2014 Annual IEEE India Conference (INDICON)

Interestingness Measure on Privacy Preserved Data with Horizontal Partitioning

S KumaraSwamy[1],Manjula S H[1], K R Venugopal[1], L M Patnaik[2]
Department of Computer Science and Engineering
[1]University Visvesvaraya College of Engineering, Bangalore University, Bangalore 560 001
[2]Honorary Professor, Indian Institute of Science, Bangalore, India
email:kumar.aruna@gmail.com

***Abstract*—Association rule mining is a process of finding the frequent item sets based on the interestingness measure. The major challenge exists when performing the association of the data where privacy preservation is emphasized. The actual transaction data provides the evident to calculate the parameters for defining the association rules. In this paper, a solution is proposed to find one such parameter i.e. support count for item sets on the non transparent data, in other words the transaction data is not disclosed. The privacy preservation is ensured by transferring the x-anonymous records for every transaction record. All the anonymous set of actual transaction record perceives high generalized values. The clients process the anonymous set of every transaction record to arrive at high abstract values and these generalized values are used for support calculation. More the number of anonymous records, more the privacy of data is amplified. In experimental results it is shown that privacy is ensured with more number of formatted transactions.**

Keywords—Association, Support count, formatted transactions, Generalization and Privacy factor.

I. INTRODUCTION

Association learning is a process of identifying the frequent item sets from the large database. The item sets are said to be frequent based on the bondage existing between them. The bondage indicates the interestingness factor of these item sets. For example, in market basket analysis how frequently people take milk if they take bread and butter together. This kind of analysis enables the shops to make decisions on the sales and take strategic decisions on the product promotion. The association between the item sets are based on the transaction database which involves the calculation of support count and confidence between the item sets. The *support count* for an item set is the number of transactions specifying the item set purchased together. And hence it gives the probability of item sets bought together in specific transactions. The probability values helps to analyze the interestingness measure for the item sets. The other parameters used to form association rules are *confidence, lift and conviction*. The confidence is the accuracy of the interestingness measure based on the support count and hence it depends on support count. *Lift* is a ratio of observed support to that of expected if item sets are independent to each other. The major challenge exists when the transaction data source is distributed and privacy preservation is emphasized. Here, the solution is proposed in order to predict the interestingness measure between the item sets on the horizontally partitioned datasets. The privacy of data is ensured using X-number of anonymous records for an individual record. The anonymous records help generalizing the data from the local data provider perspective and preserve the privacy. These most generalized set of data requires appropriate prediction to arrive at specific values for support count for every item set. After identifying the support count on each side the stochastic gradient is applied to tune the data to get accurate support count. The confidence between the required item sets is defined; both support and confidence are used to define associate rules. This paper concentrates only on calculation of support count on privacy preserved data with privacy amplification approach. The a number of anonymous transactions are incremented polynomially and the relative generalization is ensured as the derivative of polynomial function.

A. Motivation

Finding the interestingness factors(support/confidence) between the item sets enables to decide the frequency of those item sets appearing together. This is helpful for market basket analysis, fraud detection etc. The analysis is made with lesser difficulty If the transaction data centralized in specific location where as if the transaction data are dispersed or distributed and privacy preservation is required. The existing generalization and suppression technique ensures the privacy of data even if the data is exchanged. The privacy data can be highly abstracted using generalization or some information are not disclosed/suppressed. The privacy preservation techniques and association rules provides motivation to come up with the combined approach to perform the prediction of the interestingness factor for determining the association rules between the item sets.

B. Contribution

In this work a new approach is proposed for efficient prediction of interestingness measure between the frequent item sets. The generalized data is fetched from a set of anonymous records of every transaction record.

978-1-4799-5364-6/14/$31.00 ©2014 IEEE

C. Organiztion

Here, the contents organization of paper is briefed. Section II describes the Related work and Section III defines the problem. Section IV defines mathematical model. Section V describes the algorithms proposed for solution. Section VI details the system architecture. Section VII demonstrates the experimental results and data sets used for the process. The paper concludes by mentioning the enhancement that can be incorporated in prediction process along with the suppression and the list of references considered by the authors.

