An Approach for Privacy Preservation of Distributed Data in Peer-to-Peer Network using Multiparty Computation

Hemanta Kumar Bhuyan[1], Narendra Kumar Kamila[2] and Sanjit Kumar Dash[3]

**[1] Dept of CSE, Mahavir Institute of Engineering & Technology, Biju Patanaik University of Technology
Bhubaneswar, Odisha, India**

**[2] Dept of CSE, C V Raman College of Engineering, Biju Patanaik University of Technology
Bhubaneswar, Odisha, India**

**[3] Department of IT, College of Engineering & Technology, Biju Patanaik University of Technology
Bhubaneswar, Odisha, India**

Abstract: Use of technology for data collection and analysis has seen an unprecedented growth in the last couple of decades. Individuals and organizations generate huge amount of data through everyday activities. This data is either centralized for pattern identification or mined in a distributed fashion for efficient knowledge discovery and collaborative computation. This has raised serious concerns about privacy issues. The data mining community has responded to this challenge by developing a new breed of algorithms that are privacy preserving. The main objective of data mining is to extract the identified pattern for efficient knowledge discovery with centralized or decentralized collaborative computation. This paper focuses on developing secure computational model for preserving the privacy of the distributed data by performing multiparty computation in peer-to-peer network. However this approach requires that participating parties are attached to the coordinator of the peer-to-peer network through a specified path and maintain privacy by performing certain application specific computation on their local site. The computation is performed by taking the distributed data-set of a particular scenario through centralized and decentralized fashion.

Keywords: Distributed data, distributed data mining, privacy preservation, secure evaluation, peer-to-peer network, perturbation

1. INTRODUCTION

Advances in technology have enabled collection of a huge amount of data about individuals, groups or organizations from a wide variety of sources. The data collection and subsequent data mining leads to an issue of privacy. Privacy preserving data mining is a growing field of research that tries to address the issue of privacy in the context of data mining. The objective of the field of privacy preserving data mining is to modify the data or the data mining protocols in such a way that the 'privacy' of the subject is preserved while providing utility in terms of the mining results. When the private data is distributed across multiple data repositories owned by different parties, privacy preservation becomes a d ifferent kind of challenge due to personal preferences while doing distributed data mining. Preserving the privacy of user data is always a challenging task. The performance of traditional algorithm, methods, models are poor in maintain privacy in distributed data mining. More advancement is required in the existing technology to meet the increasing demand of peers for preserving the privacy of their data.

With advanced technology, many people in our society maintain rich status by applying any type conspiracy. The conspiracy by adversary is more important to distort the data in distributed data mining. Data may be distributed among multiple parties to make the acquisition of knowledge. So it needs the privacy or data secrecy to prevent the collaborative computation for distributed data mining. The privacy preserving distributed data mining requires algorithm, models, methods to perform collaborative computation for disclosing only computed result but without revealing any party's original data and also only disclose with the each party's computational result. In this paper, a modified technique is applied to the distributed data mining for privacy preserving of user's data. All users are involved with same computation except Coordinator of the network. The perturbation techniques are used to protect the user's data. It is also used both centralized and decentralized distributional computation. It may be cost effective but it covers large number of users in isolated area.

The rest of this paper is organized as follows: Section 2 presents some existing research in this area. Section 3 presents the problem definition. In Section 4, we have designed our framework and theoretically analyze the algorithm. Section 5 gives our experimental result. Finally, section 6 concludes our work.

2. Related work

Both peer-to-peer network and centralized network techniques are required to develop the technologies which are derived in this paper. Generally preservation of privacy of user's data is worked upon as security or protected communities in distributed data mining and also Privacy (no discloser of original data) and H igh communication efficiency (high speed of data exchange among the different participating parties) are required to develop the advanced technologies. But several approaches and models have already been implemented for privacy preservation in distributed data mining with collaborative computation among parties [1, 2]. The game theory is also used in privacy preserving distributed data mining (PPDDM) with multiparty computation [3]. Recently some new techniques have also evolved like the semi-trusted concept used for PPDDM [4] where several perturbation techniques are used [5, 6, 7].

