IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555
Vol. 2, No.6, December 2012

Techniques for Privacy Preserving Association Rule Mining in Distributed Database

Jayanti Dansana[1], Raghvendra Kumar[1] and Jyotirmayee Rautaray[1]
[1]School of Computer Engineering, KIIT University, ODISHA, INDIA
jayantifcs@kiit.ac.in , raghvendraagrawal7@gmail.com, jyotirmayee.1990@gmail.com

Abstract— Data mining technology has emerged as a means of identifying patterns and trends from large quantities of data. Mining association rule is important data mining problem. Association rule mining algorithms are used to discover important knowledge from databases. Privacy concept occurs when the data is distributed in environment and association rule also depend on the distribution of data in environment. There are mainly three type of distribution occurs in database like vertical partition or horizontal partition or hybrid partition among different site. So that it increases the demand of finding the global Association result that may be frequent or infrequent it's depending on the data in different sites for finding global result without disclosing information of their own data to other parties. So we need protocol for provide security in horizontal partitioning and vertical partitioning as well as hybrid partitioning of data. Here in this paper we gave a survey on various privacy preserving technique using association rule mining for vertical, horizontal and hybrid database.

Keywords*- Horizontal partitioned database;Distributed Association Rule Mining; Privacy preserving protocols.*

1 Introduction

The immediate evolution of computer technology in the last few decades has provided asset professionals with the capability to right to use and consider tremendous amounts of financial data. In addition, the World Wide Web, email, and bulletin boards make it possible for people around the sphere to access this information fast, as well as providing a means for people to voice their opinions and interact. As a result, some of the more intriguing topics of debate in recent years have revolved around the practice and consequences of data mining. Data mining involves searching through databases for correlation and pattern that differ from results that would be predictable to occur by chance or in arbitrary conditions. Data Mining is also used by advertisers and marketing firms to aim consumers. But possibly the most notorious group of data miners are stock market researchers that seek to guess outlook stock price movement. Most if not all Stock Market Anomalies have been at least documented via data mining of past prices and sometimes dissimilar variables. Data mining techniques are available to recover useful information from large database. Guess and sketch are the two fundamental goal of data mining. To full fill these goals many data mining techniques exists such as association rules, classification, Clustering and so on. Among these, association rule has wide applications to find out interesting relationship among attributes in huge databases. Association rule mining is used to find the rules which satisfy the user particular minimum support and minimum confidence. In the process of result association rules, the set of frequent item sets are computed as the first step and then association rules are generated based on these frequent item sets. Two types of database environments are present namely centralized and distributed. In contrast to the centralized data base model, the distributed data base model assumes that the data base is partitioned into put out of joint fragments and each fragment is assigned to one site.

The issue of privacy arises when the data is distributed among multiple sites and no site landlord wish to provide their private data to other sites but they are interested to know the global results obtained from the mining process. Keeping in view of the inspiration to hole in privacy in data mining techniques to protect the confidential data of the user, there evolved a latest watercourse in data mining period that is privacy preserving in data mining. 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 .a shows a method for mixed partitioned in which data is first partitioned vertically and then horizontally. Shows another mixed method in which data is partitioned horizontally and then vertically partitioned. Vertically partitioned database is further partitioned into horizontal. Horizontally partitioned database is further partitioned into vertical. In data mining, association rule mining is a accepted and fine researched method for discovering interesting relations between variables in large databases. When data is distributed among unlike sites, finding the global association rules is a demanding task

as the privacy of the individual site's data is to be conserved. In this paper, a model is proposed to find overall association rules by preserving the privacy of individual sites data when the data is partitioned horizontally or vertically or hybrid among n number of sites.

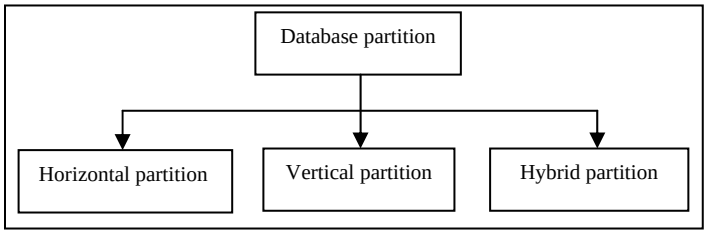


Fig. 1. Partitioning of database

In privacy preserving data mining is an important concept because when the data is distributed then its compulsory to provide security to that data so that another party will never know there particular data so that several approach come for provide security to data mining is describe below.

