ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

Privacy Preserving Secure Mining of Association Rules in Relational Databases

Maheshkumar Ramrao Gangasagare[1], Rafik Juber Thekiya[2]

Post Graduate Student, Dept. of C.S.E., M.P.G.I., School of Engineering, Nanded, MH, India.[1]

Assistant Professor, Dept. of C.S.E., M.P.G.I., School of Engineering, Nanded, MH, India.[2]

ABSTRACT: Data mining is a vital approach in the Knowledge Discovery from data (KDD) to find relevant data to our task. It is available to explore helpful hidden information from large databases. The vital role of this implementation is to find union of private subsets that each of the interacting players holds. The second most important component are set of rules that tests the inclusion of an element held by one player in a subset held by another. These set of rules uses the fact that the fundamental problem is of interest only when there is the number of players is greater than two. Besides, association rule mining has wide applications to discover interesting relationships relevant to task among attributes. While safekeeping the large databases becomes a serious issue when that data shared on the network against multiple or unauthorized access.

KEYWORDS*:* Secure Multiparty Computation, Privacy Preserving Association Rule, Horizontally Partitioned Database, Hash Based Secure Sum.

I. INTRODUCTION

Day by day the expansion of demands for knowledge discovery in all industrial sections, it is important to store all the unfinished data and to provide useful patterns or relevant data with respective to the user needs. Generally, the storage of all unfinished data will be done in a central database maintained by respective organizations. These techniques are available to get useful information from large database. Forecast the rules and their detailed descriptions are the two essential goals of data mining.

To get these goals many data mining methods exists such as classification, association rules, clustering and so on. Between such types of association rules has wide covering or thrust to discover interesting relations or relevant data to analysis task among attributes in large databases. Such type of association rules mining is utilized to find the rules which fulfill the user specified minimum support and minimum confidence. In the outgrowth of finding association rules, the set of frequent item sets are calculated as the primary step and then association rules are produced based on these frequent item sets. The enhancement of computing technology from last many years are used to handle complex data that includes huge transaction financial data, bulletins, emails etc. At last information has become a resource that made sophesticated for user to get their opinions and interacts. As a conclusion revolves around the practice, data mining came into sites. These rule mining is one of the Data Mining techniques used in distributed database to get relevant data for analysis task.

In distributed database the data may be divided among multiple fragments and every fragment is attached to one site. The problem of privacy emerged when the data is divided among multiple sites and no any other party like to provide their individual private data to their sites but vital step is to get the global result obtained by the mining process. At such time the privacy preserving data mining came into the picture as a solution. As I divided the database, multiple users can approach it without mention of one another. In distributed prototype, database is divided into disjoint fragments and every site has of only one fragment. Data can be divided into three different ways, that is, vertical partitioning, horizontal partitioning and mixed partitioning.

II. Related work

In privacy conservation distributed data mining, how the data is divided between different sites is very important. The three most important partitioning ways in distributed database setting are horizontal, vertical and mixed mode. When looking to the horizontal partition, the same schema is used to maintain the data at each site where in vertical partition; multiple schemas are used at different sites, that is, multiple types of data on the same entities. Third method is mixed partitioning where data is divided horizontally and then each part is further divided into vertical and vice versa. Privacy preserving association rule mining algorithms can be sorted into three groups according to privacy protection technologies. The three groups are heuristic-based techniques, cryptography-based techniques and reconstruction-based techniques. In this paper Fast distributed mining algorithm is taken into consideration to get global association rules by preserving the privacy when no party can be treated as trusted party. This approach is much popular because of the following two reasons:

- It has a well-known and definite model intended for privacy which can really present good number of methodologies for verifying and validating purpose.
- Fast Distributed Mining algorithm has a wide variety implementation to take privacy into consideration while data mining.

