International Journal of Scientific Research in Computer Science and Engineering
Research Paper Volume-3, Issue-1 ISSN: 2320-7639

Secure Multiparty Protocol for Distributed Mining of Association Rules

M.Manigandan[1*] and K. Aravind Kumar[2]

[1*,2] Dept. of Computer Science and Engg., Manonmaniam Sundaranar University, Tamilnadu, India
manimano2@gmail.com[1], aravindkumarmk@gmail.com[2]

Available online at www.isroset.org

Received: 10 Dec 2014 Revised: 28 Dec 2014 Accepted: 22 Jan 2015 Published: 28 Feb 2015

Abstract **-** Association rule mining is one of the data mining tasks to find the association between the items or among the item sets. The privacy concept arises when the database is in the distributed environment in which each database holder is interested to discover the globally supported association rules by participating themselves in mining process without revealing its individual locally supported association rules. The present leading protocol consumes computational overhead as well as communication overhead in order to preserve the privacy in mining. A new protocol is proposed to reduce the communication cost and computational cost for secure mining in homogeneous databases. The proposed protocol uses the secure sum computation and set union computation for determining global association rules. It is significantly more secure and efficient as compared with the existing protocol.

Keywords- *Privacy Preserving Data Mining; Frequent Item Sets; Association Rules; Secure Multiparty Protocol; Distributed Mining; Homogeneous Databases; Secure Sum Computation*

I. INTRODUCTION

Data mining extracts the useful information from huge amount of business data sets. Association rule mining discovers the association rules from a large set of data items. Finding association rules among huge amount of business transactions can help in making many business decisions. Association rules are the implication statements that help to uncover relationships between unrelated data in a database, relational database or other information repository. Association rules are used to find the relationships between the objects which are frequently used together. For example, if the customer buys bread then he may also buy butter. There are two basic criteria that association rules uses, support and confidence.

Distributed database is defined as collection of logically related database which are joined with each other in a network. It can be of three modes. They are homogeneous distributed database, heterogeneous distributed database and mixed mode of them. Each party has its own individual database, applications and operating system. Privacy is to be maintained while sharing of individual data to mine the global information. Discovering knowledge through a combination of different databases raises security issue. Although data mining results usually do not violate secrecy of individuals, it cannot be assured that an unauthorized person will not access the data that is partitioned over different sites, it is impossible to derive new knowledge about the other sites. Privacy preserving mining offers the solution for it.

Corresponding Author: *M.Manigandan*

Privacy preserving data mining try to discover the patterns or information, which are unknown and hard to uncover by individuals in distributed environment. However to find such disclosure of patterns, the mining process has to access and use individual information. Privacy preserving data mining deals with how to extract the useful knowledge from the database. Afterwards, it essentially deals with the privacy of the database instances. The main goal of privacy preserving data mining is to agree to useful aggregate computations on the complete data set through preserving the privacy of the individual party's data or information. Particularly in distributed data mining, privacy preserving mining plays a vital role.

Each party can mine the locally supported association rules in horizontally distributed database environment. In order to discover the globally supported association rules, each party is interested to work together in obtaining them. The work of this paper is to propose a new protocol that discovers the globally supported association rules without violating the privacy of individual party with low computational cost and communication cost.

II. BACKGROUND AND RELATED WORK

Cheung, Ng, C. Fu and Y. Fu [4] proposed an algorithm for distributed association rule mining namely Fast Distributed Mining (FDM). But it is unsecure version of distributed mining algorithm. Privacy preserving mining was studied in the work of paper [1]. The aim of that paper was to protect the individual information from other database holder. For sharing the information, each database holder performs the data perturbation for anonymizing the data before releasing the data. The idea behind is that the

released information can be used to discover the global information without revealing original information. But it leads to generate the false association rules.

The goal is to mine the correct data whereas keeping the secrecy of each database holder from the other database holder. The work of paper [9] considered the problem in the horizontal setting in large-scale systems in which there is no collusions occurred between the network parties. Their work is different from the paper [7]. However, it consumes the communication overhead between the different network parties whereas going for cryptographic methods in order to keep the privacy of individual information.

