Privacy Preserving Association Rule Mining in Horizontally Partitioned Databases Using Cryptography Techniques

N V Muthu lakshmi,Dr. K Sandhya Rani
Dept of Computer Science,S.P.M.V.V,Tirupati, Andhra Pradesh, INDIA

***Abstract*— Data mining techniques are used to discover hidden information from large databases. Among many data mining techniques, association rule mining is receiving more attention to the researchers to find correlations between items or items sets efficiently. In distributed database environment, the way the data is distributed plays an important role in the problem definition. The data may be distributed horizontally or vertically or in hybrid mode among different sites. There is an increasing demand for computing global association rules for the databases belongs to different sites in a way that private data is not revealed and site owner knows the global findings and their individual data only. In this paper a model is proposed which adopts a sign based secure sum cryptography technique to find global association rules with trusted party by preserving the privacy of the individual's data when the data is distributed horizontally among different sites.**

***Keywords*—DataMining, Distributed Database, Privacy Preserving Association Rule Mining, Cryptography Technique**.

I. INTRODUCTION

Data mining has been viewed as a threat to privacy because of the widespread proliferation of electronic data maintained by corporations. This has lead to increased concerns about the privacy of the underlying data. Data mining techniques find hidden information from large database while secret data is preserved safely when data is allowed to access by single person. Now a days many people want to access data or hidden information using data mining technique even they are not fully authorized to access. For getting mutual benefits, many organizations wish to share their data to many legitimate people but without revealing their secret data.

In large applications the whole data may be in single place called centralized or multiple sites called distributed database. Methodologies are proposed by many authors for both centralized as well as distributed database to protect private data. This paper deals with privacy preserving in distributed database environment while sharing discovered knowledge/hidden information to many legitimate people.

In distributed environment, database is a collection of multiple, logically interrelated databases distributed over a computer network and are distributed among number of sites. As the database is distributed, different users can access it without interfering with one another. In distributed environment, database is partitioned into disjoint fragments and each site consists of only one fragment. Data can be partitioned in different ways such as horizontal, vertical and mixed.

In horizontal partitioning of data, each fragment consists of a subset of the records of a relation R where as vertical partitioning of data, each fragment consists of a subset of attributes of a relation R. The another partitioning method is mixed fragmentation where data is partitioned horizontally and then each partitioned fragment is further partitioned into vertical fragments and vice versa [1]. Figure 1.a shows a method for mixed partitioned in which data is first partitioned vertically and then horizontally. Figure 1.b shows another mixed method in which data is partitioned horizontally and then vertically partitioned.

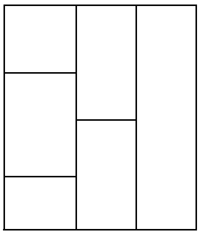

Figure1.a:Vertically partitioned database is further partitioned into horizontal

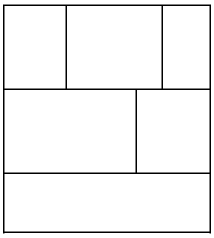

Figure1.b: Horizontally partitioned database is further partitioned into vertical.

In data mining, association rule mining is a popular and well researched method for discovering interesting relations between variables in large databases. When data is distributed among different sites, finding the global association rules is a challenging task as the privacy of the individual site's data is to be preserved. In this paper, a model is proposed to find global association rules by preserving the privacy of individual sites data when the data is partitioned horizontally among n number of sites.

II. DISTRIBUTED ASSOCIATION RULE MINING

Association rule mining technique is receiving more attention among data mining techniques to the researchers to explore correlations between items or item sets. These rules can be analyzed to make strategic decisions to improve the performance of the business or quality of the organization service and so on. Association rule mining was introduced by *Agarwal* [3]. An association rule can be defined formally as follows.

