Secure Distributed Data Mining

Priyanka Khairnar
Computer Department,
MET BKC Adgaon, Nashik, Savitribai Phule Pune University, Maharashtra India.

ABSTRACT
Security is the important paradigmin data rule mining projects. This project addresses the problem of secure distributed association rule mining over the horizontally distributed database. Through mining, interesting relations and patterns between variables of large database can be observed securely using cryptographic techniques and the mining algorithms. Round robin technique is used for Horizontal distribution of Data sets to reduce the data skew. Security concerns may prevent the sites from direct sharing of data and some type of information about the data. The paper introduces cryptographic techniques to provide security in order to minimize the information shared in mining.

Keywords
Distributed Mining, RSA, Distributed Apriori algorithm, multiparty computation

1. INTRODUCTION
The problem of secure distributed association rule mining is studied here. In this problem there are several sites that hold homogeneous Database, this database is distributed horizontally over different sites participating in transaction.

Here goal is to mine data for finding all association rules with support count at least s and confidence count at least c, for given minimal support size s and confidence c, that hold in the unified database. The main and important part of project is minimizing the information disclosed about the private Database held by sites in transaction. The information that we going to protect here is individual transactions in the different Database at each site, and also global information like association rules supported locally by each of those Database at different sites [1].

Purpose- Here the design of an alternative protocol has been proposed to the securely compute the union of private subsets. The system offers simplicity and efficiency as well as privacy. The system does not depend on commutative encryption that means all are encrypted in the same manner.[4][5]

2. EXISTING SYSTEM
In the existing system the protocol for securely computing the union of private subsets at each site in the transaction is studied. In the existing system a multi-partycomputation is considered, which is the most costly part of the system and in its implementation cryptographic techniques like encryption, decryption, commutative encryption, and hash functions are used. [1], [9].

The use of such cryptographic techniques improves communication cost and computation cost. In the existing system though these techniques are used it causes some leakage of information, Therefore it is not perfectly secure. Thus the union of private subsets is not perfectly calculated, so the system is proposed to overcome with this problem.

3. PROPOSED SYSTEM
In the proposed system the problem of secure computation of union of private subsets of sites is addressed. Here it has been proposed that the database is distributed horizontally among various sites in transaction. Round robin technique is used for Horizontal distribution of Data sets to reduce the data skew.

The input is synthetic database and the output produced will be list of association rules. The proposed system is implemented using DM algorithm and encryption based techniques. Fast Distributed Mining is the distributed version of apriori algorithm. The proposed system improves in terms of communication cost, computation cost, efficiency as well as security.

The goal takes us to the secure multiparty computation, which can be best understood with the example of two merchants, they wish to find which one of them is having more money. They want to find it without involving any third party, also without revealing their actual money to each other. In the same way in our system we are going to find the globally frequent itemset without revealing private itemsets of each site.

4. PROPOSED SYSTEM ARCHITECTURE
Each transaction D is divided into partitions and in each local partition frequent itemsets are find out. After finding local frequent itemsets all local frequents itemsets are combined to find candidate itemset. In last stage global frequent itemsets are found.

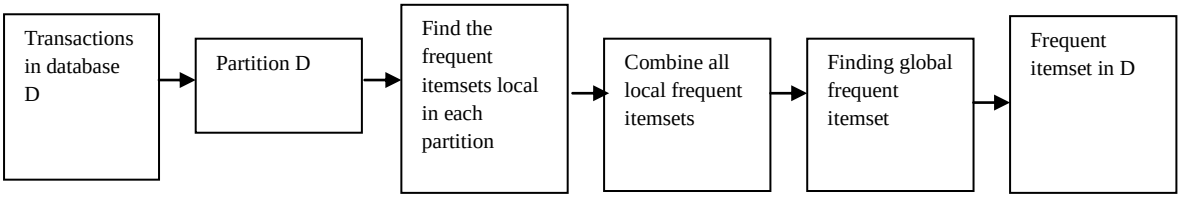

Figure 1: Block diagram

Figure 1 show how frequent itemsets can be generated whereas Figure 2 shows the architecture of proposed method. There are four sites site1,site2,site3, site4,that hold homogeneous Database, i.e., Database that are having same information but exist on different sites. The main aim of proposed system is to find association rules with support count of at least s and confidence count at least c, for some given minimal support count s and confidence level of c, and to minimize the information disclosed about the private Database held by those sites. The information that we are going to protect here is the individual transaction information of site as well as the global information about the at different sites, In figure 2 Lk is the set of frequent item set generated using flow show in figure 1.TD is the transactional database By communicating with each site a frequent tem set is generated by each Local site using distributed mining algorithm.

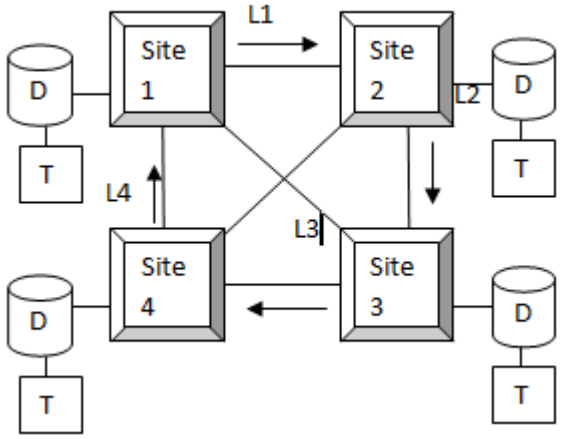

Figure 2 Architecture

5. ALGORITHMS

5.1 Distributed Mining Algorithm

The DM algorithm proceeds as follows:

1) Initialization
2) Site ItemSets Generation - Each site will generate its frequent itemset. Check weather frequent itemset

   is Locally frequent in own site and itemset is globally frequent.
3) Local Pruning-Retains Locally frequent item sets.
4) Identification of the candidate item sets – Each site broadcasts its itemset.
5) Computation of local supports - Compute local supports of all itemsets.
6) Broadcast Mining Results - Here it is identified that each locally frequent item is subset of globally frequent itemset. Algorithm Proceeds until it finds no (k+1) item are longest globally frequent itemsets. Here k is number of itemsets [5].