Freedman, Nissim and Pinkas [6] proposed a secure protocol for set intersection operations. It can be used to perform the union of private sets through its complements. It can be applied for the unification of frequent item sets. Privacy preserving set operations that are closely related to the threshold function was studied in the work of [8] and can be applicable to the proposed system.

Brickell and Shmatikov [3] resulted in communication overhead that is in the logarithm size of candidate item sets. It is suitable for the case of two parties. The generic solution for the multiparty case was proposed in [2]. The problem of privacy preserving mining on heterogeneous database was studied in [11]. It gives the scalar product protocol with reasonable communication cost. That paper work was applicable only for vertical data setting and two party cases. It is not suitable for multiparty case and homogeneous databases.

Kantarcioglu and Clifton [7] proposed a protocol for secure distributed mining of association rules. The main part of the protocol is the secure computation of the union of frequent item sets that are held by the different database holders. It consumes cryptographic and hash functions overhead. Apart from that it leaks the excess information. The semi trusted party was included in their protocol. It is not adapted to the case where there is no semi-trusted party or trusted party. Their protocol improves efficiency and privacy with respect to the algorithm in the paper [4].

Tamir [10] proposed an alternative protocol named unified protocol that is the current leading protocol for unifying the frequent item sets. The unified protocol improves efficiency and privacy with respect to that in [7]. The unified protocol calls two computation function named set inclusion function and threshold function. So it leads to the overhead of computational cost and communication cost.

III. PROBLEM DEFINITION

In horizontally partitioned databases, there are several parties that hold homogeneous databases, i.e., databases that contain the information in same schema on different entities. The goal is to find all global association rules with given minimum support percentage s % and minimum confidence percentage c % that hold in the database of all parties, without disclosing the information about the individual information.

In the presence of a trusted party, the parties involved in the process of global computation could surrender their individual information to the trusted party, and then the trusted party performs the global computation and sends to them the resulting output. To maintain the privacy of information, no trusted party is considered. In such cases, a protocol is needed to perform all global computation in which all parties can run on their own information in order to arrive at the global output. Such a protocol is considered secure if no party can know the privacy information of others and it can know only the global information. The proposed protocol provides the solution for that problem with low communication and computation cost as compared with other existing protocols. The overhead cost can be still reduced by removal of set inclusion function and the simplification of the computation.

IV. SECURE MULTIPARTY PROTOCOL

The proposed protocol securely computes the union of frequent item sets of all parties. It is same as the problem of computing the union of private subsets. Indeed, if the ground set is $\Omega = \{\omega_1, \omega_2,...\omega_n\}$, then any subset B of Ω may be described by the characteristic local binary vector $b = (b_1,b_2,.....,b_n)$ where $b_i = 1$ if and only if $\omega_i \in B$. The proposed protocol is much simpler to understand and more efficient than the existing solutions. It utilizes to compute in a privacy preserving manner unions of private subsets with the help of secret sharing scheme.

The main idea behind the protocol is to use the secure summation protocol in order to compute the sum of shares and then use those shares to securely verify the threshold conditions in each component. Each player has its own individual shares that should not be revealed to the others. Each player starts by creating a random share and divides its share to the number of parties. Each player sends its random share to the next party. The next one adds its respective share to the random share for resultant. The summand is sent continuously after adding its respective share by the subsequent party. This process is continued until it reaches the originator. When the secret summand reaches to the originator, the originator party removes its random share and adds its original share. Players involved in the protocol send the sum vector to player P_1. Player P_1

receives the shares of all players and then performs the union operations on the received shares to obtain the global share. The Secure Multiparty Protocol is shown in figure 4.1 for secure mining of association rules on horizontally distributed database. The secure multiparty protocol is described in the module 2 section of System Methodology.

