryption function g : D ×M → M are computable functions in polynomial time, defined on finite computable domains, that satisfy all properties listed below. We denote fe(x) ≡ f(e, x), gd(x) ≡ g(d, x), and use ∈r to signify is chosen uniformly at random from.

1.For all e1, e2…en∈ E, d1, d2, ・ ・dn ∈ D, each m ∈ M, we have

fei1 (fei2 (… (fein (m)) …)) = fej1 (fej2 (….. (fejn (m)) …)) ≡ T and gds1(gds2 (……. (gdsn (T)) ….)) = gdt1 (gdt2 (….(gdtn (T)) ….)),

where (i1, i2,……. in), (j1, j2…… jn), (s1, s2… sn) and (t1, t2,…….. tn) are four permutations of (1, 2.. n).

2. For all e1,e2,・ ・ ・ ,en ∈ E, all m1,m2 ∈ M such that m1 = m2 and big enough k, we have

pr(fe1 (fe2 (· · · (fen(m1)) · · ·)) = fe1 (fe2 …(fen(m2)) · · ·))) <1/2k

3. For x, y, z ∈ r M, e∈ r E, the distribution of < x, fe(x), y, fe(y) > is indistinguishable from the distribution of < x,fe(x), y, z >.

Informally, Property 1 denotes that the ciphertext of compositely commutative encryption is identical regardless of the encryption order.

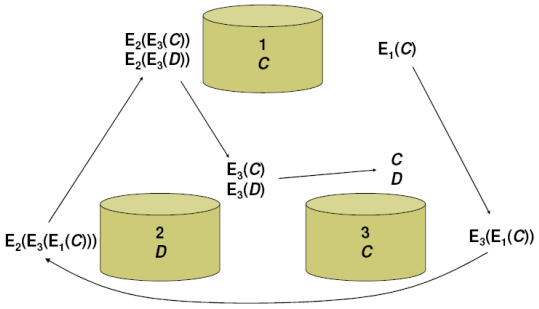
Property 2 signifies that two different plain messages will never have the same encrypted messages. Property

3 guarantees that the encryption is secure. Specifically, it states that given a plain message x and its randomly chosen encryption way fe(x), for a new plain message y, an adversary

2.5 Secure Set Union

Secure set union is a method of privacy preserving in data mining rules[6], local large item-sets etc. local large itemset can be broadcasted in the distribution system, but owners of broadcasted information can't be exposed. Secure set union executes its function basing on the thought of permuting encrypt item.

Fig- 3 shows the principle to set up secure set union.



The first phase uses commutative encryption. Each party encrypts its own frequent itemsets . The encrypted itemsets are then passed to other parties until all parties have encrypted all itemsets. These are passed to a common party to eliminate duplicates and to begin decryption. This set is then passed to each party, and each party decrypts each itemset. The final result is comprised of the distinct itemsets (C and D in the figure).

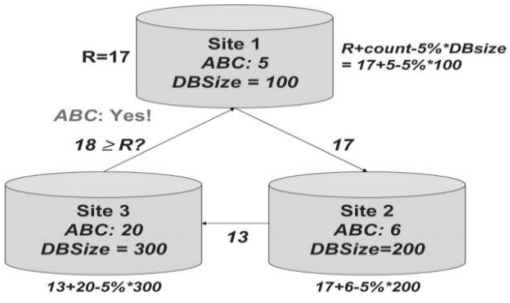


Fig-4 Locally itemset is tested

In the second phase each of the locally supported itemsets is tested to see if it is supported globally.

Determining if itemset support exceeds the 5 percent threshold [13] that the itemset ABC is known to be supported at one or more sites, and each computes its local support. The first site chooses a random value R and adds to R the amount by which its support for ABC exceeds the minimum support threshold. This value is passed to site 2, which adds the amount by which its support exceeds the threshold This is passed to Site 3, which again adds its excess support. The resulting value (18) is tested using a secure comparison to see if it exceeds the random value (17). If so, itemset ABC is supported globally.

3. 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 [8] which didn't violate privacy constraints.

Both MHS [4] and EMHS [9] 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.

4. 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 use a different technique to provide a privacy. Like a Secure sum, Secure Union, 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 provide the advantage and disadvantage of this method. Improve Computation and communication cost in multi-party horizontally partition data over malicious model

5. 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.