II. RELATED WORK

Ming-Jun et al., [1] studies privacy preserving decision tree classification algorithm to solve a distributed computation problem that the participant parties jointly build a decision tree over the data set distributed among them, and they do not want their private sensitive data to be revealed to others during the tree-building process. The paper proposes a solution to privacy preserving C4.5 algorithm based on secure multi-party computation techniques, which can securely build a decision tree over the horizontally partitioned data with both discrete and continuous attribute values using a secure two-party bubble sort algorithm to solve the privacy preserving. Our proposal is to share X:1 transaction records to central node to generalize the data, preserve the privacy and interestingness study among the transaction record attributes.

Agrawal et al.,[2] proposes to build classifiers whose accuracy is comparable to the accuracy of classifiers built with the original data considering that the resulting data records look very different from the original records and the distribution of data values is also very different from the original distribution. It is not possible to accurately estimate original values in individual data records, a proposal of reconstruction procedure to accurately estimate the distribution of original data values.
Here our approach is based on the getting max value for each attribute of anonymous transaction records and then the resulting formatted transaction record is considered for association analysis.

Shuguo Han et al., [3] have proposed a preliminary formulation of gradient descent with data privacy preservation for solving many optimization problems with two approaches. Namely, stochastic approach and least square approach with different assumptions. Authors are proposed four protocols for the two approaches incorporating various secure building blocks for horizontally and vertically partitioned data. Authors conducted experiments to evaluate the scalability for secure building blocks and the accuracy and efficiency of the protocols with different scenarios.

Gabriel Ghinita et al., [4] have proposed a method to solve the problem of anonymizing sparse and high-dimensional transactional data is solved by local NN-search and global data reorganization. LSH-based anonymization method outperforms the RCM method in terms of data utility and incurs slightly higher computational overhead, but Gray code sorting is best compared to all other methods with respect to both data utility and anonymization overhead. The requirement for proposed anonymizing transactional data is recently emphasized with the release of the "Netflix Prize" data which contains movie ratings of 500,000 subscribers. In [5] an attacker can reidentify 80 percent of the subscribers depending on knowledge about six reviewed movies. This proposed method fails to address the problem of anonymization of high-dimensional data for non binary databases.

Mathew G and Obradovic Z [6] proposed a framework for distributed knowledge-mining, which is helpful for clinical decision support tool in decision tree form. This proposed framework helps for knowledge building using statistics based on patient data from multiple sites which satisfies a certain filtering condition, without the need of actual data to leave the participating sites. The proposed information retrieval and diagnostics supporting tool accommodates both heterogeneous data schema associated with participating sites and it also supports prevention of personally identifiable information leakage and preservation. These are important security concerns in managing of clinical data transactions.

Zhu Yu-quan et al.,[7] have proposed a method which is effective to find frequent item sets on vertically distributed data to resolve the problems on existing protocol of secure two-party vector dot product computation. This proposed method uses semi-honest third party to participate in the calculation for the converted data of the parties to a third party to calculate. Advantage of this method is, it results better algorithm efficiency and accuracy compared to the original Vector dot product algorithm.

III. PROBLEM DEFINITION

If the data is distributed using horizontal partitioning among data custodians then there is a need for analyzing the interestingness calculation for better business vision and strategy formulation. Now the major concern is about data privacy and sensitivity compromise while analyzing the data to a central location by all the communicating parties' agreement.
Assumptions: All the communicating parties are aware of the data partition scheme and abide by association calculation model.

IV. MATHEMATICAL MODEL

Definitions

Transactions: The successful transported data records from individual data provider, a communicating party, to the centralized server is termed as transaction. All these transactions serve the basis for associatively analysis.