1. The data is misrepresented earlier than delivering it to the data miner.
2. The data is distributed between two or more sites, which assist using a semi-honest protocol to learn global data mining results without revealing any information about the data at their individual sites.
3. While using a model to catalog data, the classification results are only revealed to the designated party, who does not learn anything else other that the classification results, but can confirm for presence of certain rules lacking revealing the rules.

In this paper, an impressive level overview of some of the generally used tools and algorithms for privacy preserving data mining is presented. In this paper we try to find that probability of data leakage is zero and provides high security to database in every site. In section 2 we discuss association rule mining. In section 3 we give brief detail of data mining partitioning. In section 4 we give a concept of Secure multi party computation .In sections 5 we discuss about relevant data mining and security techniques.

2 Distributed Association Rule Mining

In this paper we describe the Privacy Preserving association rule [8] mining technique for a horizontally partitioned or vertical partitioned or mixed partitioned data set across multiple sites. Let I = {i_1, i_2, .i_n} be a set of items and T = {T_1, T_2,… T_n} be a set of transactions where each $T_i \subseteq I$. A transaction T_i contains an item set $X \subseteq I$ only if $X \subseteq T_i$. An association rule implication is of the form $X \Rightarrow Y (X \cap Y = 0)$ with support s and confidence C if S% of the transactions in T contains $X \cup Y$ and C% of transactions that contain X also contain Y. In a horizontally partitioned database, the transactions are distributed among n sites.

Support (X Y) = Provability(X Y) /|Total Number of Transaction|

The global support count of an item set is the sum of all local support counts.

$$\text{Support g (X)} = \sum_{i=1}^{n} Support_i(x).$$

Confidence of rule (X Y) = Sup (X Y) / Sup(X)
The global confidence [8] of a rule can be expressed in terms of the global support.

Confidence g $(X \Rightarrow Y)$ = Support g $(X \cup Y)$ /Support g(X)

The aim of the privacy preserving association rule mining is to find all rules with global support and global confidence higher than the user specified minimum support and confidence. The following steps, utilizing the secure sum and secure set union methods described earlier are used. The basis of the algorithm is the Apriori algorithm [3] [4] [5] [13] which use the (k-1) sized frequent item sets to generate the k sized frequent item sets. The problem of generating size 1 item sets can be easily done with secure computation on the multiple sites.

1. *Candidate Set Generation*: Overlap the globally frequent item set of size (k-1) with locally frequent (k-1) item set to get candidates. From these, use the apriori algorithm to get the candidate k item sets.
2. *Local Pruning:* For each X in the local candidate set, scan the local database to compute the support of X. If X is locally frequent, it's included in the locally frequent item set.
3. *Item set Exchange*: Calculate a Secure union of the large item sets over all sites.
4. *Support Count*: Compute a Secure Sum of the local supports to get the global support.

3 Horizontal Partitioning

Horizontal partitioning divides [3] a table into several tables. Every table then contains the same number of columns, but fewer rows. For example, a table that contains 1 billion rows could be partitioned horizontally into 12 tables, with each smaller table representing one month of data for a specific year. Some queries requiring data for a specific month only reference the suitable table. Determining how to partition the tables horizontally depends on how data is analyzed. We partition the tables so that queries reference as only some tables as possible. Otherwise, excessive UNION queries, used to merge the tables sensibly at query time, can affect performance. For more information about querying horizontally partitioned tables, see Scenarios for Using Views. For providing security in data mining for horizontal partition many Privacy Preserving protocol [7] [8] are used one by one is describe below in this paper.

4 Secure Multi Party Communication

Approximately all Privacy Preserving data mining techniques rely on Secure multi party communication protocol. Secure multi party communication is defined as a computation protocol at the last part of which no party involved knows anything else except its own inputs the outcome, i.e. the view of each party during the execution can be effectively simulated by the input and output of the party. In the late 1980s [1], work on Secure multi party communication verified that an extensive class of functions can be computed securely under reasonable assumptions without involving a trusted third party. Secure multi party communication has commonly concentrated on two models of security. The semi-honest model assumes that every party follows the rule of the protocol, but is free to later use [2] [12] what it sees during execution of the protocol. The malicious model assumes that parties can arbitrarily cheat and such cheating will not compromise moreover security or the outcome, i.e. the results from the malicious party will be correct or the malicious party will be detected. Most of the Privacy Preserving data mining techniques assume an intermediate model, Preserving Privacy with non-colluding parties. A malicious party May dishonest the results, but will not be able to learn the private data of other parties without colluding with another party. This is a practical hypothesis in most cases.