SECURE MULTIPARTY PROTOCOL

Input: Each Player P_i has frequent-k-itemset $F_s^{k,i}, \forall 1 \leq i \leq M$

Output: Global bit vector $b: = T_t(b(1), b(2), .., b(n))$

Procedure:

1. Each player P_i encodes its frequent $F_s^{k,i}$ of length $n_{k..}$
2. P_i partitions the local bit vector into M shares s_i.
3. P_i selects a random bit vector m_i and sends to $P_{i+1.}$
4. P_{i+1} adds its respective share s_i to m_i for summand $S_{i.}$
5. P_{i+1} sends S_i to continue with next party.
6. P_{i+1} verifies the integrity where P_i and P_{i+1} agree on $h_K(.)$.
7. This process is continued until it reaches the initiator $P_{i.}$
8. P_i computes the global partitioned vector b_i as in follows.
   calculate $S_i = (S_i - m_i) + s_i$.
   for each index x in b_i do
   if $(S_i \geq t)$ then
   set $b_i(x) = 1$;
   else
   set $b_i(x) = 0$;
   end if
   end for
9. Each player P_i sends its global partitioned vector b_i to P_1.
10. P_1 computes $b:=b_1 \cup b_2 \cup b_3$ ..... b_i and broadcasts it.

Fig. 4.1 Secure Multiparty Protocol

V. SYSTEM METHODOLOGY

The inputs are the individual databases and the output is the globally supported association rules that hold in the united database with the given minimum support s and minimum confidence c. The architecture of proposed system is shown in figure 5.1. The proposed system consists of three modules. They are

1. Generation of Local Frequent Item Sets
2. Generation of Global Frequent Item Sets
3. Discovery of Association rules

A. Generation of Local Frequent Item Sets

Each party mines the data from their database to compute the locally supported frequent item sets using AprioriTid Algorithm. The support of frequent item sets is inputted from the user. Once the local frequent-k-itemsets are computed then it is encoded into local bit vector and check whether they are globally frequent or not as described in Module 2. Then proceed to determine the local frequent-$(k+1)$-itemsets using the global frequent-k-itemsets and check whether they are globally frequent or not. This process is repeated until there is no more frequent-k-itemsets locally.