List of notations used

Table 1: lists the notations used while explanation

Notations	Meaning
T	Set of all the transactions
T_r	A transaction record
T_{rf}	A formatted transaction record
X	No of anonymous transaction records
P_f	Privacy factor, a set of X anonymous transaction records for a generalized T_{rf} calculation
A_r	The set of attributes; generally its subset is used in a transaction record
$S_{A_r' \rightarrow A_r''}$	The support count between attribute sets A_r' and A_r'' for the formatted transaction records
$C_{A_r' \rightarrow A_r''}$	The confidence function for attribute sets A_r' and A_r'' calculated by the central node
$A_f(T_i)$	Function to fetch the attribute set of i^{th} transaction

Transaction Record: A transaction record is information about individual transaction at the end of a data provider is defined as following:

$$T_r = \{ E/E=\{v_1, v_2, \cdots, v_n\},\ n \leq |A_r| \ \wedge\ (\forall i : 1 < i \leq n \, v_i \in R^+) \} \quad (1)$$

Privacy Factor: A set of X anonymous transaction records from T which are used to calculated a generalized formatted transaction record, T_{rf} , can be defined as follows:

$$P_f = \{ (T_r)_i \ / \ 1 < i \leq X \} \quad (2)$$

Formatted Transaction Record: This is the formatted transaction record by formatting unit at the client end i.e. data provider end, is transported to the centralized server for the association calculation among transaction attributes.

$$T_{rf} = \{ max((T_r)_i \cdot V_j) \ / \ T_r \in P_f \wedge 1 < i \leq |P_f| \wedge 1 < j \leq |A_r| \} \quad (3)$$

Each formatted transaction record has all the attributes of A_r . If there is any attribute missing in the original transaction record, T_r , is represented as value 0 in the transformed data record. Each attribute value in the formatted transaction record is the max of the X anonymous records attribute value and if a particular value is missing it is considered as 0.

Interestingness calculation among transactions: The interestingness is measured using association rules, support count and confidence functions, over all the transactions. The support count and confidence is determined for attribute sets A_r' and A_r'' among transactions, which can be defined as following:

$$\forall T_i \in T : \left(\left(A_f(T_i) \cap A_r' = A_r' \right) \wedge \left(A_f(T_i) \cap A_r' = A_r'' \right) \right) \Rightarrow A_r' \rightarrow A_r'' \quad (3)$$

This propositional function for a transaction, T_i , can be demonstrated using a truth table as:

A_r'	A_r''	$A_r' \rightarrow A_r''$
T	T	T
T	F	F
F	T	F
F	F	F

(4)

The set of transactions having association between attribute sets A_r' and A_r'' can be defined as following:

$$T_{A_r' \rightarrow A_r''} = \{ t / (t \in T) \wedge \left(A_f(t) \cap A_r' = A_r' \right) \wedge \left(A_f(t) \cap A_r'' = A_r'' \right) \} \quad (5)$$

The support count between attribute sets A_r' and A_r'' among all the transactions can be defined as following:

$$S_{A_r' \rightarrow A_r''} = \frac{\left(T_{A_r' \rightarrow A_r''} \right)}{(T)} \quad (6)$$

Before defining the confidence function between attribute sets A_r' and A_r'' among all the transactions; we need to define a set for transactions having attributes A_r' and its support function. The set definition follows as:

$$T_{A_r'} = \{ t / (t \in T) \wedge \left(A_f(t) \cap A_r' = A_r' \right) \} \quad (7)$$

The support count function for A_r' attribute definition follows as:

$$S_{A_r'} = \frac{\left(T_{A_r'}\right)}{(T)} \quad (8)$$

The confidence function for attribute sets $A_{r'}$ and $A_{r''}$ can be defined as following:

$$C_{A_r' \rightarrow A_r''} = \frac{\left(S_{A_r' \rightarrow A_r''}\right)}{\left(S_{A_r'}\right)} \quad (9)$$