A description of data mining techniques and a detailed analysis of mining association rules are depicted by me and this approach is done keeping in mind that database researcher's point of aspect based upon the data mining techniques. I also studied multiple classes of data mining techniques and its distinct features exist between them [1]. Secure two party computations ides was first studied by Yao [2], and later widespread to multi party computation. In [3], the publisher presented ID3 classification when two parties with horizontally partitioned data by applying secure protocols to get complete zero knowledge outflow. The authors proposed in [4], four efficient methods namely secure set union, scalar product, secure size of set intersection and secure sum for privacy preserving data mining in distributed setting. In [5], the authors studied the issues of privacy preserving data mining of association rules when the data is partitioned horizontally. They implemented algorithm which consider three basic facts such as secure computation, encryption of site results and randomization. The implementation approach in the area of privacy preserving data mining techniques is implemented [6]. The authors also studied about privacy preserving algorithms and classifications of privacy preserving techniques such as reconstruction based technique, cryptography-based techniques and heuristic-based techniques. A model for implementation of privacy preserving data mining algorithms and based on this model effort one can evaluate the multiple features of privacy preserving algorithms as per the different assessment criteria [7]. An enhanced kantarcioglu and Clifton's approach is considered by authors in [8], which is a two segment for privacy preserving distributed data mining. In [9], the authors studied the problem of distributed data bases while considering privacy preservation of data mining. They recommended a new model based on two different entities a calculator and a minor both aren't taking any parts of the database. They also demonstrated three algorithms based on this prototype one for any data mining method, one for vertically partitioned data and one for horizontally partitioned data. The publisher in [10], implemented a new algorithm for mining association rules in distributed homogeneous databases which have a base on semi honest prototype and insignificant collision probability. In [10], the publisher studied clustering of various association rule mining algorithms, an extended description and classification. They also proposed future research ways of privacy preserving association rule mining algorithms by examining the existing work.

III. Proposed algorithm

A. Definitions, Notations and its Description

Let D be a transaction database. In paper [5], we see D as a binary matrix of L columns and N rows, where each row consist of transaction over some set of items, A = $\{a_1, . . . , a_L\}$, and each column contain one of the items in A. (In simple words, the $\{i , j\}$ entry of D equals 1 if the i^{th} transaction contain the item a_j, and 0 if no item present.) The database D is divided horizontally between M players, referred $P_1, . . . , P_M$. Player P_m contain the partial database D_m that have N_m=| D_m | of the transactions in D, $1 \leq m \leq M$. The united database is union of D = $D_1 \cup . . . D_M$, and it includes N := $\sum_{m=1}^{M} N_m$ transactions. An universal itemset A have a subset X.

Let,

supp (X) = Number of transactions in D that have a global support.
$\mathrm{supp}_m(X)$ = Number of transactions in D_m that have a local support.

Distinctly concluded, supp(X) $=\sum_{m=1}^{M} supp_m\ (X)$.

Let, s means threshold be a real number between 1 and 0 that is used for a required support threshold. If supp(X) ≥ ${}_sN$ then that item set X is called s-frequent. If supp_m (X) ≥ ${}_sN_m$ then it is called locally s-frequent at D_m. For every item 1 ≤ k ≤ L.

Let ,

F_s^k = Set of all k-item sets that are s-frequent globally for i-itemsets.
$F_s^{k,m}$ = Set of all k-item sets that are locally s-frequent at D_m.

Our main step is to find, for a given threshold support 0 < s ≤ 1, then we distinctly conclude that the set of all s-frequent item sets, $F_s := \bigcup_{k=1}^{L} F_s^k$. Then we continued to find all (s, c)- association rules, that possess support at least ${}_sN$ and confidence at least c to find all association rules. (Again considering that if X and Y are two disjoint subsets of universal set A, the support of the related association rule X ⇒ Y is confidence is supp(X ∪ Y)/supp(X) and its supp (X ∪ Y)).

B. The Fast Distributed Mining Algorithm

The set of rules of [5], as well as in my approach, are depends on the Fast Distributed Mining (FDM) algorithm, that algorithm is an unsecured distributed version of the Apriori algorithm. Its main theme is that locally s-frequent in at least one of the sites must also search in other sites also. Here, I have to locate all globally s-frequent item sets, each player bring out his locally s-frequent item sets and then the players will verify each of them to see if these are s-frequent globally also. The FDM algorithm implemented as follows:

1. **Initialization:** It is assumed that the players have already jointly calculated F_s^{k-1} . The goal is to proceed and calculate F_s^k.
2. **Candidate Sets Generation:** Each player P_m computes the set of all (k - 1)- item sets that are globally frequent in his site and also locally frequent; namely, P_m computes the set $F_s^{k-1,m} \cap F_s^{k-1}$. I then applied on that set the Apriori algorithm in order to generate the set $B_s^{k,m}$ of candidate k-item sets.
3. **Local Pruning:** For each X ∈ $B_s^{k,m}$, P_m computes supp_m(X). He then retains only those item sets that are locally s-frequent. We denote this collection of item sets by$C_s^{k,m}$.
4. **Unifying the candidate item sets:** Each player broadcasts his $C_s^{k,m}$ and then all players compute $C_s^k := \bigcup_{m=1}^{M} C_s^{k,m}$.
5. **Computing local supports**: All players compute the local supports of all item sets in C_s^k.
6. **Broadcast mining results:** Each player broadcasts the local supports that he computed. Among that, each one can work out the global support of every item set in C_s^k . Lastly, F_s^k is the subset of C_s^k that depends on all globally s-frequent k-item sets.

In the first iteration, when k = 1, the set $C_s^{1,m}$ that the m^{th} player computes (Steps 2-3) is just $F_s^{1,m}$, namely, the set of single items that are s-frequent in D_m. The complete FDM algorithm proceeds from by ending all single items that are globally s-frequent. It then started from getting all 2-item sets that are globally s-frequent, and so on, till it ends the longest globally s-frequent item sets. If the length of such item sets is K, then in the next (k + 1)th iteration of the FDM it will search no (k +1)- item sets that are s-frequent globally, in that case only it terminates.

C. A Running Example

Let D be a database of N = 18 item sets over a set of L = 5 items, A = {A, B, C, D, E}. It is partitioned among M = 3 players and the respective partial databases are:

D_1 = {AB, ABCDE, ABD, ABDE, AD, ADE, BCE, BD, BD},
D_2 = {ABCD, ACD, BC, BCD, BCDE},
D_3 = {ABCD, ABD, ACD, BC}.

While solving to example D_1 contains N_1 = 9 transactions, the third partial database of which (in alphabetic order) consists of three items—A, B and D.

Taken into granted s = 1/3 (threshold value), an item set which is globally s-frequent in D then it is supported by at least 6 = sN of its transactions. In this case,

$$F_s^1 = \{A, B, C, D\},$$
$$F_s^2 = \{AB, AD, BC, BD, CD\},$$
$$F_s^3 = \{ABD\},$$
$$F_s^4 = F_s^5 = \emptyset.$$

and $F_s = F_s^1 \cup F_s^2 \cup F_s^3$. For example, the item set CD is indeed globally s-frequent since it is contained in 7 transactions of D. At last, it is locally s-frequent only in D_3 and D_2.

In the first round of the FDM algorithm, the three players compute the sets $C_s^{1,m}$ of at each player m all 1-item sets that are locally frequent at their partial databases:

$$C_s^{1,1} = \{A, B, D, E\},$$
$$C_s^{1,2} = \{A, B, C, D\},$$
$$C_s^{1,3} = \{A, B, C, D\}.$$

Hence, C_s^1 = {A, B, C, D, E}. Consequently, all 1-item sets have to be patterned for existence globally frequent; that check reveals that the subset of globally s-frequent 1-item sets is F_s^1 = {A, B, C, D}.

In the second round, the candidate item sets are:

$$C_s^{2,1} = \{AB, AD, BD\},$$
$$C_s^{2,2} = \{AC, AD, BC, BD, CD\},$$
$$C_s^{2,3} = \{AB, AC, AD, BC, BD, CD\}.$$

(Note that AE, BE, DE are locally s-frequent at D_1 but they are not included in $C_s^{2,1}$ since E were already found to be globally infrequent.) Hence, C_s^2 = {AB, AC, AD, BC, BD, CD}. Then, after verifying global frequency, we are left with F_s^2 = {AB, AD, BC, BD, CD}.

In the third round, the candidate item sets are:

$C_s^{3,1}$ = {ABD}, $C_s^{3,2}$ = {BCD}, $C_s^{3,3}$ = {ABD}.