Let $I = \{i_1, i_2, \ldots, i_m\}$ be the set of attributes called items. The item set X consisting of one or more items. Let $DB = \{t_1, t_2, \ldots, t_n\}$ be the database consisting of n number of boolean transactions, and each transaction t_i consisting of items supported by i^{th} transaction. An item set X is said to be frequent when number of transactions supporting this item set is greater that or equal to the user specified minimum support threshold otherwise it is said to be infrequent. An association rule is an implication of the form X→ Y where X and Y are disjoint subsets of I, X is called the antecedent and Y is called consequent. An association rule X→Y is said to be strong association rule only when its confidence is greater than or equal to user specified minimum confidence.

Association rule generation has two steps. Computation of frequent item sets from the database based on user specified minimum support threshold is the first step and this process is difficult since it involves searching all combinations of item sets. In the second step, the association rules can be easily generated based on user specified minimum confidence threshold for the frequent item sets which are generated in the first step.

Now a days, more people who prefer to provide mutual benefit to their partners want to get access over association rules which are derived from large database even though they are neither owners nor possessing privileges to access. The database owners also wish to share their derived results that is association rules to get some benefits from them but they do not want to provide their secret data and also leakage of secret data may cause damage or loss. Sharing of knowledge is the main concern in some application for the mutual benefits in knowledge discovery system while preserving privacy of individual is another concern.

Distributed association rule mining is nothing but association rules computed globally from n number of sites in distributed environment by satisfying privacy constraints. Global support count of an item set X, can be computed as follows.

$$\text{X. support} = \sum X_i.\text{sup}$$

An item set, X is globally frequent when its support value computed from all sites is ≥ user specified minimum support threshold * |DB|.

An item set X which is locally frequent in a site may be infrequent globally and an item set can be locally infrequent in one or more sites may be globally frequent since global support is computed by considering the support value of the item set at all sites. The aim of the global association rule mining is to find all rules for each global frequent item set where global confidence is greater than or equal to the user specified minimum confidence. A global association rule can be expressed as follows.

A→ B, where A and B are disjoint subsets of I.

The global confidence of a rule is global support (AUB) / global support (A).

The rule A→B is said to be strong only when its global confidence ≥ user specified minimum confidence * |DB| otherwise it is weak rule.

In distributed environment, the challenging task is how efficiently one can provide accurate knowledge to their partners to have goodwill while no single secret data is revealed to them. This issue makes the researchers to study further to propose methods for privacy preserving association rule mining. The proposed model for privacy preserving association rule mining for horizontally partitioned distributed databases is explained in the following sections.

III. PRIVACY PRESERVING ASSOCIATION RULE MINING FOR HORIZONTALLY PARTITIONED DISTRIBUTED DATABASE

Many researchers proposed many methods for privacy preserving association rule mining for both centralized and distributed databases. The state of the art in the area of privacy preserving data mining techniques is discussed by the authors in [3]. This paper also describes the different dimensions of preserving data mining techniques such as data distribution, data modification technique, data mining algorithms, data or rule hiding and approaches for privacy preserving data mining techniques. The survey of basic paradigms and notations of secure multiparty computations and their relevance to the field of privacy preserving data mining are presented by the authors in [4]. They also discussed the issue of efficiency and demonstrate the difficulties involved in constructing highly efficient protocols. In [5], the authors proposed a framework for evaluating privacy preserving data mining algorithms and based on their frame work one can assess the different features of privacy preserving algorithms according to different evaluation criteria.

In [6], Secure mining of association rules over horizontally partitioned database using cryptographic technique to minimize the information shared by adding the overhead to the mining process is presented. In [7], authors proposed an enhanced *kantarcioglu* and *Clifton's* schemes which is a two- phase for privacy preserving in distributed data mining. They presented two protocols to increase the security against collusion in the communication environment with or without trusted party.

A new algorithm which is the modification of the existing algorithm and based on a semi-honest model with negligible collision probability is proposed in [8]. They also used cryptography techniques to preserve the privacy. Privacy preserving in data mining by using cryptographic role based access control is presented in [9]. They proposed a new solution by integrating the advantages of the first approach which protects the privacy of the data by using an extended role based access control approach and the second approach which uses cryptographic techniques with the view of minimizing loss of information and privacy. In [10], authors addresses the problem of association rule mining in vertically partitioned database by using cryptography based approach and also presents the analysis of security and communication. The history of secure multi party computation in two milliner's problem, in which they want to know who is richer without disclosing their wealth is addressed in [11]. Protocols are also proposed for two milliner's problem as well for m-party case.