Several types of techniques have already been developed for altering data in distributed data mining with the privacy preserving protocols [8, 9, 10, 11]. Distributed data mining deals with the problem of data analysis with regards to distributed data, computing nodes, communication cost, computational task and user problem. It has been involved with both centralized and decentralized environment with different sites and has already been developed for homogenous and heterogeneous data distribution [12, 13].

Peer-to-peer data mining can be worked upon as massive network with decentralized administrator upon participating autonomous nodes and monitoring all of their activities .It requires high scalable and communicational efficient algorithm for distributed data mining with approximate or exact techniques [14].

3. Problem statement

The secure multiparty computation techniques are involved with many secure technique algorithms such as secure sum, secure scalar product, secure tree operations etc, to provide security to data of different parties in different computational model. The above techniques may solve the security issue in computational models but it may not work perfectly for any computational model with large number of users with collaborative computation and with huge amount of data. The question arises how to solve the different computational problem of large number of users participating in the system. In distributed environments, the computation and communication plays the important role to solve this type of problem. The proposed methodology has been introduced to meet the following limitations of various techniques:

1) In ring network, secure sum computation provides privacy but it can be applied to few number of peers.
2) The peers in ring network work in a decentralized manner which involves more overhead.

In our methodology we have combined both centralized and decentralized environment over large number of distributed nodes. In our approach, the distributed data is collected from various sites through a specific route after performing the computation at the local site. The computed result of all routes is further computed at the coordinator to prepare the final result of whole system. Then the final result is sent to all nodes.

4. **Our Approach for Privacy Preserving Distributed Computation Model**

4.1 Framework

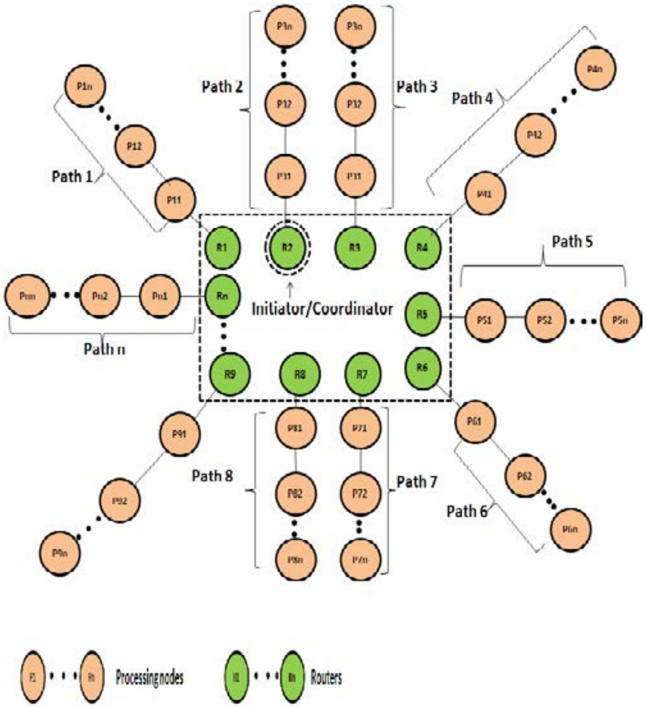


Fig. 1 Decentralized distributed peers over centralized Ring network

The proposed framework has been prepared with both synchronous and asynchronous technique where multiple parties can compute several task. The framework contains three types of participating nodes and communication link between the nodes.