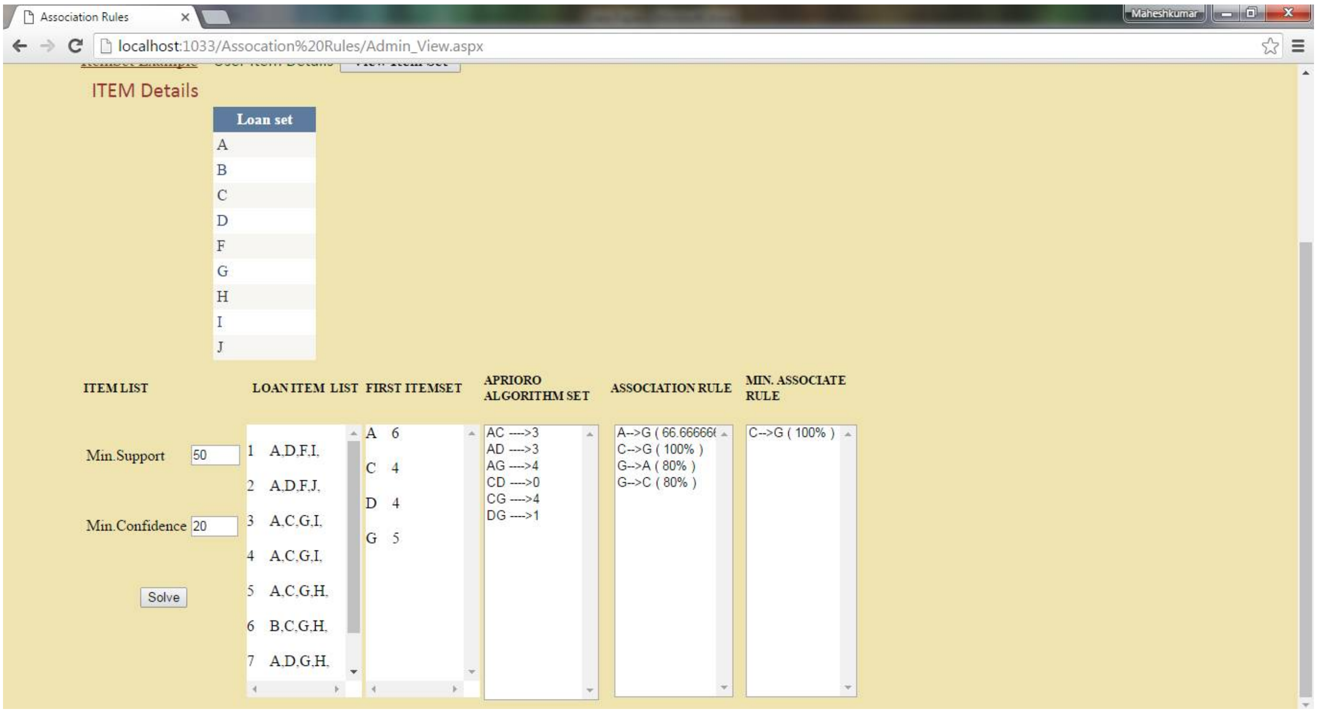
So, C_s^3 = {ABD, BCD} and, then, F_s^3 = {ABD}. There are no more frequent item sets.

D. Implementation Overview

Our main goal is to calculate or find all association rules which are relevant to our task. In my case I used 9 variables for my consideration. In that case (A, B) variables stands to consider that user is Male or Female respectively. Then (C, D) variables stands for age of user < 25 or >25 respectively. Then (F, G) variables stands for finding that user are student (Yes) or not (No) respectively. At last (H, I, J) variables stands for finding that user purchased which type of loan Education, Car and Housing respectively.

These three slots of variables belong to different tables. Our main interested key is to find association rules which are divided among multiple tables. Every user information is on the basis of priority it is separated among multiple tables.

From below snapshot we conclude that for the 1-itemset the result is shown into the second column. Then for the next iteration i.e. for second iteration 2-itemset is considered and its result is shown in the third column. The number of iteration depends upon the number of columns contained by the tables. In the association rule column it will show all the rules along with its values calculated from given input. In the last column it represents all association rules which can be calculated from all given input. Also in our implementation we can depicts the rule form the graph also for the different vales i.e. for Deposit, Account Holder, Fund Transfer, Loan Request etc.