Privacy Amplification: In order to strengthen the privacy preservation, the transactions are formatted in terms of power of anonymous transactions. The privacy preservation is improved with increase in the power factor and hence the power amplification is a function of anonymous transactions and hence defined as,

$$g(x) = x^{\eta} \quad (11)$$

Where x is the number of anonymous transactions that every node decides to generate and this number is amplified at every node before transforming to central node. In every consecutive amplification more generalization data is produced and hence more privacy is emphasized. The relative privacy amplification between every 'x'(number of anonymous transaction record) is a derivative of g(x) and hence it is defined as,

$$\delta g(x) = \eta x^{(\eta-1)} \quad (12)$$

V. ALGORITHMS

This complete process consists of three steps, formatting of data, transportation of formatted data to centralized node for association analysis and report of analyzed data to the interested parties. The algorithm for associatability calculation is detailed below in Table 2:

These are the following state events used in the protocol:

[1] **CON_CHK_PING:** The local data custodian initiates the communication sending a message segment 'PING' to the central node.

[2] **CON_PING_ACK:** Central node acknowledges the connection initiation request by sending a message 'SEND ME DATA ATTRIBUTE|FEATURE' to the data provider. This marks the communication channel establishment between the two nodes.

[3] **TXF_RECORD_ATT_REQ:** Data provider sends a string message, formatted attributes list as delimited string, to the central node. This message formatting is mutually understood by both the communicating parties.

Table 2 Algorithm for formatting a transaction record

Input:

T_{rf} ; The formatted transaction record at the data provider end.

Output:

$S_{A_r' \rightarrow A_r''}$, $C_{A_r' \rightarrow A_r''}$

Process:

1. Local data provider initiates communication using CON_CHK_PING message to the central node; Central node acknowledges as CON_PING_ACK.

2. Data provider node sends the formatted transaction record attributes using TXF_RECORD_ATT_REQ message, a string message, to the central node and central node responds with TXF_RECORD_ATT_RES message to start sending the data now.

3. Data provider node starts selecting P_f , X anonymous no of transactions from transaction records and transforms it to a formatted message, T_{rf} , which are transported to the central node as TXF_RECORD unless all are sent. Central node keeps acknowledging the transaction record formatting with TXF_RECORD_ACK message back to the data provider node.

4. Once the client node sends all the formatted transactions to the central node; it sends the CON_TERM_REQ to the central node to aware of data being transmitted.

Central acknowledges this request by CON_TERM_RES and triggers 'Data Reporter' component of the associative analyzer.

5. Central node calculates the associability among the received formatted transactions after reformatting the data. The mathematical model explained above is used and it produces the output for $S_{A_r' \rightarrow A_r''}$ and $C_{A_r' \rightarrow A_r''}$.

[4] **TXF_RECORD_ATT_RES:** Central node reformats the string and if it is in compliance with the transaction record exchange format, it sends the message segment 'SEND ME DATA OR BYE TO DISCONNECT' to the private data node.

[5] **TXF_RECORD:** Data provider sends a formatted transaction record to the central node.
[6] **TXF_RECORD_ACK:** Central node reformats the received transaction record and if it complies with the transaction record formatting protocol; central node acknowledges in this segment.
[7] **CON_TERM_REQ:** Data provider node requests for termination of the established communication channel to the central node.
[8] **CON_TERM_RESP**: Central node provides the acknowledges the request by terminating the established channel.