5.2 Rivest-Shamir-Adleman (RSA) Algorithm

The step number 5 of DM algorithm can be implemented using RSA algorithm. Rivest-Shamir-Adleman (RSA). Is a public key encryption algorithm invented byRivest.RSA is an algorithm for providing public key encryption [6].

The algorithm works as follows:

1) Select two sets P and Q.
2) Calculate N = P x Q.
3) Choose the public encryption key E such that it is not a factor of (P – 1) and (Q -1).
4) Select the private decryption key D such that,

   (D x E) mod (P – 1) x (Q – 1) = 1.
5) For encryption, calculate the cipher text from the plain text as follows:

   CT =E (PT)
6) Send Cipher text to the receiver site.
7) For decryption, calculate the plain text from the cipher text as follows:

   PT =D (CT)

Where,

CT=Cipher Text

PT=Plain Text

6. MODULES

6.1 User Module-

In this module, different sites participating in transaction are considered as users. The problem definition is of interest if number of sites participating in transaction is greater.

6.2 Admin Module-In this module, Site details can be verified. Also views the item set using association rule.

6.3 Association Rule- Association rules are rules that help to understand relationships, patterns between variables of database. Association rule mining is import paradigm of mining in distributed environment.

6.4 Apriori Algorithm- Apriori algorithm is used to operate on Database containing transactions on different sites. The main aim of the Apriori Algorithm is to find an association rule that is patterns or interesting relations between different datasets. It is rendered as "Market Basket Analysis". In this each data set is having number of transaction. The output of Apriori algorithm is sets of rules that show how frequent item sets of data are generated at each of the site.

7. METHODOLOGY

The proposed method can be implemented as follows: In implementation of system the database is distributed horizontally among various sites in the transaction. Round robin technique is used for Horizontal distribution of Data sets to reduce the data skew. The Join key is present at all the sites where the database is distributed. While implementation, one database at one site in the transaction is rendered as primary and it is considered as "Initiator" of the process or system. The Database on other sites will act as "Responder" of process.

The main Moto is to find association rules involving attributed except join key. Also security should be maintained while doing this mining process. Therefore for maintaining security the encryption algorithms like RSA, key hashed functions like HMAC can also be used.[6],[7],[8].

8. CONCLUSION

In this paper the interesting properties between locally frequent and globally frequent itemsets are observed. The distributed version of Apriori algorithm is applied for distributed mining of association rules. The Cryptographic tools can enable us for securely performing association rule mining. We have given techniques to mine distributed association rules on horizontally partitioned database. Round robin technique is used for Horizontal distribution of Data sets to reduce the data skew.

It is expected that distributed association rule mining can be done efficiently through security assumptions. It is possible to mine globally valid results from distributed data without revealing private information. Such Secure distributed association rule mining can be done with a reasonable cost. Research will expand the scope of Secure distributed association rule mining that will enable most or all association rule mining methods to be used.

9. REFERENCES

[1] T.Tassa Secure Mining of association rules in horizontally distributed Database.2014

[2] G. Alex and A. Freitas, "Scalable, high-performance data mining with parallel processing,"in Principles and Practice of Knowledge Discovery in Databases, (Nantes, France),1998.

[3] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large Database. In VLDB, pages 487499, 1994.

[4] A.V. Ev_mievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In KDD, pages 217228, 2002.

[5] D.W.L. Cheung, J. Han, V.T.Y. Ng, A.W.C. Fu, and Y. Fu. A fast distributed algorithm for mining association rules. In PDIS, pages 3142, 1996.

[6] R.L. Rivest, A. Shamir, and L.M. Adleman, "A Method for Obtaining Digital Signatures and Public-Key Cryptosystems," Comm.ACM, vol. 21, no. 2, pp. 120-126, 1978.

[7] A. Ben-David, N. Nisan, and B. Pinkas, "FairplayMP - A System for Secure Multi-Party Computation," Proc. 15th ACM Conf. Computer and Comm. Security (CCS), pp. 257-266, 2008.

[8] M. Bellare, R. Canetti, and H. Krawczyk, "Keying Hash Functions for Message Authentication," Proc. 16th Ann. Int'l Cryptology Conf. Advances in Cryptology (Crypto), pp. 1-15, 1996.

[9] M. Kantarcioglu and C. Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned data. IEEE Transactions on Knowledge and Data Engineering, 16:10261037, 2004.

[10] T. Tassa and E. Gudes. Secure distributed computation of anonymizedviews of shared databases. Transactions on Database Systems, 37, Article 11, 2012.

[11] T. Tassa, A. Jarrous, and J. Ben-Ya'akov. Oblivious evaluation ofmultivariate polynomials. Submitted.

[12] J. Brickell and V. Shmatikov. Privacy-preserving graph algorithms inthe semi-honest model. In ASIACRYPT, pages 236–252, 2005.