1. Coordinator (C) who will control and coordinates the task among the network. Any router can act as a coordinator for a specific application. It comprises of following things: Routers (**R_1, R_2, ...R_n**) , IP address of nodes, Port number and a token for every router (**T_1, T_2, ...T_n**). The token keeps the value of the number of nodes present in a particular route.
2. Routers (**R_1, R_2, … R_n**) are also nodes present in the network. They are given special names in order to indicate that each router identifies a particular path. All nodes in particular path are connected to the coordinator via routers.
3. Nodes (**N_{11}, N_{12}... N_{1n}, N_{21}...N_{2n},....N_{ij}**) -: Each node in the network recognized as user who are participating in the network, they compute same task with own data. The number of nodes attack with each route depending upon the distance nearer to route.
4. A duplex communication link exists among the nodes.

The frame work can be explained as follows:
Let N_{11}, **N_{12}... N_{1n}, N_{21}...N_{2n},....N_{ij}** are large number of nodes present in a particular network. The existing nodes in a particular path are connected with the coordinator through a specific router. The coordinator assigns the routers a number of nodes with a specified path. The participating nodes compute their result and sent their computed result to the coordinator through their neighbors on the direction of the coordinator. The result is finally available at the coordinator after performing computation at every intermediate node. The cost of communication and computation of each route depend on the number of participating nodes in a particular route. Coordinator then performs the final computation on results received from different router by considering a number of predefined features. In this way Coordinator collects the data from different routers which are subsequently received from various nodes.

4.2 Working

The working of the proposed framework can be described as follows:

1) Coordinator creates a number of routes by assigning them a router R_1, R_2,.....$R_{n.}$
2) Select the peers for each route on the basis of shortest distance from the route & collect the IP Address, Port number of each peer.
3) After assigning the peers to the route, the coordinator assigns a token to each router. The token (T_1, T_2...$T_{n)}$ stores the number of peers present in a particular route by counting the number of nodes.
4) Coordinator sends a particular token for particular route to count the number of peers start from 0 to max.
5) Each peer sends the data to its neighbor peer on the path to router by performing secured computation.
6) Finally the router receives the computed result of the specified route, after processing of the data and result at every intermediate node.
7) Then coordinator collects the computed result from each router present in the network.
8) It then performs secured computation on the received result and finally the computed result will be sent back to the peers through the specified path.

4.3 Secure Evaluation

Let each node P_{ij} can be recognized as a participating party where 'i' corresponds to route number and 'j' corresponds to node number in a particular route. The parties P_{i1}....P_{in} are participating in a particular route $R_{i,}$ which is subsequently connected to the coordinator. So R_1 contains P_{11}, P_{12}, P_{13}.....P_{1n}, R_2 contains R_{21}, R_{22}, ...R_{2n} and R_n contains P_{n1}, P_{n2},... P_{nn}. So total numbers of nodes present in the network is

$$N=\sum_{i=1}^{n}\sum_{j=1}^{n} P_{ij}$$

Let each Party has own data X_{ij}. Each party performs secure evaluation on its own data and converts it to another form. This conversion can take place by using the following equation:

$$X_{ij}{*}\ C+Z=Y_{ij} \qquad \text{---------} \ (1)$$

Where X_{ij} is individual nodes' data and C and Z are fixed numbers for every party in a particular route. The value of C and Z may change from one route to another. Y_{ij} value is the resultant value that each node will send towards router after secured evaluation. The route receives the summation of all Y_{ij} which can be described as

$$\sum Y_{ij} = \left(\sum_{j=1}^{N} X_{ij}\right){*}\mathrm{C+nZ} \qquad \text{--------} \ (2)$$

Where i remain constant for a particular route and n is the number of nodes present in a particular route. Then the router will convert them into original data by following the equation given below:

$$R_i = \sum_{j=1}^{N} X_{ij} = \left(\sum Y_{ij} - nZ\right)\Big/ C \qquad \text{--- (3)}$$