In distributed environment when data is partitioned horizontally among different sites with a trusted party is considered in this paper. In horizontally partitioned distributed database, different set of records with same set of attributes of whole database are placed at different sites.The associations between items or item sets can be found correct only if rules are determined from results of total set of records from all sites. But no single site owner wishes to provide no single record to any site and this makes the issue a challenging one that is extracting necessary information from all sites data without accessing individual records to generate association rules.

In this paper, new methodology is proposed to find privacy preserving association rule mining for horizontally partitioned distributed database with Trusted Party (TP). The proposed method solves this problem by using a cryptography technique that is sign based secure sum. Cryptography technique protects data/ information from others accessing in a distributed database environment for semi honest model. In this paper to protect one's local frequent item sets from others accessing, public private key algorithm is used. The proposed sign based secure sum concept enhances the security by transmitting ambiguous results between sites in the process of generating frequent item sets and rules. In this method, a special site is designated and this site owner is called Trusted Party and initiates the process of finding association rules without knowing any one's individual data/information but by taking processed results from all the sites in secure manner.

A Proposed Model

In the horizontally partitioned distributed database model, there is n number of sites and every site owner has local autonomy over their database and one special site called Trusted Party (TP) who has special privileges to perform certain tasks.

The proposed method consisting of many tasks, performed by both TP as well as sites to find global association rules while preserving individual's private data. The following diagram shows the communication between TP and sites in the proposed model.

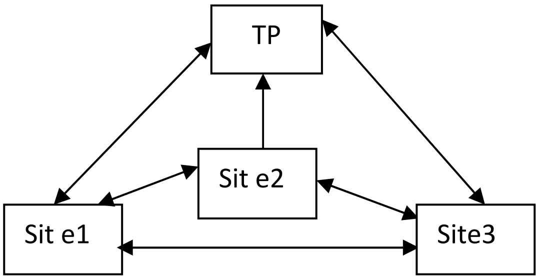

Figure2:Communication between sites and Trusted Party

In this proposed model, distributed database consists of n number of partitioned distributed databases and available in n number of sites termed as $Site_1$,$Site_2$, …, $Site_n$. The $Site_i$ maintains a database DB_i whose length is $|DB_i|$ where $1 \le i \le n$. Total number of transactions in all sites (|DB|) is

$$|DB| = \sum_{i=1}^{n} |DBi|$$

Every site requires global frequent item sets and supports to generate global association rules. So the goal is to determine global frequent item sets with supports based on the databases at all sites. Any item set is said to be globally frequent only when sum of support value of item sets at all sites is greater than or equal to minimum number of transactions required to support this item globally. An item set can be globally frequent only when it exists in at least one or more sites as frequent. Similarly an item set can be globally infrequent only when the item set is infrequent in at least one or more sites.

It is very clear that no one is willing to reveal their local frequent item sets, supports and database size to any site owner as well as to TP. To solve this problem the method provides special privileges to TP to capture local frequent item sets without taking the value of supports from all sites to determine all sites frequent item sets. Every site owner accepts to provide local frequent item sets in encrypted form to TP to whom they trusts to generate merged frequent item set list.

To find global association rules in horizontally partitioned distributed databases of size n (> 2), several tasks should be performed by both TP as well as site owners. In this model several terms are used and are shown in table I.

TABLE I TERMS USED IN THE PROPOSED MODEL

Terms	Description
AS_j	Actual Support
GES_j	Global Excess Support for item set X_j
PS_{ij}	partial support of item set Xj at $Site_i$
RN_i	random number for $Site_i$
$Sign_i$	Sign used with random number for $Site_i$
SignSumRN	sum of random numbers along with respective signs
$TotalPS_{ij}$	Sum of PS_{ij} of item set $X_{j,}$ where i indicates site number varies from 0 to n
TP	Trusted Party
MinSup	Minimum support threshold
MinConf	Minimum confidence threshold

The proposed model adopted sign based secure sum cryptography method to find global association rules by preserving the privacy. The various steps in the proposed model are as follows.

