Privacy Preservation in Distributed Data Mining Using Secured Multiparty Computation

Dillip Kumar Swain[1] Sarojananda Mishra[2] Subhendu Bhusan Rout[3]

[1]Department of CSE&A, IGIT Sarang, Odisha, India, dkswain_41@yahoo.com
[2]Department of CSE&A, IGIT Sarang, Odisha, India, sarose.mishra@gmail.com
[3]Department of CSE&A, IGIT Sarang, Odisha, Indi, subhendu.as@gmail.com

Abstract

Data mining is the theory deals with the knowledge discovery or gaining the meaningful data from a data base. Data sharing among two parties, institutions or organizations is very common for many applications like sharing of information, business planning or marketing. As these types of data sharing are folded into many times in the last few decades so Big Data Analytics is becoming very popular in order to process these data and to capture the meaningful information. Looking into this matter privacy preservation is a major issue for a distributed data mining. In a communication generally several nodes from all over the globe are participating. Each node is having different intention to share or receive data over the network. Security and trust in cooperative communication is a very big challenge for this type of sharing. So privacy preservation as well as Security of the shared data is very much necessary for this aspect. Though several research and techniques are already carried out in this topic still misuse of data and lost of information is a big challenge for today and for upcoming days. In this paper we have discussed various techniques that are being developed for the privacy preservation in a distributed data mining and we have also proposed a secured multiparty computation technique for the privacy preservation in distributed data mining. Our proposed algorithm is based upon a multiparty computation upon a distributed data mining. This technique may provide a better security to reduce the lost, stolen and misuse of data over a distributed data mining.

Index Terms- *Big Data Analytics, Data Mining, Privacy preservation, Communicating nodes, Cooperative communication*

I. INTRODUCTION

Privacy preservation in distributed data mining is one the big issue and big challenge for the current researchers. Staring from a simple Node to node connection to Internet of Thing each and every network is associated with sharing of data and information upon the network. Data mining methodologies can be defined as identifying the trends and patterns from huge data base or from a large amount of data. Data warehousing and data mining both terms are very associated with each other. This technique is mostly gathering huge amount of data into a common platform and applying an algorithm to find the useful, fruitful and meaningful information.

Privacy preservation means preventing information from disclosure due to legitimate assessment of the data and information. Some time privacy preserving is little bit different with conventional data security, encryption technology and access control that tries to prevent information disclosure against the unauthorized means. The main objective in privacy preserving data mining is a two way path. First of all, sensitive raw data such as names, identifiers, addresses, should be modified or secured from the original database, in the way that for the recipient of the data should not compromise with another person's privacy. Secondly, the sensitive knowledge that can be mapped from a database by using data mining algorithms should be excluded; as such type of knowledge can equally compromise with data privacy. As privacy is a major issue in all type of data sharing so privacy preservation is very much required for all these type of sharing which may be a simple data sharing to sharing of a big data set or data mining.

Secured Multiparty technique belongs to securing the whole data base or data mining in the presence of several users or applying security technique at different user's side. In this scenario more than one node can be chosen as the allegiant nodes for a neighbor node. In order to choose the „Allegiant" nodes for a parent node we have taken the „linier regression model"

technique. The neighbor nodes will take part for the privacy preservation and security of the parent node. For a particular parent node there may be a number of allegiant nodes should be chosen. In this way privacy preservation for a particular node or network can be maintained in a distributed data mining using a secured multiparty cooperation or computation. As the linier model takes several parameters and makes huge assumptions about structures so linier models can be chosen to capture the allegiant nodes.

This paper is organized as follows. In section II we have provided a brief review of data mining, distributed data mining and its applications. In section III we have discussed about the necessity of privacy preservation in data mining as well as distributed data mining. In section IV we have provide a detail algorithm and a new proposed technique for this privacy preservation problem in data mining. The section V provides the details simulation environment as well as simulation result of the algorithm and proposed techniques of the previous section. Finally the paper concludes in section VI with conclusion and future work.

II. DISTRIBUTED DATA MINING

Data mining is the process of extracting important data from large data collection. Sometimes these are splits over large number of nodes. Data warehousing is the process of bringing the data from multiple resources under a single authority which may increase the privacy violation. In other way privacy matter may obstruct user to share data over the network directly. In observing this issue distributed data mining provides a mean to address this issue, particularly to avoid the disclosure of any information before the final publish. Data mining generally focuses on producing a general model instead of focusing on a specific model or individual. The data which is very sensible need more secure when it is considered upon a single platform to share over a network or over large number of nodes. In order to process huge amount of data that are being distributed over a large number of nodes or networks need more precise and sharp securing method before final result is being published. So in this case as multiple nodes take part in the data processing and sharing fraudulent activities between the nodes and data lost is a common issue. So in this issue distributed data mining provides a good platform to process the data as well as to share the data in a common platform to find out the desired result.