5 Secure Sum Protocol

Distributed data mining algorithms [5] [6] frequently evaluate the sum of values from individual sites. Pretentious three or more parties and no collusion, the subsequent method securely computes such a sum.

Let $v = \sum_{i=1}^{s} v_i$ is to be computed for s sites and v is known to lie in the range [0...N]. Site 1, designated as the master site generates a random number R and sends $(R+v_1) \bmod N$ to site 2. For every other site l = 1,2, 3, and 4 … s, the site receives: $V = (R + \sum_{i=1}^{l-1} v_j) \bmod N.$

.

Site l computes:

$$(V + v_l) \bmod N = (R + \sum_{i=1}^{l} v_j) \bmod N$$

This is passed to site (l+1). At the end, site 1 gets:

And knowing R, it can compute the sum V. The technique faces an obvious problem if sites collude. Site(l-1) and (l+1) can compare their inputs and outputs to determine v_l .The method can overall to work for an honest majority , Each site divides v_l into shares .The sum of each share is computed individually. The path used is permuted for each share such that no site has the identical neighbors twice.

5.1 Ck Secure Sum Protocol

The technique for Ck-Secure Sum Protocol [10] is that we adjust the neighbors in each round of segment computation. Accordingly it is positive that no two semi honest parties can know all the data segments of a wounded party. Inside this protocol all of the parties breaks the data block into K = N-1 segments where N is the number of parties involved in secure sum computation. We select P1 as the protocol designer. The position of the protocol initiator is kept fixed in all the rounds of computation. For the first round of the computation parties are arranged serially as P1, P2… Pn. The protocol initiator starts computation to get the sum of first segments of each party. For this computation our k-Secure Sum protocol is used. Now, P2 exchanges its location with P3 and second round of computation is performed. Now, P2 exchanges its location with P4 and so on. Figure1shows Ck Secure sum protocol. Formally, In i^{th} round of the computation P2 exchanges its location with Pi+1 until Pn is reached. In each round of computation, segments are added and the partial sum is passed to the next party until all the segments are added. Finally, the sum is announced by the protocol initiator party. The Ck-Secure Sum Protocol provides privacy against two colluding neighbors. Its analysis shows that when more than two parties collude, they can know the data of some party. The protocol initiator can be attacked by more than two parties that mainly want to know secret data of the protocol initiator. But for that also a specific combination of the parties must join against the protocol initiator. Any party who moves its position cannot be attacked by any collection of the parties.

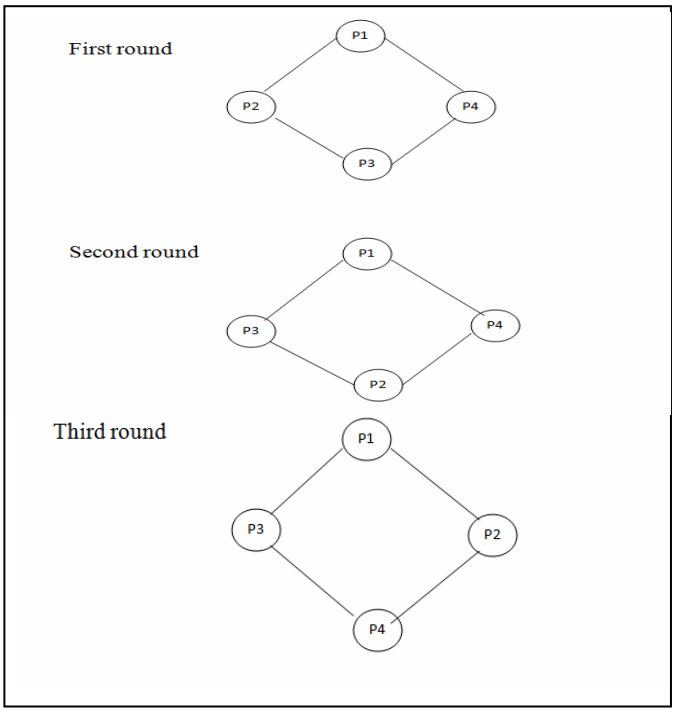


Fig. 2: Ck Secure Sum Protocol

5.2 Modified Ck-Secure Sum Protocol

Two modifications to the Ck-Secure Sum Protocol [10] are ended. Figure 2 shows the Modified Ck Secure sum protocol.

Step1. The number of segments k is kept equal to the number of the parties' N.
Step2. The protocol initiator party moves through the ring.