Step1. The first task is initiation done by the TP, and sends request to find frequent item sets to all sites by sending public key, minimum supports.

Step2. On receiving the public key and threshold, each site finds Local frequent item sets for their data by using the apriori algorithm. For their generated set of frequent item sets, every site applies encryption algorithm to convert frequent item sets into encrypted form using the public key and send it toTP.

Step3. TP then decrypts the each site's encrypted data by using Private key and prepares a merged list which consists of all site's local frequent item sets after eliminating duplicates. For each site, TP generates a unique random numbers and a sign (+ or -). The merged list along with respective random number (RN_i) and a sign ($Sign_i$) are sent to each site. The Sign field indicates whether the random number is to be added or subtracted from its partial support value (PS_{ij}).

Step4. Each site computes partial support for each item set in the merged list which is received from TP by using the formula

$PS_{ij} = X_j.sup - MinSup \times | DB_i | + (Sign_i) RN_i$

where i indicates the i^{th} site, ranges from 1 to n and j indicates j^{th} item set in the merged list, ranges from 1 to k. Each site then broadcast its computed PS_{ij} values for all the item sets in the merged list to all other sites.

Step5. Each site computes $TotalPS_{ij}$ for each item set X_j by using the formula.

$$\text{Total } PS_{ij} = \sum_{i=1}^{n} PS_{ij} \quad \text{for each } j = 1 \text{ to } k$$

and sends to the TP.

Step6. TP receives the $TotalPS_{ij}$ from all the sites for each item set X_j.

Step7. TP verifies the uniqueness of receiving Total-PS_{ij} from all sites. If any discrepancy exist then TP request all owners to perform step 5 once again to get the correct results.

Step8. TP computes Global Excess Support (GES_j) for each item set X_j by using the formula

$GES_j = TotalPS_{1j} - SignSumRN$

where SignSumRN is computed by adding all the random numbers with their signs by TP. If the computed value of $GES_j \geq 0$ then the item set X_j is globally frequent otherwise it is globally infrequent.

Step9. For each global frequent item set X_j, TP finds Actual Support (AS_j) as

$AS_j = GES_j + MinSup * |DB|$

$$\text{Where } |DB| = \sum_{i=1}^{n} |DB_i|$$

Step10. TP broadcast a list which consists of all global frequent item sets and their values to all sites.

Step11. Each site can generate association rules with various confidence values by using the globally frequent item sets and support values received from TP. The above process is explained with an example in the next section.

IV. IMPLEMENTATION OF THE PROPOSED MODEL WITH SAMPLE DATA

The proposed model is illustrated by using three horizontally partitioned distributed databases for finding privacy preserving association rule mining. In this sample model, the horizontally partitioned databases called fragments such as DB_1, DB_2 and DB_3 are placed in $Site_1$,$Site_2$ and $Site_3$ respectively. Apart from these three sites, there exist a special site called Trusted Party site. Sample databases at $Site_1$, $Site_2$and $Site_3$ are given below.

TABLE II.A DATABASE DB1, AT $SITE_1$

T-Id \Item	A_1	A_2	A_3	A_4	A_5
$Site_1$ has the following database					
T_1	1	0	0	1	0
T_2	1	1	0	1	1
T_3	0	1	1	0	1
T_4	0	0	1	1	1
T_5	1	1	0	1	1

TABLE II.B DATABASE DB2 AT $SITE_2$

T-Id \Item	A1	A_2	A_3	A_4	A_5
$Site_2$ has the following database					
T_1	0	1	1	1	1
T_2	0	0	1	1	1
T_3	1	1	1	1	0
T_4	1	1	0	1	1
T_5	1	1	0	0	1

TABLE II.C DATABASE, DB3 AT $SITE_3$

T-Id \Item	A_1	A_2	A_3	A_4	A_5
$Site_3$ has the following database					
T_1	1	0	0	1	1
T_2	1	1	1	0	1
T_3	1	0	1	1	1
T_4	1	0	1	1	0
T_5	1	0	1	1	1

TP request three sites to send encrypted form of local frequent item sets by sending two values such as minimum support threshold and public key. Each site computes local frequent item sets for their database by using minimum support threshold value 40% which is sent by the TP. The local frequent item sets (LF) of sites Site1, $Site_2$and $Site_3$, are given below.

