Privacy Preserving Association Rules Mining in Horizontally Distributed Databases Using FDM and K&C Algorithm

Gayatri K. Chaturvedi[1], Ranjit M.Gawande[2]
[1] Post Graduate Student, [2]Assistant professor
Department of Computer Engineering, MCOER, Pune University, Maharashtra, India.

***Abstract* - Data mining is the most fast growing area today which is used to extract important knowledge from large data collections but often these collections are divided among several parties. This paper addresses secure mining of association rules over horizontally partitioned data. This method incorporates a protocol is that of Kantarcioglu and Clifton well known as K&C protocol. This protocol is based on an unsecured distributed version of the Apriori algorithm named as Fast Distributed Mining (FDM) algorithm of Cheung et al. The main ingredients in our protocol are two novel secure multi-party algorithms one that computes the union of private subsets that each of the interacting players hold and another that tests the whether an element held by one player is included in a subset held by another. This protocol offers enhanced privacy with respect to the earlier protocols. In addition, it is not complicated and is importantly more effectual in terms of communication cost, communication rounds and computational cost. We present a two multiparty algorithm for efficiently discovering frequent item sets with minimum support levels without either player (site) revealing it to all players.**

***Keywords* - Security, Privacy, Data Mining, Frequent Item sets, Association Rules, multi-party**

I. INTRODUCTION

Data mining can extract important knowledge from large data collections but sometimes these collections are split among various parties. Privacy liability may pre-vent the parties from directly sharing the data, and some types of information about the data. Data mining technology has become prominent as a means of identifying patterns and trends from large quantities of data. Data mining and data warehousing co-jointly: most popular tools operate by gathering all data into a central site then running an algorithm against that data. However, privacy liability can prevent building a centralized warehouse data may be distributed among several custodians none of which are allowed to transfer their data to another site.

In Horizontally partitioned database there are several players that hold homogeneous database. The goal is to find all association rules with support at least s and confidence at least c, for some given minimal support size s and confidence level c, that hold in the unified database, while minimizing the information disclosed about the private databases held by those players. That goal defines a problem of secure multi-party computation.

If there existed a trusted third party, the players could surrender to him their inputs and he would perform the function evaluation and send to them the resulting out-put. In the absence of such a trusted third party, it is needed to devise a protocol that the players can run on their own in order to arrive at the required output y. Such a protocol is considered perfectly secure if no player can learn from his view of the protocol more than what he would have learnt in the idealized setting where the computation is carried out by a trusted third party.

In previous year various techniques are applied for secure mining of association rules in horizontally partitioned database. These approaches use various techniques such as data perturbation, homo-morphic encryption, keyword search and oblivious pseudorandom functions etc. These privacy preserving approaches are inefficient due to

- Homo-morphic encryption
- Higher computational cost
- In some of the techniques data owner tries to hide data from data miner.

Our proposed protocol based on two novel secure multiparty algorithm using these algorithms the protocol provides enhanced privacy, security and efficiency as it uses commutative encryption.

In this project we propose a protocol for secure mining of association rules in horizontally distributed database. This protocol is based on: FDM Algorithm which is an unsecured distributed version of the Apriori algorithm. In our protocol two secure multiparty algorithms are involved:

1. Computes the union of private subsets that each interacting players hold.
2. Tests the inclusion of an element held by one player in subset held by another.

In Horizontally partitioned database there are several players that hold homogeneous database. Our protocol offers enhanced privacy with respect to the current leading K and C protocol simplicity, more efficient in terms of communication rounds, communication cost and computational cost.

In our problem, the inputs are the partial databases and the required output is the list of association rules that hold in the unified database with support and confidence no smaller than the given thresholds s and c, respectively. The paper is organized as follows. Section 2 overviews the details about your proposed work. This section includes the details about Algorithms, flowchart etc. Sections 3 presents explain all the result of work in the form of graph, figure, chart etc. Section 4 explains analysis part of the application and feature scope of current work. Section 5 states the possible follow-ups of this work and draws the conclusions.