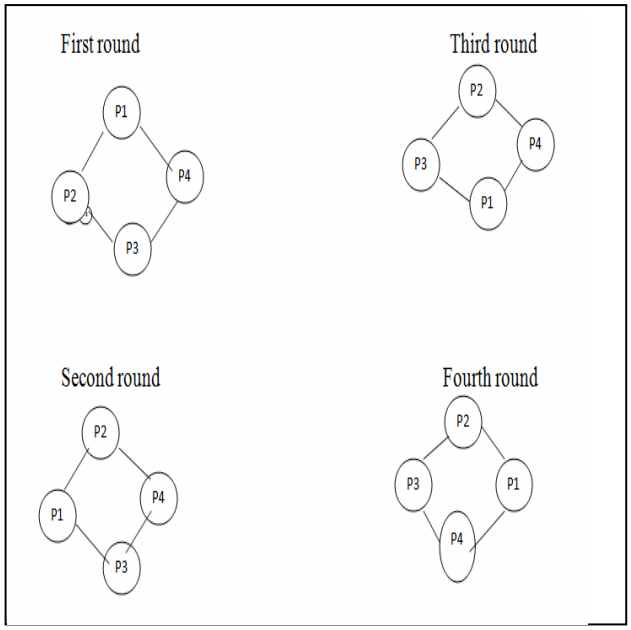


Fig. 3: Modified Ck Secure Sum Protocol

5.3 D-k Secure Sum Protocol

In this protocol each party divides its data into number of segment .There also exist distribution paths for distributing the data segments to other parties earlier than computation P1, P2, …, Pk are k parties involved in supportive Secure Sum computation where each party is skilled of infringement its data building block into a fixed number of segments such that the sum of all the segments is equal to the assessment of the data block of that party. Architecture of Dk Secure Sum Protocol before redistribution shown in figure 3. In proposed protocol number of segments in a data building block is kept equal to the number of parties. The values of the segments are arbitrarily selected by the party and it an undisclosed of the party. If k be the number of segments then in this scheme each party holds any one segment with it and k-1 segments are sent to k-1 parties, one to each of the parties. Thus at the end of this rearrangement each of the parties holds k segments in which only one segment belongs to the party and other segments belong to remaining parties, one from each. A snapshot of four party Secure Sum computations after distribution of segments is shown in figure 4. Now, Ck-Secure Sum Protocol can be functional to get the sum of all the segments. In this protocol, one of the parties is commonly selected as the protocol initiator party which starts the computation by distribution the data segment to the next party in the ring. The in receipt of party adds its data segment to the received partial sum and transmits its result to the next party in the ring. This process is recurring in anticipation of all the segments of all the parties are added and the sum is announced by the protocol initiator party. Dk-Secure Sum Protocol [10] for first round of segment computation is shown in the figure 5. Now even if two adjacent parties spitefully cooperate to know the data of a middle party they will be able to know only those k segments of a party which belong to every party. The sum of these segments is a refuse value and thus valueless for the hacker party.