Local frequent item sets at Site1

LF_1= { A_1 ,A_2,A_3 ,A_4,A_5,(A_1,A_2),(A_1,A_4),(A_1,A_5), (A_2,A_4),(A_2,A_5),(A_3,A_5),(A_4,A_5),(A_1,A_2,A_4), (A_1,A_2,A_5),(A_1,A_4,A_5),(A_2,A_4,A_5),(A_1,A_2,A_4,A_5)}

Local frequent item sets at $Site_2$

LF_2= { A_1 ,A_2 ,A_3 ,A_4 ,A_5,(A_1,A_2),(A_1,A_4), (A_1,A_5), (A_2,A_3), (A_2,A_4), (A_2,A_5), (A_3,A_4), (A_3,A_5),(A_4,A_5),(A_1,A_2,A_4),(A_1,A_2,A_5),(A_2,A_3,A_4) (A2,A_4,A_5),(A_3,A_4,A_5)}

Local frequent item sets at $Site_3$

LF_3= {A_1 ,A_3 ,A_4 ,A_5,(A_1,A_3),(A_1,A_4), (A_1,A_5), (A_3,A_4), (A_3,A_5),(A_4,A_5),(A_1,A_3,A_4,A_5) }

After receiving the encrypted form of local frequent item sets from the sites, TP prepares a merged frequent item list after eliminating duplicates. The merged list is as follows.

{A_1,A_2,A_3,A_4,A_5,(A_1,A_2),(A_1,A_3),(A_1,A_4),(A_1,A_5), (A_2,A_3),(A_2,A_4),(A_2,A_5),(A_3,A_4),(A_3,A_5),(A_4,A_5), (A_1,A_3,A_4),(A_1,A_3,A_5),(A_1,A_4,A_5),(A_1,A_2,A_4), (A_1,A_2,A_5),(A_2,A_3,A_4)(A_2,A_4,A_5),(A_3,A_4,A_5), (A_1,A_2,A_4,A_5),(A_1,A_3,A_4,A_5)}

The following are the random numbers and signs sent by TP along with merged list to the three sites.
Site$_1$ received $RN_1 = 20$, $Sign_1 = ('+')$.
Site$_2$ received $RN_2 = 39$, $Sign_2 = ('-')$.
Site$_3$ received $RN_3 = 41$, $Sign_3 = ('-')$.

Each site computes partial support and broadcast to all other sites in order to find the total partial supports. All three sites broadcast total partial supports for all the item sets in the merged list. TP finally declares global frequent item sets by comparing global excess support (GES) of an item set with zero where GES$_i$ is computed by subtracting SignSumRN from TotalPS$_i$.

The following steps illustrate the process of finding whether the two item sets in the merged list are globally frequent or not. Consider the two item sets $\{(A_3, A_5), (A_3, A_4, A_5)\}$ from the merged list.
Let $X_{1=}$ (A_3, A_5) and $X_2 = (A_3, A_4, A_5)$
From the tables 2.1, 2.2 & 2.3, length of databases at three sites are given below

$| DB_1 | = 5, | DB_2 | = 5, | DB_3 | = 5$ Global database size is $| DB | = \sum_{i=1}^{3} | DB_i | = 15$

TP computes SignSumRN by adding three random numbers along with signs using the formula
SignSumRN = (+) 20 + (-) 39 + (-) 41 = - 60
Partial supports for X_1 at different sites are computed as follows.