Where R_i is the summation of all originals values at a particular router. Now all the routers present in the network will perform secure evaluation by using the equation given below:

$$R_i * P + Q = R_j \qquad \text{---------- (4)}$$

Where R_j is the perturbed result at each router and P & Q are constants which is fixed by the coordinator. After performing the secured evaluation, all routers will send their result to the coordinator. The coordinator receives the summation of all R_j values from respective routers by the following equation:

$$\sum R_j = \left(\sum R_i * P\right) + nQ \qquad \text{---------- (5)}$$

Then the coordinator retrieves the original data from the received result given in equation (5) by using the following equation:

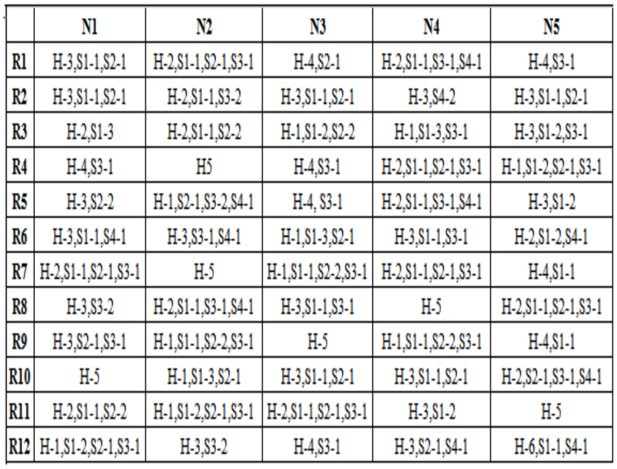
$$R = \sum R_i = \left(\sum R_j - nQ\right)\Big/ P \qquad \text{------- (6)}$$

Where R is the final result. Then the coordinator sends the final result to all nodes present in the network through their corresponding router.

5. Experimental Result

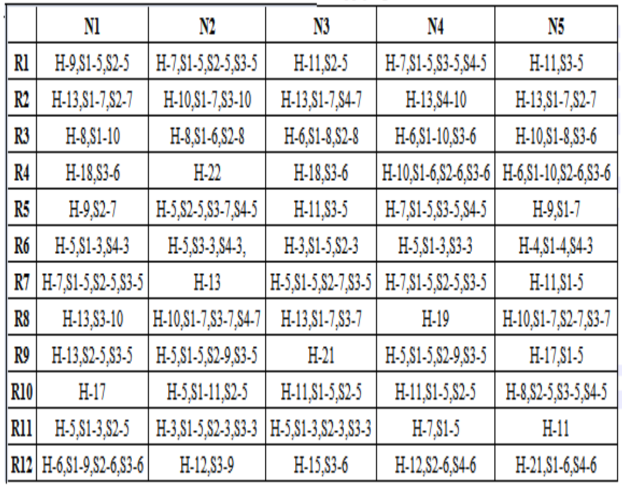
The above methodology is experimented on Heart disease data sets taken from the UCI machine learning repository. Our experiment is based on class distribution with several instances. In our experiment we have taken 12 routes, each having 5 nodes. We are considering only 14 features out of 76 features which are actually derived from 300 instances from the dataset. Table 1 contains the original data present at each node. Each node contains maximum five classes of data that is healthy (H), Seek-1(S1), Seek-2(S2), Seek-3(S3) and seek-4(S4). The result indicates that whether patient is healthy or sick. Then the node evaluates its own data by using the equation (1) and the computed result is shown in Table 2. Nodes will send their computed result to the corresponding router by a specified path through their neighbor nodes. Table 3 c ontains the summation of all nodes perturbed data. Then router retrieves the original data by equation (3). The summation of all original data received from all nodes is given in Table 4. Then the router sends the perturbed data to the coordinator which is given in Table 5. Table 6 finally shows the original data which has been retrieved by the coordinator from the data sent by routers.