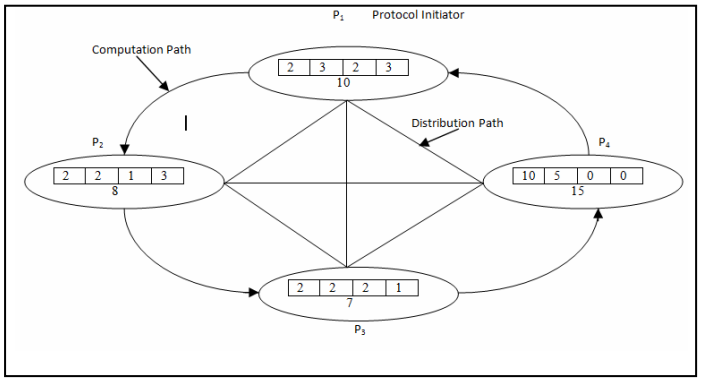


Fig. 4. Architecture of Dk-Secure Sum Protocol before redistribution

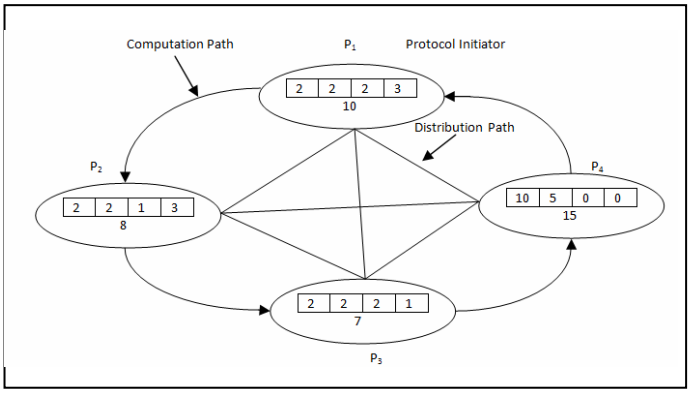


Fig. 5.Dk-Secure Sum Protocol after redistribution

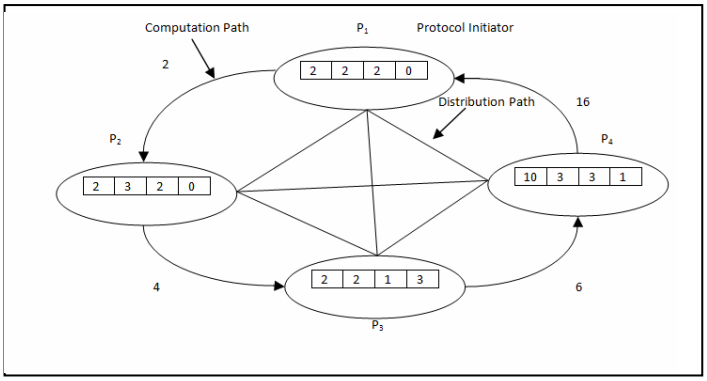


Fig. 6.Dk-Secure Sum Protocol for first round of segment computation

5.4 Secure Sum Protocol Using Hybrid Technique

Hybrid model of Secure Sum Computation is proposed as shown in figure 6. In Hybrid model [11] third party and individual parties both do computation partially at their end. In this protocol each party divides its data in three segments and with each segment parties adds different random number. In this protocol N parties and one third parties exist. N Parties compute the Sum of their data by means of the help of third party. Third party is not trusted so for privacy and security of data, data is partition into segments. Fragment of the data is done on the parties side, no method is proposed for the segmentation of data, it is on party how they partition their data in segments only number of fragment is previously announced . All party partitions their data in three segments. Computation of these segments is done by announcement between parties and third party. For more Security and Privacy random numbers are added with the segments. After computation of sum result is announced by third party to all the parties.

Working steps of hybrid technique using Secure Sum protocol

Step1. Every party send its sum of initial fragment D11, D21, D13,….Dn1 and random no. r11, r21, r31…..rn1 to third party.

Step2.

(i) Third party do sum of the entire initial fragment received from all the parties P1, P2, P3….Pn i.e. S.

(ii) Third party send sum S to party P1.

Step3. Party Pi subtracts its random no. ri1 and adds its second fragment Di2 and its random no. ri2 and then sends sum to next party Pi+1. This step repeat till I=N.

Step4. Party Pn send sum S to Pn-1.

Step5. Party Pn-1 subtracts its random no. ri2 and adds its third fragment Di3 and its random no. ri3 and send sum to earlier party Pi-1. This step repeat till I=1

Step6. Party P1 send sum S to TP and TP send this sum to Pn.

Step7. Party Pn subtracts its random no. rn2 and add its third fragment Dn3 and send sum to Pn-1.

Step8. Party Pi-1 subtracts its random no. ri3 and send sum to Pi-2. Repeat this step till I=1.

Step9. Party P1 sends sum S to third party.

Step10. Third party broadcast the sum S to P1, P2, P3,….Pn.

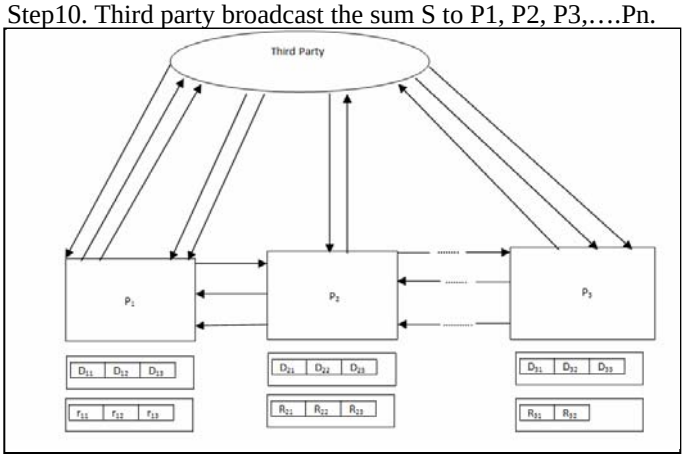
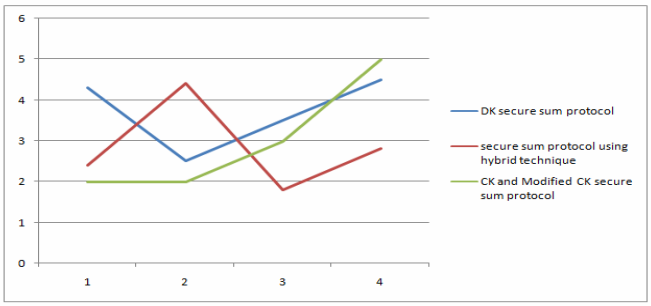