At Site$_1$
$PS_{11} = X_1.Sup - 40\%$ of $DB_1 + (Sign_1)\ RN_1$
$PS_{11} = 2 - 2 + 20 = 20$
At Site$_2$
$PS_{21} = X_1 .sup - 40\%$ of $DB_2 + (Sign_2)\ RN_2$
$PS_{21} = 2 - 2 - 39 = - 39$
At Site$_3$
$PS_{31} = X_1.sup - 40\%$ of $DB_3 + (Sign_3)\ RN_3$
$PS_{31} = 3 - 2 - 41 = - 40$
Site$_1$ broadcasts 20 to Site$_2$ and Site$_3$, Site$_2$ broadcasts -39 to Site$_1$ and Site$_3$, and Site$_3$ broadcasts -40 to Site$_1$ and Site$_{2.}$ TotalPS$_{ij}$ are computed at all sites.
$TotalPS_{11} = PS_{11} + PS_{21} + PS_{31} = 20 + (- 39 - 40)$
$= -59$
$TotalPS_{21} = PS_{21} + (PS_{11} + PS_{31}) = - 39 + (20 - 40)$
$= -59$
$TotalPS_{31} = PS_{31} + (PS_{11} + PS_{21}) = -40 + (20 - 39)$
$= -59$
TP receives -59 as total support of an item set X_1 from three sites which ensures the computations performed by all sites is correct. TP then calculates Global Excess Support (GES$_1$) by subtracting SignSumRN from TotalPS$_{11.}$
$GES_{1=} TotalPS_{11}$ - SignSumRN
$= -59 - (-60) = 1$
The value of GES$_1$ is 1 which is greater than or equal to 0, so (A_3,A_5) is declared as globally frequent by TP and actual support(AS$_1$) of X_1 is computed by adding minimum support of the total database to GES$_1$.
$AS_1 = GES_1 + MinSup * |DB| = 1 + 6 = 7$ where $|DB| = 15$.

Hence, the global frequent item set (A_3,A_5) support is 7.
Let us find whether the item set X_2 is globally frequent or not.
Partial support for X_2 at three sites are computed as follows.
At Site$_1$
$PS_{12} = X_2.Sup - 40\%$ of $DB_1 + (Sign_1)\ RN_1$
$= 1 - 2 + 20 = 19$
At Site$_2$
$PS_{22} = X_2 .sup - 40\%$ of $DB_2 + (Sign_2)\ RN_2$
$= 2 - 2 - 39 = -39$
At Site$_3$
$PS_{32} = X_2.sup - 40\%$ of $DB_3 + (Sign_{3)}\ RN_3$
$= 2 - 2 - 41 = -41$
Site$_1$ broadcasts 19 to Site$_2$ and Site$_3$, Site$_2$ broadcasts -39 to Site$_1$ and Site$_3$, and Site$_3$ broadcasts -41 to Site$_1$ and Site$_{2.}$ TotalPS$_{i2}$ are computed at all sites and as follows
$TotalPS_{12} = PS_{12} + PS_{22} + PS_{32} = 19 + (- 39 - 41)$
$= - 61$
$\therefore TotalPS_{12} = TotalPS_{22} = TotalPS_{32} = - 61$
Each site sends its computed TotalPS$_{i2}$ (total support of X_2) to TP. TP then finds GES$_{2.}$
$GES_2 = TotalPS_{12}$ - SignSumRN
$= 59 - (-60)$
$= -1$
The value of GES$_2$ is -1 which is lower than zero, so (A_3, A_4, A_5) is declared as globally infrequent by TP even though it is frequent at Site$_2$ and Site$_{3.}$
The above procedure is repeated for all the item sets in the merged list to find whether they are globally frequent or not. Finally TP prepares a list which consists of global frequent item sets and their support values, TP then broadcast this list to three sites. This information is given in the following table.

TABLE III GLOBAL FREQUENT ITEM SETS AND SUPPORTS

Item Set	Sup	Item Set	Sup	Item Set	Sup
A_1	11	(A_1,A_2)	6	(A_4,A_5)	9
A_2	8	(A_1,A_4)	9	(A_3,A_4)	7
A_3	9	(A_3,A_5)	7	(A_1,A_4,A_5)	6
A_4	12	(A_1,A_5)	8		
A_5	12	(A_2,A_5)	7		

Even though the merged list consists of 25 item sets only 13 item sets are globally frequent.