TABLE 1 Original Data at each node



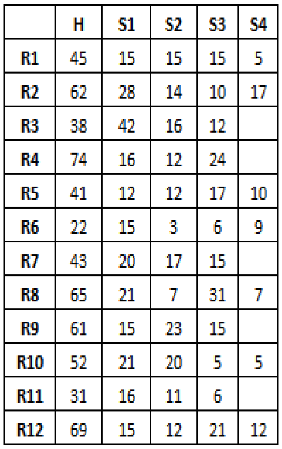
	N1	N2	N3	N4	N5
R1	H-3,S1-1,S2-1	H-2,S1-1,S2-1,S3-1	H-4,S2-1	H-2,S1-1,S3-1,S4-1	H-4,S3-1
R2	H-3,S1-1,S2-1	H-2,S1-1,S3-2	H-3,S1-1,S2-1	H-3,S4-2	H-3,S1-1,S2-1
R3	H-2,S1-3	H-2,S1-1,S2-2	H-1,S1-2,S2-2	H-1,S1-3,S3-1	H-3,S1-2,S3-1
R4	H-4,S3-1	H5	H-4,S3-1	H-2,S1-1,S2-1,S3-1	H-1,S1-2,S2-1,S3-1
R5	H-3,S2-2	H-1,S2-1,S3-2,S4-1	H-4, S3-1	H-2,S1-1,S3-1,S4-1	H-3,S1-2
R6	H-3,S1-1,S4-1	H-3,S3-1,S4-1	H-1,S1-3,S2-1	H-3,S1-1,S3-1	H-2,S1-2,S4-1
R7	H-2,S1-1,S2-1,S3-1	H-5	H-1,S1-1,S2-2,S3-1	H-2,S1-1,S2-1,S3-1	H-4,S1-1
R8	H-3,S3-2	H-2,S1-1,S3-1,S4-1	H-3,S1-1,S3-1	H-5	H-2,S1-1,S2-1,S3-1
R9	H-3,S2-1,S3-1	H-1,S1-1,S2-2,S3-1	H-5	H-1,S1-1,S2-2,S3-1	H-4,S1-1
R10	H-5	H-1,S1-3,S2-1	H-3,S1-1,S2-1	H-3,S1-1,S2-1	H-2,S2-1,S3-1,S4-1
R11	H-2,S1-1,S2-2	H-1,S1-2,S2-1,S3-1	H-2,S1-1,S2-1,S3-1	H-3,S1-2	H-5
R12	H-1,S1-2,S2-1,S3-1	H-3,S3-2	H-4,S3-1	H-3,S2-1,S4-1	H-6,S1-1,S4-1

TABLE 2 Data at each node after applying Perturbation



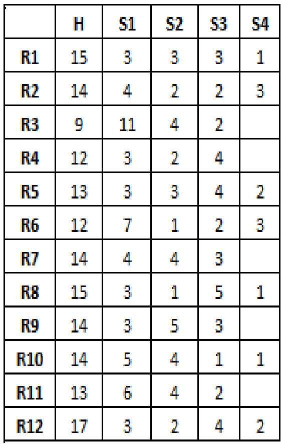
	N1	N2	N3	N4	N5
R1	H-9,S1-5,S2-5	H-7,S1-5,S2-5,S3-5	H-11,S2-5	H-7,S1-5,S3-5,S4-5	H-11,S3-5
R2	H-13,S1-7,S2-7	H-10,S1-7,S3-10	H-13,S1-7,S4-7	H-13,S4-10	H-13,S1-7,S2-7
R3	H-8,S1-10	H-8,S1-6,S2-8	H-6,S1-8,S2-8	H-6,S1-10,S3-6	H-10,S1-8,S3-6
R4	H-18,S3-6	H-22	H-18,S3-6	H-10,S1-6,S2-6,S3-6	H-6,S1-10,S2-6,S3-6
R5	H-9,S2-7	H-5,S2-5,S3-7,S4-5	H-11,S3-5	H-7,S1-5,S3-5,S4-5	H-9,S1-7
R6	H-5,S1-3,S4-3	H-5,S3-3,S4-3,	H-3,S1-5,S2-3	H-5,S1-3,S3-3	H-4,S1-4,S4-3
R7	H-7,S1-5,S2-5,S3-5	H-13	H-5,S1-5,S2-7,S3-5	H-7,S1-5,S2-5,S3-5	H-11,S1-5
R8	H-13,S3-10	H-10,S1-7,S3-7,S4-7	H-13,S1-7,S3-7	H-19	H-10,S1-7,S2-7,S3-7
R9	H-13,S2-5,S3-5	H-5,S1-5,S2-9,S3-5	H-21	H-5,S1-5,S2-9,S3-5	H-17,S1-5
R10	H-17	H-5,S1-11,S2-5	H-11,S1-5,S2-5	H-11,S1-5,S2-5	H-8,S2-5,S3-5,S4-5
R11	H-5,S1-3,S2-5	H-3,S1-5,S2-3,S3-3	H-5,S1-3,S2-3,S3-3	H-7,S1-5	H-11
R12	H-6,S1-9,S2-6,S3-6	H-12,S3-9	H-15,S3-6	H-12,S2-6,S4-6	H-21,S1-6,S4-6