ITEM Details

Loan set
A
B
C
D
F
G
H
I
J

ITEM LIST

Min.Support 50

Min.Confidence 20

Solve

LOAN ITEM LIST

1 A,D,F,I,
2 A,D,F,J,
3 A,C,G,I,
4 A,C,G,I,
5 A,C,G,H,
6 B,C,G,H,
7 A,D,G,H,

FIRST ITEMSET

A 6
C 4
D 4
G 5

APRIORO ALGORITHM SET

AC ---->3
AD ---->3
AG ---->4
CD ---->0
CG ---->4
DG ---->1

ASSOCIATION RULE

A-->G (66.66666(
C-->G (100%)
G-->A (80%)
G-->C (80%)

MIN. ASSOCIATE RULE

C-->G (100%)

Figure: Association Rule Computation Details

IV. Conclusion and Future Work

The vital role of this implementation is to find union of private subsets that each of the interacting players holds. The second most important component are set of rules that tests the inclusion of an element held by one player in a subset held by another. These set of rules uses the fact that the fundamental problem is of interest only when there is the number of players is greater than two.

Our main goal of preserving privacy in association rule mining when the database is divided horizontally among multiple sites when no trusted party is considered. A prototype which considers a privacy preserving technique to find the global association rules which is relevant to our proposed task. In this paper we preserved the privacy of itemsets by applying some constraints data distribution. The implemented approach effectively finds global frequent item sets though can be treated as trusted. By looking towards sample databases, working of the proposed model is explained.

REFERENCES

1. Ming-Syan Chen, Jiawei Han,Yu, P.S. (1996), Data mining: an overview from a database perspective, IEEE Transactions on Knowledge and Data Engineering, Vol. 8 No. 6, pp 866 – 883.
2. A.C Yao(1986), How to generate and exchange secrets, In proceedings of the 27th IEEE Symposium on Foundations of Computer Science, pp 162-167.
3. Y Lindell and B pinkas (2000), Privacy preserving data mining, In Proc. O CRYPTO'00, pp36- 54. Springer-Verlag2000.
4. Chris Clifton, Murat Kantarcioglu, Jaideep Vaidya, Xiaodong Lin, and Michael Y. Zhu(2003), Tools for privacy preserving distributed data mining, SIGKDD Explorations, Vol. 4, No. 2 pp 1-7.
5. M. Kantarcioglu and C. Clifto (2004). Privacy-preserving distributed mining of association rules on horizontally partitioned data. In IEEE Transactions on Knowledge and Data Engineering Journal, volume 16(9), pp. 1026-1037.
6. Verykios, V.S., Bertino, E., Nai Fovino, I., Parasiliti, L., Saygin, Y., and Theodoridis, Y. (2004), State-of- the-art in privacy preserving data mining, SIGMOD Record, 33(1):50–57.
7. Elisa Bertino , Igor Nai Fovino Loredana Parasiliti Provenza (2005), A Framework for Evaluating Privacy Preserving Data Mining Algorithms, Data Mining and Knowledge Discovery, Vol. 11, 121–154.
8. Chin-Chen Chang, Jieh-Shan Yeh, and Yu-Chiang Li (2006), Privacy-Preserving Mining of Association Rules on DistributedDatabases, IJCSNS International Journal of Computer Science and Network Security, Vol.6 No.11.

9. Alex Gurevich, Ehud Gudes (2006), Privacy preserving data mining algorithms without the use of secure computation or perturbation, 10th international database Engineering and Applications Symposium IDEAS06 IEEE.
10. Jian Wang, Yongcheng Luo, Yan Zhao, Jiajin Le(2009), A Survey on Privacy Preserving Data Mining, First International Workshop on Database Technology and Applications, pp. 111-114.

BIOGRAPHY

Maheshkumar Ramrao Gangasagare, received the B.E. degree in Computer Engineering from Savitribai Phule Pune University, Pune in 2012 and diploma in Diploma in Advanced Computing from C-DAC institute Pune in 2013. He is a M.E. candidate at Swami Ramanand Teerth Marathwada University, Nanded.

Rafik Juber Thekiya, received the B.E. degree in Computer Science and Engineering from Swami Ramanand Teerth Marathwada University, Nanded in 2006. Also, he received M.Tech. in Computer Science and Engineering from Jawaharlal Nehru Technological University, Hyderabad in 2012. He served as a Lecturer for 3 years and Assistant Professor for 2 years.