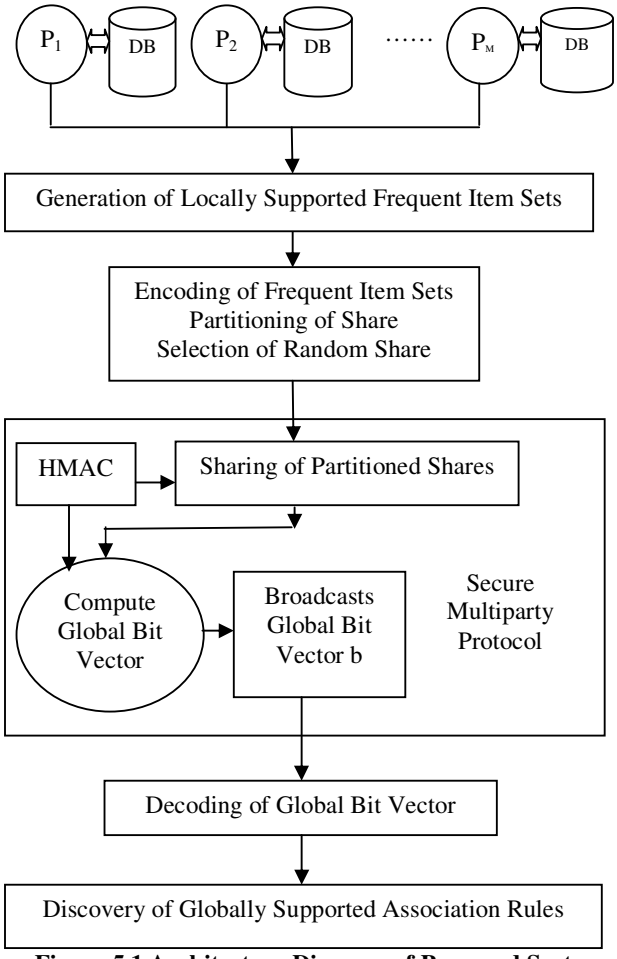

Figure 5.1 Architecture Diagram of Proposed System

B. Generation of Global Frequent Item Sets

After the local frequent-k-itemsets are computed, the proposed protocol is invoked for computation of global frequent-k-itemsets. The proposed protocol uses secure sum computation to sum the bit vectors of individual party. Each party partitions its local bit vector into local partitioned bit vector s_i according to the number of parties involved in the computation process and selects a random bit vector m_i. Each player P_i initiates the process for the

computation of global partitioned bit vector b_i by sending m_i to P_{i+1}. Upon receiving a random bit vector m_i from P_{i-1}, each party P_i adds its own corresponding partitioned bit vector s_i to the random number m_i and sends the summand S_i to P_{i+1}. Upon receiving a summand bit vector S_i, each party P_i adds its own corresponding partitioned bit vector s_i to the summand S_i. This process is continued until the secret summand S_i reaches the originator party P_i. When this process gets back to P_i, it removes its own random bit vector m_i from whatever is got from P_{i-1} and adds its own corresponding partitioned bit vector s_i to the summand S_i. Then apply the threshold function on summand S_i to get the global partitioned bit vector b_i. The threshold value t is agreed by all parties. The above computation process is repeated for computing all global partitioned bit vectors.

Each party P_i has the responsibility to compute the global partitioned bit vector b_i that is assigned. Global bit vector b for frequent-k- itemsets is computed by $b= b_1 \cup b_2 \cup b_3 \ldots \cup b_m$, where m is the number of players. The party P_1 broadcasts the global bit vector to all parties involved in the process. Then each party decodes into the global frequent-k-itemsets. The final global bit vector is determined when there is no more global frequent-k-itemsets. For verifying the integrity of message, it utilizes Hash based Message Authentication Code (HMAC). SHA-2 is utilized for Hash Function algorithm with Message Authentication Code.

C. Discovery of Association Rules

Once the final global bit is received, it is decoded into the final globally supported frequent item sets. Each player generates the combination of frequent item sets to form the association rules. The association rules must be satisfied with support level at least s % and confidence level at least c %. For item sets $X, Y \in F_s$, the association rule $X \rightarrow Y$ has the confidence level at least C if the following equation (1) is satisfied. The condition can be expressed as

$$\frac{Sup\ (X \cup Y)}{Sup\ (X)} \geq C \qquad (1)$$

The association rules which satisfies the equation (1) are only discovered for the given confidence value.

VI. Experimental Evaluation

A. Experimental Setup

The synthetic database is used in this experimental evaluation. The generated synthetic database D is divided into M partially distributed databases, D_m. The mining was implemented in Java. Microsoft SQL Server 2012 is used for storing database. The experiments were executed on an Intel Core i7 CPU, 4 GB of RAM and Windows 8.1.

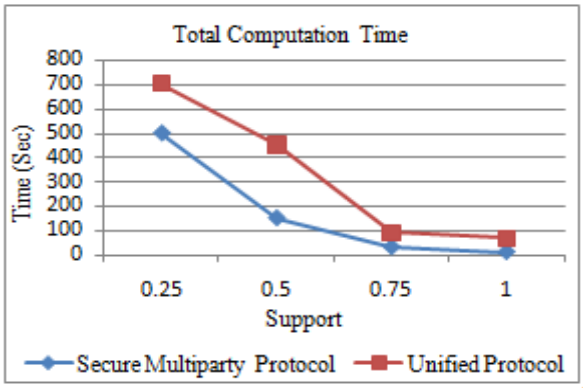

Fig. 6.1 Computational Cost

B. Comparison with Existing Methods

Performance of the existing protocol and proposed protocol is evaluated in terms of communication cost and computational cost. The computational cost involves the time needed for discovering the final global frequent item sets. The performance analysis of computation time of the proposed protocol across various support levels is shown in figure 6.1. The protocol reveals no false association rules, in addition, all globally supported association rules are generated; it gives efficient and secure mining in comparison to the existing protocol. The communication cost can be evaluated by calculating the number of communication rounds. For, $1 \leq k \leq K$, the proposed protocol takes $M\ (K+1)$ communication rounds, where M is the number of parties and K is the number of iterations.