TABLE 3 Perturbed data at each route



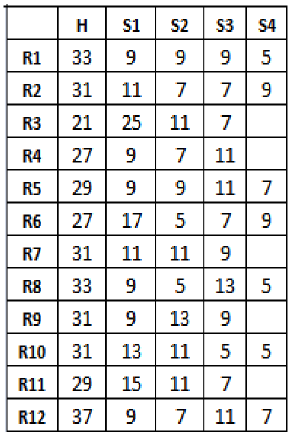
	H	S1	S2	S3	S4
R1	45	15	15	15	5
R2	62	28	14	10	17
R3	38	42	16	12	
R4	74	16	12	24	
R5	41	12	12	17	10
R6	22	15	3	6	9
R7	43	20	17	15	
R8	65	21	7	31	7
R9	61	15	23	15	
R10	52	21	20	5	5
R11	31	16	11	6	
R12	69	15	12	21	12

TABLE 4 Original data at router after retrival



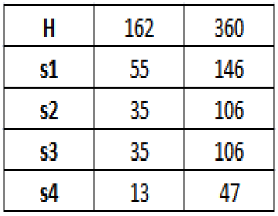
	H	S1	S2	S3	S4
R1	15	3	3	3	1
R2	14	4	2	2	3
R3	9	11	4	2	
R4	12	3	2	4	
R5	13	3	3	4	2
R6	12	7	1	2	3
R7	14	4	4	3	
R8	15	3	1	5	1
R9	14	3	5	3	
R10	14	5	4	1	1
R11	13	6	4	2	
R12	17	3	2	4	2

TABLE 5 Perturbed data sent by router



	H	S1	S2	S3	S4
R1	33	9	9	9	5
R2	31	11	7	7	9
R3	21	25	11	7	
R4	27	9	7	11	
R5	29	9	9	11	7
R6	27	17	5	7	9
R7	31	11	11	9	
R8	33	9	5	13	5
R9	31	9	13	9	
R10	31	13	11	5	5
R11	29	15	11	7	
R12	37	9	7	11	7

TABLE6 Original & Perturbed data at coordinator



H	162	360
s1	55	146
s2	35	106
s3	35	106
s4	13	47

Figure 2 represents the chart concerning the perturbed data at 12 routers. Figure 3 represents the reperturbed data sent by routers to coordinator and Figure 4 represents the final result available at coordinator.