VI. SYSTEM ARCHITECTURE

As shown in fig 1,

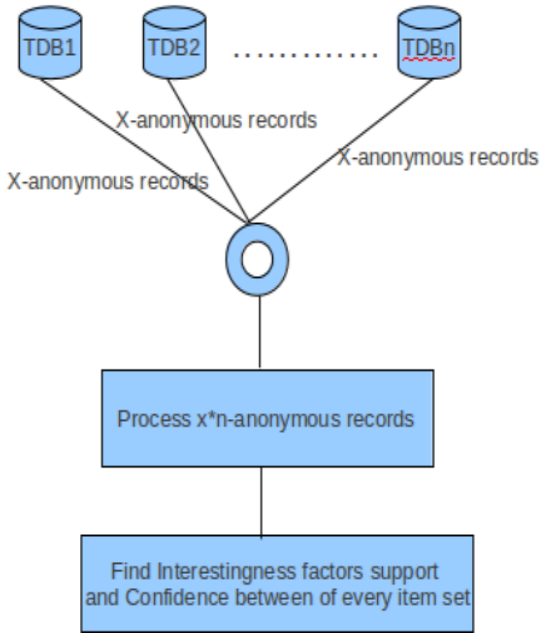

Fig. 1. Privacy prservation glimpse

1. Every communicating party has their own set of transaction records forming transaction database namely TDB_1, TDB_2,...,TDB_n for communicating parties CP_1, CP_2,.....,. CP_n .
2. The transaction databases are updated frequently for every transaction performed at the market place. Every communicating party chooses X-anonymous transaction records.
3. The anonymous transaction records are used to create a formatted transaction record extracting upper bound value for each attribute out of the anonymous transaction records.
4. Every communicating party knows the item sets between whom the interestingness measures needs to be determined. Each one of them calculates the support count and confidence for the item sets to define association rule.
5. The support count and confidence is determined between every pair of item sets. The supersets of all item sets are known in prior to all communicating parties.
6. The support count indicates the minimum number of transactions or percentage of transactions supports the presence of the item sets together.

As shown in Figure 2, the local data custodians are depicted as Nodes. The data records are horizontally partitioned i.e. each record is disjoint and has same set of attributes. To derive the associability from all these local data, transaction records, 'Node Agent' computation model is proposed. This agent behaves as client for the data providers and server for the centralized node, where the association analysis is performed. This model consists of three components which are detailed further.

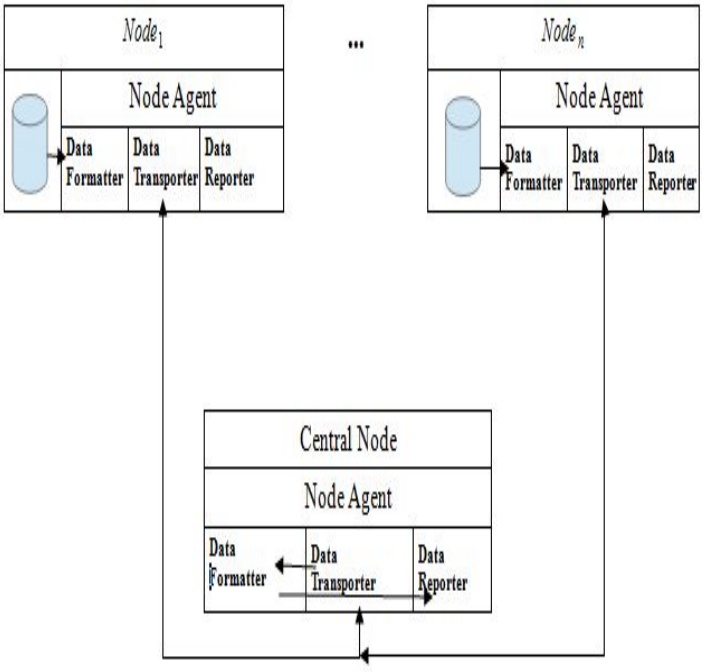

Fig. 2. Privacy Preservation Architecture

Data Formatter: This component is responsible for privacy preservation of data for the local data providers. Each data custodian stores transaction records related to its business or organization and strives for its privacy and unethical access. The generalized privacy is attained using the anonymous count, X, which is used to create a formatted transaction record for such X records and the probability of privacy is 1/X. The max value for each attribute is chosen from P_f set; it results in a formatted transaction record, T_{rf} .

Data Transporter: This module is responsible for communication between the local data provider and central node. It follows a state based protocol as explained above in the section III for reliable and secure data transfer between two parties. The scope of this module is over once the local data provider node sends the connection termination request and the central node acknowledges it.

Data Reporter: This module is the core of associability calculation and confined to the central node. The accuracy of the associability depends upon the probability of privacy. This module is triggered once the formatted transaction records from all the data providers are cached on central node. It basically consists of two parts, support count calculation and confidence calculation. The confidence calculation depends upon the support count results.