II. METHODOLOGY

2.1 Process Design

Let D be a transaction database. The database is partitioned horizontally between P_1, P_2 . . . , P_m players, denoted 1 M. Player Pm holds the partial database Dm that contains Nm = |Dm | of the transactions in D, 1≤m ≤M . The unified database is D =D_1 U···U D_M.

An itemset X is a subset of A. Its global support, supp(X), is the number of transactions in D that contain it. Its local support, sup (X), is the number of transactions in Dm that contain it.

Support - The rule X ⇒ Y holds with support s if s% of transactions in D contain X ∪ Y. Rules that have a s greater than a user-specified support is said to have minimum support or threshold support. The support of a rule is defined as,

sup(X) = no of transactions that contain X / total no of Transactions.

Confidence - The rule X ⇒ Y holds with confidence c if c% of the transactions in D that contain X also contain Y. Rules that have a c greater than a user-specified confidence is said to have minimum confidence or threshold Confidence. The confidence of a rule is defined as,

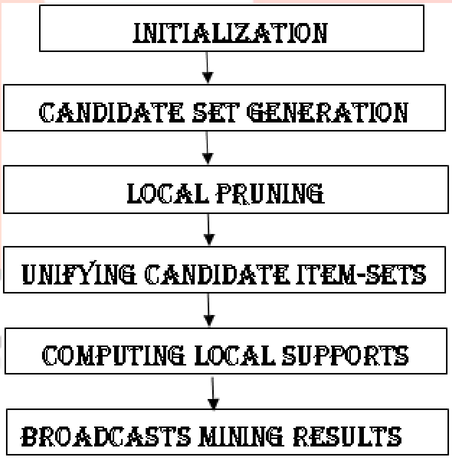
conf(X =>Y) = sup(X U Y)/ supp(X).

FDM Algorithm

Fast Distributed Mining (FDM) algorithm is an unsecured distributed version of the Apriori algorithm. Its main idea is that any s-frequent itemset must be also locally s-frequent in at least one of the sites. Hence, in order to find all globally -frequent itemsets, each player reveals his locally s-frequent itemsets and then the players check each of them to see if they are s-frequent also globally.

In the first iteration of FDM algorithm, when k=1, Cs1,m the set that the mth player computes (Steps 2-3) is just Fs1,m , namely, the set of single items that are s-frequent in Dm. The complete FDM algorithm starts by finding all single items that are globally s-frequent. It then proceeds to find all 2-itemsets that are globally s-frequent, and so forth, until it finds the longest globally s-frequent itemsets. If the length of such itemsets is k, then in the (k+1)th iteration of the FDM it will find no (k+1)-itemsets that are globally s-frequent, in that case it terminates.

FDM algorithm steps are as follows:



Unifi-KC(FDM-KC)

The input that each player P_m has at the beginning of Protocol UNIFI-KC is the collection $C_s^{k,m}$, as defined in Steps 2-3 of the FDM algorithm. Let Ap(F^{k-1}) denotes the set of all candidate k-itemsets that the Apriori algorithm generates from F_s^{k-1} s . The output of the protocol is the union $C_s^k = \bigcup_{m=1}^{M} C_s^{k,m}$.In the first iteration of this computation, and the players compute all -frequent 1-itemsets (here F_S^0 = s {∅}). In the next iteration they compute all s-frequent 2-itemsets, and so forth, until the first ≤ in which they find no s-frequent k-itemsets.

After computing that union, the players proceed to extract from C_s^k the subset F_s^k that consists of all k-itemsets that are globally s-frequent; Finally, by applying the above described procedure from k=1 until the first value of k≤L for which the resulting set F_s^k is empty, the

Players may recover the full set of $F_s = \bigcup_{k=1}^{L} F_s^k$ all globally –frequent item sets.

Protocol UNIFI-KC works as follows: First, each player adds to his private subset $C_s^{k,m}$ fake item sets, in order to hide its size. Then, the players jointly compute the encryption of their private subsets by applying on those subsets a commutative encryption, where each player adds, in his turn, his own layer of encryption using his private secret key. At the end of that stage, every item set in each subset is encrypted by all of the players; the usage of a commutative encryption scheme ensures that all item sets are, eventually, encrypted in the same manner. Then, they compute the union of those subsets in their encrypted form. Finally, they decrypt the union set and remove from it item sets which are identified as fake.