Each site can generate global association rules for each global frequent item set based on the specified minimum confidence threshold. The following computations illustrates that how a rule can be declared as strong or weak rule based on the user specified minimum confidence threshold value (65%).
For the item set (A_1, A_4, A_5), the various rules that can be generated are $\{A_1 \rightarrow (A_4, A_5),\ A_4 \rightarrow (A_1, A_5),\ A_5 \rightarrow (A_1, A_4),\ (A_1, A_4) \rightarrow A_{5,}\ (A_1, A_5) \rightarrow A_{4,}\ (A_4, A_5) \rightarrow A_1\}$.
All these rules need not be strong rules. A rule can be declared as strong only when the confidence of the rule is greater than minimum confidence threshold value.
For the rule $A_1 \rightarrow (A_4, A_5)$
Confidence of this rule is
$Sup (A_1, A_4, A_5) / Sup(A_1)$
$= 6/11 = 54\%$

The rule, $A_1 \rightarrow (A_4,A_5)$ is a weak rule since rule's confidence is lower than minimum confidence value of 65%.

For the rule $(A_1, A_4) \rightarrow A_5$

Confidence of this rule is

$$\text{Sup } (A_1, A_4, A_5) / \text{Sup}(A_1, A_4) = 6/9 = 66\%$$

Hence, $(A_1,A_4) \rightarrow A_5$ is a strong rule as its confidence is greater than minimum confidence.

For the rule $(A_4, A_5) \rightarrow A_1$

Confidence of this rule is

$$\text{Sup } (A_1, A_4, A_5) / \text{Sup}(A_4, A_5)$$

$$= 6/9 = 66\% \text{ M} \geq \text{MinConf}$$

Hence, $(A_4, A_5) \rightarrow A_1$ is a strong rule as its confidence is greater than minimum confidence

For the rule $(A_1, A_5) \rightarrow A_4$

Confidence of this rule is

$$\text{Sup}(A_1, A_4, A_5) / \text{sup}(A_1,A_5) = 6/8 = 75\ \%$$

The rule, $(A_1,A_5) \rightarrow A_4$ is a strong rule as its confidence is greater than minimum confidence.

V. PRIVACY PRESERVATION IN THE PROPOSED MODEL

A new model is proposed in this paper to find efficiently privacy preserving association rule mining in horizontally partitioned databases. The proposed model can be applied to any number of sites and for any number of transactions in the databases of the sites. Many tasks such as findings of locally frequent item sets, partial supports and total supports for each item set in the merged list are performed independently at different sites. Hence the computation time of the proposed model is less. The efficiency of the proposed method in terms of privacy and communication is discussed as follows.

- Privacy is ensured by using encryption and decryption techniques at the time of transferring the frequent item sets from different sites to trusted party. From this, trusted party can know only local frequent item sets of each site but he does not know the supports of any item and cannot predict any thing related to sites database.
- At the time of calculation of Partial Supports of an item set at each Site_i, MinSup * DB_i is subtracted and the value of sign * random number is added to the supports of the item at that site. So Partial Supports are in disguised form and broadcast to the sites securely. Each site is not having any idea about the sign, random number which are assigned by trusted party to other sites and the database size of other sites is also not known. So from the Partial Supports, no site can predict other sites data/information. In this way, partial supports of item sets can be broadcast to all other sites by preserving privacy of individual data. Hence, the sign based secure sum concept which is used in the computation of partial supports enhances the privacy.
- Trusted party receives total partial support of each item set from all sites in order to find the global frequent item sets. By having these total supports, trusted party cannot find sites data/information since the database size of any site and local supports of any item at any site is not known by trusted party. Although trusted party assigned random numbers, signs to all sites and total database size is known, he cannot predict any site's private data.
- Finally results that are global frequent item sets and their supports are broadcasted by trusted party to all sites. With these results, no site owner can predict local support of any global frequent item sets, as global frequent item sets may not be frequent in all sites and any site owner can not predict the contribution of other sites database which makes the item set globally frequent.