VII. Experimental Results

This experiment demonstrates the relationships among the components involved in associability computing; Data Formatter, Data Transporter and Data Reporter. The experimental setup is prepared in line with the proposed system in this paper. The software system is designed using client-server technology. The data formatting and transporting components of the local data provider system is simulated using the client module for this experiment and the data reporter, the module responsible for data analysis and centrally located, is simulated as the server module. The software system is designed in such a way that it can act as client and server module simultaneously; this mode of deployment is not advisable in case of multiparty data providers. It should be practiced for in house data analysis, business insight or experimentation. The complete implementation is done with java, Eclipse IDE in Linux platform.

The client module chooses anonymous count, X, as privacy factor to create the privacy factor sets. The privacy factor count lets derive a set of formatted transaction records out of the chosen set of transaction records to preserve the data privacy prior transporting from client module to server module for associability calculation. The probability of privacy is inversely proportional to privacy factor.

In the experiment here, privacy factor count, X, is chosen 2 and 4 respectively to create the privacy factor sets. For X=2, there are 6 sets of transaction records of size 20, 100, 200, 1000, 2000 and 10000 are chosen; these create 6 sets of formatted transaction records of size 10,50,100,500,1000 and 5000 respectively. For X=4, the privacy factor sets of size 40, 200, 400, 2000, 4000 and 20000 are chosen to create sets of formatted transaction records of size 10,50,100,500,1000 and5000 respectively.

The privacy factor count, X, yields probability of privacy for values 2 and 4; 1/2 =0.5 and 1/4 = 0.25 respectively. Here probability of privacy, 0.25, is high in preserving privacy of data. The graph depicted below in Fig. 3 demonstrates the relation between transaction records set, formatted transaction records set and the privacy factor, P_f .

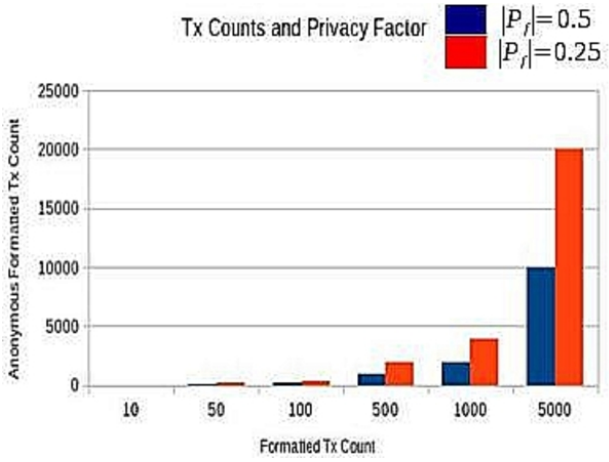

Fig. 3. Plot of formatted vs. raw transaction (Unformatted)