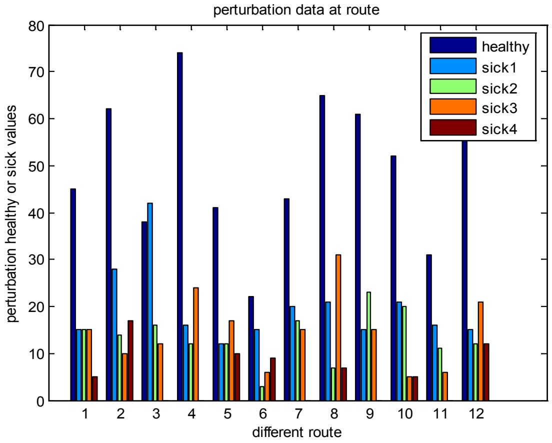


Fig. 2 Perturbed Data at Routers

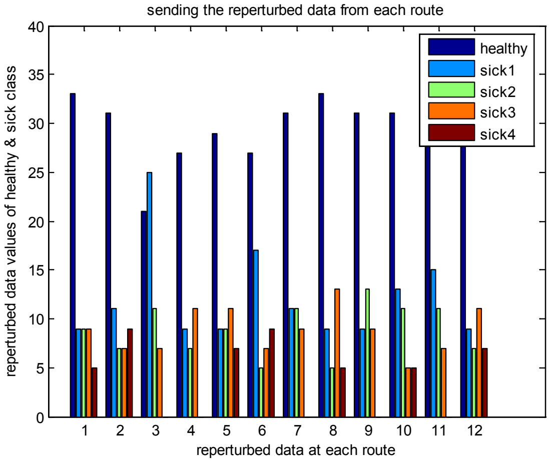


Fig. 3 Reperturbed data sent by router to coordinator

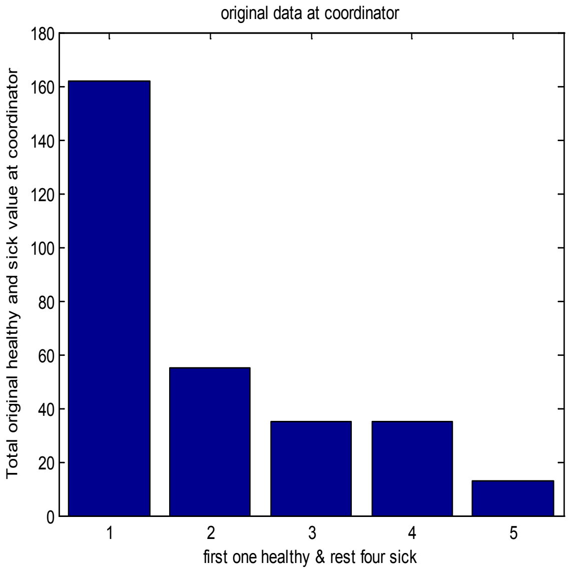


Fig. 4 Final Result available at coordinator

6. Conclusion

The proposed approach in this paper is about privacy issue in distributed data mining and knowledge protection techniques using perturbation technique. In this study, the perturbation methods are effective in preserving the privacy of the data in an effective manner and preserve the accuracy of the original dataset and also maintain the consistency of user data. In this approach, we are providing more privacy by performing double computation at node and router. This paper presents a synchronous/asynchronous method for monitoring and preserving the privacy of data. The method is efficient and exact in the sense that once the computation terminates each node in the network gets global correct result. Due to constant communication complexity and locally synchronous nature of the methodology, it i s highly scalable. This method can be applied to large scale heterogeneous distributed system and has various application that require privacy preserving data mining

References

[1]. R. Agrawal and R.Srikant. Privacy preserving data mining. In ACM SIGMOD, pages 0439-450, may 2000.

[2]. S.Jha, L.Kruger, and P. MC Daniel. Privacy preserving clustering. In ESORICS, Pages 397-417, 2005.

[3]. H. Kargupta, K. Das, K. Liu. Multiparty. p rivacy preserving distributed data mining in distributed data mining using a g ame theoretic framework. In 11th European conference on principles and practice of

knowledge discovery in data bases (PKDD),Pp 523-531,2007.