  In distributed environment, the cost of communication is measured in terms of the number of communications for data transfer among all the sites and trusted party which are involved in the process of finding global association rules.
- The efficiency of an algorithm is assessed in terms of the communication costs incurred during information exchange. The proposed model minimizes the number of data transfers by allowing the transfer of bulk of data at a time from one site to another site and trusted party to sites. For example each site sends local frequent item sets of their database in a single data transfer to trusted party and even the sites sends its partial support for each item to other sites in a single transfer instead of sending one item set's partial support in one transfer to other sites. Hence the proposed model needs less communications.
- Trusted party also broadcast all the global frequent item sets for all sites in a single transfer. Hence the proposed model is more economy in terms of communication cost as it utilizes bulk data transfers.

The above discussion clearly specifies that the proposed model is efficient for finding global association rules by satisfying privacy constraints.

6. CONCLUSION

The main threat in finding association rule mining in distributed database environment is privacy that is no site owner wish to provide database or local frequent item sets or support value to any one. However every owner wishes to access mined result by participating indirectly in the mining process by providing partial results in disguised form. The problem of preserving privacy in association rule mining when the database is distributed horizontally among n ($n > 2$) number of sites with a trusted party is considered. A model is proposed in this paper which adopts a sign based secure sum cryptography technique to find the global association rules without disclosing individual's private data/information. The trusted party initiates the process and prepares the merged list. All the sites computes the partial supports and total supports for all the item sets in the merged list using the sign based secure sum cryptography technique and based on these results finally trusted party finds global frequent item sets. The functionality of the proposed model is illustrated with an example. The performance of the proposed model in terms of privacy and communication is presented and it indicates that this model efficiently preserves the privacy of individual sites in the process of finding global frequent item sets and global association rules with minimum number of communications.

REFERENCES

[1] M tamer Ozsu Patrick Valduriez, *Principles of Distributed Database Systems* ,3rd Edition.

[2] *R Agarwal, T Imielinski and A Swamy*, Mining Association Rules between Sets of Items in Large Databases, Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, page 207-210, 1993.

[3] *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.

[4] *Y. Lindell and B. Pinkas,* Secure Multiparty Computation for Privacy-Preserving Data Mining, The Journal of Privacy and Confidentiality (2009) , 1, Number 1, pp. 59-98.

[5] *Elisa Bertino , Igor Nai Fovino Loredana Parasiliti Provenza* ,A Framework for Evaluating Privacy Preserving Data Mining Algorithms, Data Mining and Knowledge Discovery, 2005, 11, 121–154.

[6] *M. Kantarcioglu and C. Clifto.* Privacy-preserving distributed mining of association rules on horizontally partitioned data. In IEEETransactions on Knowledge and Data Engineering Journal, volume 16(9), pages 1026–1037.

[7] *Chin-Chen Chang, Jieh-Shan Yeh, and Yu-Chiang Li,* Privacy-Preserving Mining of Association Rules on DistributedDatabases, IJCSNS International Journal of Computer Science and Network Security, VOL.6 No.11, November 2006.

[8] *Mahmoud Hussein, Ashraf El-Sisi,and Nabil Ismail,* Fast Cryptographic Privacy Preserving Association Rules Mining on Distributed Homogenous Data Base, I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 607–616, 2008.© Springer-Verlag Berlin Heidelberg 2008.

[9] Lalanthika Vasudevan , S.E. Deepa Sukanya, N. Aarthi ,Privacy Preserving Data Mining Using Cryptographic Role Based Access Control Approach, Proceedings of the International MultiConference of Engineers and Computer Scientists 2008 Vol I, IMECS 2008.

[10] Vaidya, J. and Clifton, C. 2002. Privacy preserving association rule mining in vertically partitioned data, 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, pp. 639–644.

[11] A.C. Yao. Protocols for secure computations. In Proceedings of the 23rd Annual IEEE Symposium on Foundations of Computer Science, 1982

.