VII. Conclusion

The secure multiparty protocol that is proposed here significantly improves efficiency upon the unified protocol which is the leading protocol currently for secure mining of globally supported association rules in horizontally partitioned databases. It reduces the communication rounds needed to compute the global frequent item sets. The experiment results show that it decreases the computational cost compared with the existing protocol. In this study, other research problems suggests that is the proposed protocol can be adapted to the problem of discovering the generalized association rules, multilevel association rules and quantitative association rules on homogeneous distributed databases.

References

[1] R. Agrawal and R. Srikant, "Privacy Preserving Data

Mining", *Proc. ACM SIGMOD Conf.,* pp. 439-450, 2000.

[2] A. Ben-David, N. Nisan, and B. Pinkas, "Fairplay MP - A System for Secure Multiparty Computation", *Proc. 15th ACM Conf. Computer and Comm. Security* (*CCS),* pp.257-266, 2008.

[3] J. Brickell and V. Shmatikov, "Privacy Preserving Graph Algorithms in the Semi-Honest Model", *Proc. 11th Int'l Conf. Theory and Application of Cryptology and Information Security* (*ASIACRYPT),* pages 236-252, 2005.

[4] D.W.L Cheung, V.T.Y. Ng, A.W.C. Fu, and Y. Fu. "Efficient Mining of Association Rules in Distributed Databases", *IEEE Trans. Knowl. and Data Eng.,* Vol.8, No.6, 911-922, 1996.

[5] V. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. "Privacy Preserving Mining of Association Rules", *Proc. Eight ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining* (*KDD), pp*. 217-228, 2002.

[6] M. J. Freedman, K. Nissim, and B. Pinkas, "Efficient Private Matching and Set Intersection", *Proc. Int'l Conf. Theory and Application of Cryptographic Techniques (EUROCRYPT)*, pp.1-19, 2004.

[7] M. Kantarcioglu and C. Clifton, "Privacy Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data", *IEEE Trans. Knowl. and Data Eng.,* Vol. 16, No.9, 1026-1037, 2004.

[8] L. Kissner and D.X. Song, "Privacy Preserving Set Operations", *Proc. 25th Ann. Int'l Cryptology Conf. (CRYPTO),* pages 241-257, 2005.

[9] A. Schuster, R. Wolff, and B. Gilburd, "Privacy Preserving Association Rule Mining in Large Scale Distributed Systems", *Proc. IEEE Int'l Symp. Cluster Computing and the Grid (CCGRID),* pp. 411-418, 2004.

[10] Tamir Tassa, "Secure Mining of Association Rules in Horizontally Distributed Database", *IEEE Trans. Knowl. and Data Eng.,* Vol.26. No.4, 970-983, 2014.

[11] J. Vaidya and C. Clifton, "Privacy Preserving Association Rule Mining in Vertically Partitioned Data", *Proc. Eight ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining* (*KDD), pp*. 639-644, 2002.

AUTHORS PROFILE

M. Manigandan has received B.Tech. in Information Technology from Bharathidasan Institute of Technology, Trichy, India. He is pursuing his M.E. in Computer Science and Engg., from Manonmaniam Sundaranar University, Tamilnadu, India in 2015. His subjects of interest include Cloud Computing and Data Mining.

K. Aravind Kumar. M.E. is currently working as a Assistant Professor in Dept. of Computer Science and Engineering, Manonmaniam Sundaranar University, Tamilnadu, India. His subjects of interest include Computer Networks and Data Mining.