Steps For secure computations of all item sets (by K&C):

1. Cryptographic Primitive Selection
   - Player selects needed commutative cipher and corresponding private key
   - Player selects hash function to apply on all itemsets prior to encryption
   - Player compute lookup table with hash values to find preimage of given hash values.
2. All itemsets Encryption
3. Itemset Merging
   - Each odd player sends his encrypted set to player 1.
   - Each even player sends his encrypted set to player 2.
   - Player 1 unifies all sets that were sent by the odd players and removes duplicates.
   - Player 2 unifies all sets that were sent by the even players and removes duplicates.
   - Player 2 sends his permuted list of itemsets to Player 1.
   - Player 1 unifies his list of itemsets and the list received from Player 2 and then remove duplicates from the unified list. Denote the final list by EC_S^K.
4. Decryption

III. Result

Let *D* be a database of *N* = 18 item sets over a set of *L* = 5 items, *A* = *{*1, 2, 3, 4, 5*}*. It is partitioned between *M* = 3 players and the corresponding partial databases are:
*D*1 = *{*12, 12345, 124, 1245, 14, 145, 235, 24, 24*}*
*D*2 = *{*1234, 134, 23, 234, 2345*}*
*D*3 = *{*1234, 124, 134, 23*}*.
For example, *D*1 includes *N*1 = 9 transactions, the third of

Min-Support (s) or threshold support ($0 < s \leq 1$) = 0.33
Min-Confidence (c) or threshold Confidence ($0 < c \leq 1$) = 0.33.

Table 1 Result set using FDM and Unifi KC Algorithm:

SR.NO	Item-Set	Support Value	Confidence Value
1	1	0.61	-----
2	2	0.78	-----
3	3	0.56	-----
4	4	0.78	-----
5	1,2 = {1} -> {2}	0.39	0.64
6	1,4 = {1} -> {4}	0.56	0.92
7	2,3 = {2} -> {3}	0.44	0.56
8	2,4 = {2} -> {4}	0.56	0.72
9	3,4 = {3} -> {4}	0.39	0.7
10	1,2,4 = {1} -> {2,4}	0.33	0.54
11	1,2,4 = {1,2} -> {4}	0.33	0.85

IV. Discussion

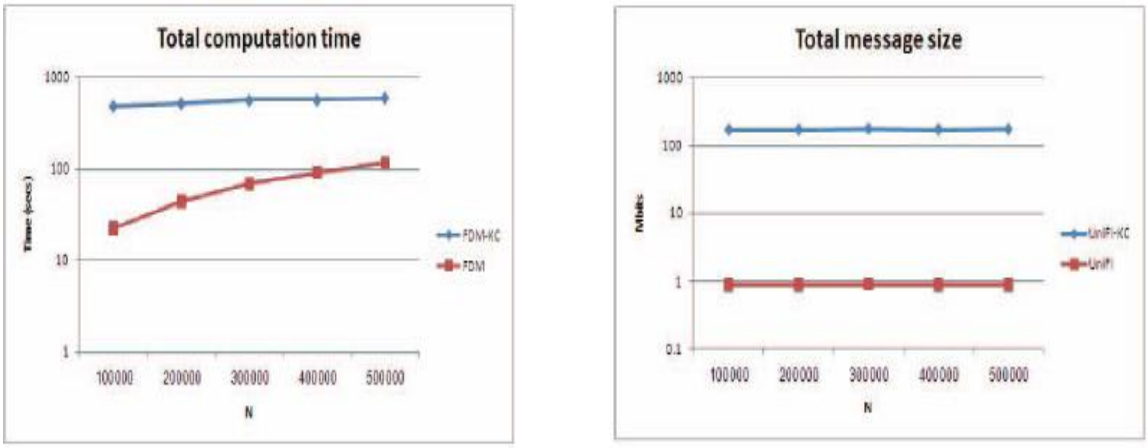


Figure 1: Computation and communication costs versus the number of transactions N