[4]. M.G.Kaosar, X.Yi. Semi-trusted mizer based privacy preserving distributed data mining for resource constrained Devices. IJCSIS, Vol.8.No.1, April 2010.

[5]. M.A.Kadampur, D.V.L.N. Somayajulu. A noise Addition scheme in Decision tree for privacy preserving data mining. Journal of computing, volume 2, issue 1, pp 137-144, January 2010.

[6]. P.Kamakshi, A.V.Babu. Preserving privacy and sharing the data in distributed environment using crypto graphic technique on perturbed data. Journal of computing volume 2, issue 4, pp 115-119, April 2010.

[7]. P.Kamakshi, A.V.Babu. Preserving privacy and sharing the data using classification on perturbed data. ICSE, Vol.02, No.03, Pp 860-864, 2010.

[8]. D.Agrawal and C.C. Agarwal. On the design and qualification of privacy preserving data mining algorithms. In proceedings of the twentieth ACM SIGACT-SIGMOD-SIGART symposium on principles of database systems. Santa Barbara, Califarnia, USA: ACM May 21-23, 2001, pp.247-255.

[9]. A. Evfimievski, R. Srikant, R. Agrawal and J. Gehrke. Privacy preserving mining of association rules. In the Eighth ACMSIGKDD, International conference on knowledge Discovery and data mining Edmonton, Alberta, Canada, July 23-26, 2006, pp.217-228.

[10]. M.Kantarcioghu and C.Clifron. Privacy reserving distributed mining of association rules on horizontally partitioned data .In the ACM SIGMOD workshop in research issues on data mining and knowledge discovery (DMKD 02), Madison, Wisconsin, June 2, 2002, Pp.24-31.

[11]. J. Vaidya and C. Clifton. Privacy preserving association rule mining in vertically partitioned data. In the Eight ACM SIGKDD international conference on knowledge Discovery and data mining Edmonton, Alberta, Canada, July 23-26,2002, Pp 639-644.

[12]. H.Kargupta and K. Sivakumar. Existential pleasures of distributed data mining. Pages 1-25. AAAI/MIT press 2004.

[13]. Mj.Zaki. parallel and distributed association mining: A survey. IEEE concurrency, 7(4):14-25, 1999.

[14]. S.Datta, K.Bhaduri, C.Giannella, R.Wolff and H.Kargupta. Distributed data mining in peer-to-peer networks. IEEE internet computing, 10(4): 18-26, 2006.

Hemanta Kumar Bhuyan is a research scholar (PhD) in the Department of Computer Science and Engineering, Sikhya 'O' Anusandhan (SOA) University, Odisha, India. He obtained his M.Tech degree in Computer Science and Engineering from Utkal University, Odisha, India in 2005. He is currently working as an Assistant Professor in the department of computer science & engineering, Mahavir Institute of Engineering and Technology, Odisha, India. He has more than 12 y ears of teaching experiences in various academic organizations. His research interests include privacy preserving, distributed data mining, feature extraction.

Narendra Kumar Kamila is a Professor in the department of computer science and engineering in C.V. Raman engineering college, Odisha, India. He obtained his master of technology in computer science from IIT, Kharagpur, PhD from Utkal University, Odisha, India. He has awarded Post-Doctorate Fellowship from USA, in 2005. He has more than 18 years of teaching experience and he h as guided several research scholars. He is a member of IEEE. His research interests include Privacy/security policy, data mining, databases, neural network, artificial intelligent, soft computing, and sensor network.

Sanjit Kumar Dash received the B.Tech. Degree in Information Technology from Biju Patnaik University of Technology, Odisha, India, in 2004 and pursuing M.Tech. Degree in Computer Science and Engineering at Institute of Technical Education and Research, Bhubaneswar, India. He is also working as a faculty member at the Information Technology Department, College of Engineering and Technology, Bhubaneswar, India. His research interests include Cloud Computing, Sensor Network, and Mobile Computing.