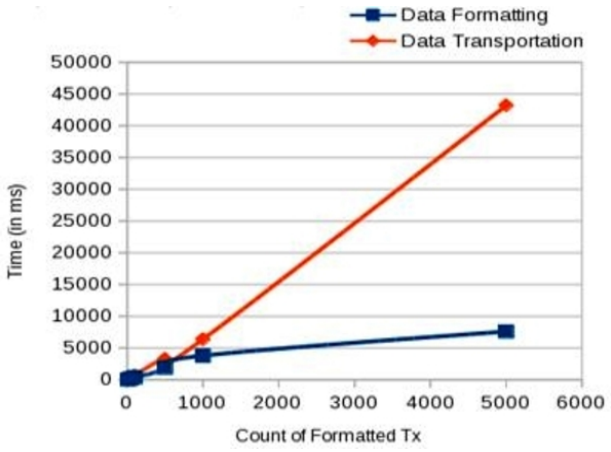

Fig. 4. Formatting of transaction at Client side

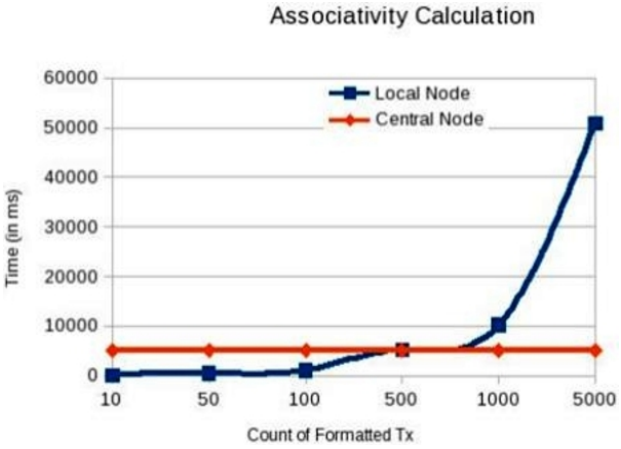

Fig. 5. Formatting of transaction at Server side

In the experiment here, there is a delay of 10 ms used as a worst case analysis for network bandwidth delay for consecutive message transportation. This is counted as part of

the transportation time from client module to server module. The graph presents the data for client module processing. From the Fig. 4 it is understood that data transportation time is always more than data formatting. Bound to improve for a commercial or enterprise hardware configuration. The client module processing, formatting and transportation, time is directly proportional to the size of formatted transaction records.

The graph presents the data for client module processing. From the Fig. 5 it is understood that data transportation time is as the formatted transaction record size is increasing, the processing time also increase significantly. In our experiment, the value is ranging for the above mentioned data set from 108 ms - 51 seconds. This processing time are also expected to improve for enterprise hardware systems.

VIII. CONCLUSION

Interestingness measure in transactions is a common task in database for prediction or any sort of analysis. This paper has proposed the solution for finding all sort of associability in transactions.

We have proposed the computation model which has two important components, formatting component and transportation component. The formatting component ensures the data privacy interest of the data provider and follows a very secure and reliable mechanism of data transferring to the central node. There is a significant tradeoff observed while transporting the data to the central node; the transportation is consuming significant time from client to central node proportional to the records count while the association calculation at the central node is almost constant in all the cases. Data reporter unit reports the association rules existing in all the transactions. The work can be enhanced to optimize the support count values by using first order optimization algorithms. This enables to define accurate parameters for association rule mining.

REFERENCES

[1] Ming-Jun Xiao, Kai Han, Liu-Sheng Huang and Jing-Yuan Li, "Privacy Preserving C4.5 Algorithm Over Horizontally Partitioned Data", Proceedings of the Fifth IEEE International Conference on Grid and Cooperative Computing, pp. 78-85, 2006.

[2] R. Agrawal and R. Srikant, "Privacy preserving data mining" In Proceedings of the ACM SIGMOD Conference on Management of Data, pp. 439-450,2000.

[3] Shuguo Han, Wee Keong Ng, Li Wan , Vincent C S Lee, " Privacy-Preserving Gradient-Descent Methods ", IEEE Transactions on Knowledge and Data Engineering, vol.22, no.6, pp 884-899,June 2010.

[4] Gabriel Ghinita, Panos Kalnis, and Yufei Tao, "Anonymous Publication of Sensitive Transactional Data", IEEE Transactions on knowledge and data engineering, vol. 23, no. 2, February 2011.

[5] A. Narayanan and V. Shmatikov, "How to Break Anonymity of the Netflix Prize Dataset," http://arxiv.org/abs/cs/0610105, 2010.

[6] Mathew G, Obradovic Z," A privacy-preserving framework for distributed clinical decision support ", Proceedings of IEEE 1st International Conference on Computational Advances in Bio and Medical Sciences (ICCABS), pp.129-134, Feb. 2011.

[7] Zhu Yu-quan ,Tang Yang, Chen Geng,"A Privacy Preserving Algorithm for Mining Distributed Association Rules", proceedings of International Conference on Computer and Management (CAMAN), pp.1-4, May 2011.