Comparatively study of FDM and FDM-KC is shown in the figure 1 and figure 2. Here in figure 1 two graphs shows the number of transaction N has little effect on the runtime of FDM-KC.
The second set of graph in figure 2 illustrates how the comm

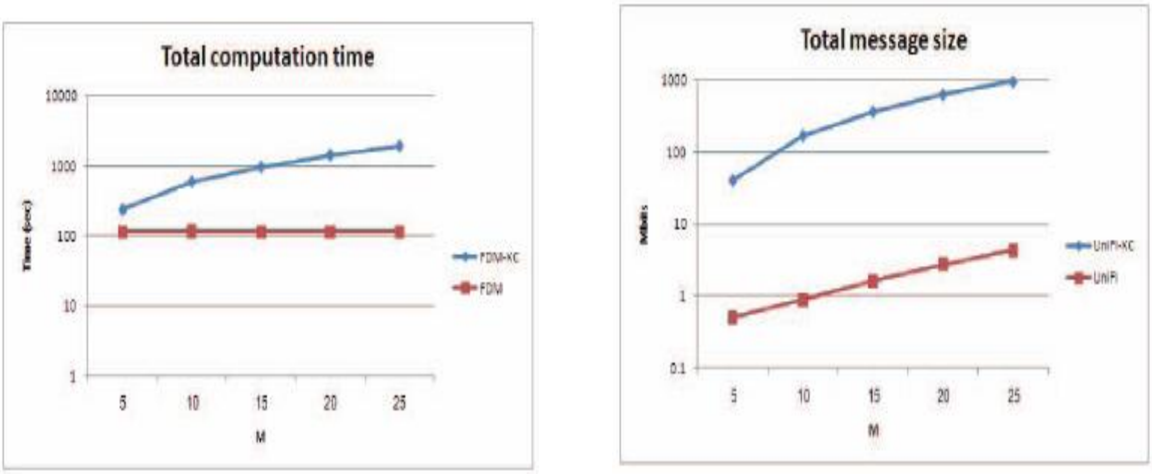


Figure 2: Computation and communication costs versus the number of players *M*

V. Conclusion

In this project we proposed a protocol for secure mining of association rules in horizontally partitioned distributed databases. The protocol is more efficient than current leading K and C protocol. The main ingredients of this protocol are two novel secure multiparty algorithms in which these two main operations are union and intersection. The protocol exploits the fact that the underlying problem is of interest only if the number of player is more than two.

In this paper, we proposed and studied an efficient and effective distributive algorithm FDM and FDM-KC for mining association rules. Some interesting properties between locally and globally frequent item sets are observed which leads to an effective technique for the reduction of candidate sets in the discovery of large item sets.

The direction to future work is to devise an efficient protocol for inequality verifications that uses the existence of semi-honest third party and another in Implementation of the techniques to the problem of distributed association rule mining in vertical setting.

References

[1] Tamir tassa, "Secure Mining of Association Rules in Horizontally Distributed Databases", IEEE transactions on knowledge and data engineering, 2013.

[2] J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data," in *The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada, July 23-26 2002, pp. 639–644.

[3] M.Kantarcioglu and C. Clifton., "Privacy-preserving distributed mining of association rules on horizontally partitioned data", *IEEE Transactions on Knowledge and Data Engineering*, 16:1026–1037, 2004.

[4] R.Agrawal and R. Srikant.,"Privacy-preserving data mining", *SIGMOD Conference*, pages 439–450, 2000.

[5] A.V. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules", In *KDD*, pages 217–228, 2002.

[6] M. Kantarcioglu, R. Nix, and J. Vaidya,"An efficient approximate protocol for privacy-preserving association rule mining", In *PAKDD*, pages 515–524, 2009.

[7] M. Freedman, Y. Ishai, B. Pinkas, and O. Reingold., "Keyword search and oblivious pseudorandom functions", In *TCC*, pages 303–324, 2005.