Fig. 7. Architecture of Secure Sum Protocol using hybrid technique

6 Conclusion and Future work

In this paper we present three main protocols for providing security to the database in every site of data with zero percentage of data leakages with high security to the database. The communicational complexity of the Secure Ck and Modified Ck Secure sum protocol is (n^2) And some modification is done in Dk Secure Sum protocol but it still the rate of (n^2) with zero percentage of data leakage but in Secure sum protocol using Hybrid technique the complexity is decreases to (n^2) with zero percentage of data leakage so in future we will try to find a protocol to reduce the complexity (n) with zero percentage of data leakage.



Communicational complexity vs. Number of parties

Figure: Communicational complexity

7 References

1. Yao,A.C.: Protocol for Secure computations, in proceedings of the 23rd annual IEEE symposium on foundation of computer science, pp. 160-164, IEEE Press, Chicago, USA, (1982).
2. Yao,A.C.C.: How to generate and exchange secrets (extended abstract).In Proceedings of the 27th IEEE Symposium on Foundations of Computer Science (FOCS), IEEE Press USA, (1986).
3. Agrawal ,R., et al.: Mining association rules between sets of items in large database. In Proc. of ACM SIGMOD'93, D.C.pp.207-216, ACM Press, Washington, (1993).
4. Agarwal, R., Imielinski, T ., Swamy A.: Mining Association Rules between Sets of Items in Large Databases, Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207-210, ACM Press, Washington, (1993)
5. Srikant, R., Agrawal, R.: Mining generalized association rules. In VLDB'95, pp.479-488, Zurich, Switzerland, (1994).
6. Agrawal, R., Srikant, R.: Privacy-Preserving Data Mining, In proceedings of the 2000 ACM SIGMOD on management of data, pp. 439-450, ACM Press, Dallas, TX USA (2000).
7. Lindell,Y., Pinkas,B.,.: Privacy Preserving data mining, In Proceedings of 20th Annual International Cryptology Conference (CRYPTO), Santa Barbara, California, USA, (2000).
8. Kantarcioglu,M., Clifto, C.,.: 'Privacy-Preserving distributed mining of association rules on horizontally partitioned data' In IEEE Transactions on Knowledge and Data Engineering Journal, vol 16(9), 1026-1037, IEEE Press, (2004).
9. Sheikh,R., Kumar,B., Mishra, D, K.,.: 'Changing Neighbors k-Secure Sum Protocol for Secure Multi-Party computation' International Journal of Computer Science and Information Security, vol 7, No. 1,239-243, USA, (2010).

10.Sheikh,R., Kumar,B., Mishra, D,K.,.: 'A Distributed k- Secure Sum Protocol for Secure Multi-Party Computations' JOURNAL OF COMPUTING, vol 2, USA,(2010)
11. Jangde,P., Chandel, G, S., Mishra, D, K.,.: 'Hybrid Technique for Secure Sum Protocol' World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 vol 1, No. 5,198-201, (2011).
12.Sugumar, Jayakumar,R., Rengarajan,C.,.:'Design a Secure MultipartyComputation System for Privacy Preserving Data Mining' International Journal of Computer Science and Telecommunications, vol 3, 101-105 (2012).
13. Han, J. Kamber,M.,.: Data mining. Concepts and Techniques. Morgan Kaufmann, San Francisco (2006).

AUTHORS PROFILE

Jayanti Dansana
School of Computer Engineering, KIIT University, ODISHA, INDIA
jayantifcs@kiit.ac.in

Raghvendra Kumar
School of Computer Engineering KIIT Univeristy,
Bhubaneswar, Odisha, India
raghvendraagrawal7@gmail.com

Jyotirmayee Rautaray
School of Computer Engineering KIIT Univeristy,
Bhubaneswar, Odisha, India
jyotirmayee.1990@gmail.com