For an example we can take an example of the result of cultivation and situation of the farmers over the different states of India. We can collect the data from the different state head quarters to create common results upon a common platform in order to provide the insurance to the farmers. In this case all data are being available with the data centre of all the state. As there may be 25-30 state take part in this research so there may be the probability that there may be the probability fraudulent activities. If we take the quantity of irrigation land(X) of a state to that of extraction of paddy(Y) a relation may establish from X and Y that is Pr(Y/X). Some states are having less amount of irrigation land and also less number of water resources is available. So in these states other type crops like pea, red lentils, and yellow lentils are being cultivated. A union of other different type of crops may be considered for producing the results and probability set up. Here the problem may be that the insurance companies should be concerned about the sharing of this data. Not only the privacy of farmers records be maintained properly, but the insurers may be unwilling to release the rules pertaining them. Considering a result indicating a high rate of disputes concerning with suicide of farmers related to different issues, the insurance company should be more precise and pinpoint about the problem so as to design the insurance policies. One solution for this problem is to avoid disclosing of data resources, while still the constructing data mining models equivalent to those that would have not finalized. The process of distributed data mining decreases the opportunity of misuse of data and information.

III. PRIVACY PRESERVATION IN DATA MINING

The internet is one of the best communication medium among several parties or several businesses in the last two decades. The privacy preservation in the private data is one of the major issue for each type of data sharing. Maintaining the privacy preservation with anonymous ID is a good technique. In[1] they have proposed a technique for privacy preservation in data sharing among several nodes with anonymous ID assignment. In this paper they have taken N nodes. They

have developed an Algorithm for sharing private data among several parties. These N nodes are assigned ID numbers starting from 1 to N. In this paper they have taken a comparison in between slot selection AIDA, Prime Modulus AIDA, Sturm's Theorem AIDA methods etc. This technique decreases the communication overhead. This also enable the use of more number of slots with the reduction in the round[1].

Sharing of data is very much essential in many areas like medical research, bioinformatics park, business hubs and also in generalized marketing. In this sharing there should be some sensitive data which should not be disclosed. If we consider the medical databases these type data require more and security in case of sharing as well as maintain the security. In [2] S. Lohiya et. al. proposes a method called a Hybrid approach for privacy preserving. In the First step they are randomizing the original data. Then they have applied generalization on randomized or modified data. This method protects the private data with better scalability and accuracy. It may reconstruct the original data and can provide data without information loss and makes the best usability of the data. Privacy preserving is very essential in case of cellular and ad-hoc network also[3][4]. In [3] D.K. Swain et. al. provides an proposed techniques to provide privacy preservation for the privileged information. The Cellular network data also carry much potential as an input to applications that are based on high-coverage with low-resolution in comparison with both space and time. So these type of network also require more potential and security for the privacy preservation of the sharable data. In [4] they have presented a compound approach for mining route information from large volumes of cellular network data. This approach addresses scalability by enabling efficient reduction of sequence data through distributed clustering, and privacy. If it needed it Groups the raw data in aggregation and after that perturbing statistics about these aggregation in order to achieve further privacy protection [4].

Cloud computing is one of the best techniques for the requirement of proving of data on an On-demand basis[5][6]. As the data mining is gaining more attention in society for the rapid growths of data sharing In [5] they have applied cloud technologies to overcome the data availability problem as well as data security issues. In this paper they have proposed a technique to provide the security by a third party service provider in the concern that including Data mining as a Service (DMaaS) for the protection of data. Though a large number of tools and techniques already being developed for the purpose of privacy preservation still Soft Computing methods is also a very powerful technique for data mining as soft computing is capable of handling partial truth and imprecision database. Soft computing models provides low cost, robust and dedicative technique over many hard computing techniques[7][8]. In [7] they have proposed a model that preserves the privacy of individuals without affecting the final results of the Neural Networks. Though several models suffer from data lost but they are getting almost same accuracy even preserving the privacy using fuzzified data.

Privacy preservation is a big challenge for many researchers till date. Though several technologies are being developed in his aspect still there is the lack of security for which the hacking of data and stolen of data creates a big problem. There are several techniques that are already developed. In [9][10][11] they have provided a brief review upon the various developed techniques in the field of privacy preservation of data base in a multiparty computation. Secure multiparty allows several parties to participate in a communication. In [12] they have applied an innovative protocol which uses both actual and idyllic model. By breaking the data blocks in to segments and by redistributing the segments among different parties with these two models they provides better security and privacy.

IV. PRIVACY PRESERVATION USING SECURED MULTI PARTY COMPUTATION

In a distributed network or distributed platform there are several nodes that participate in data communication. Though every node participates in data communication but every node is not having the same intension in the topology. Some malicious nodes may be present in the network which may hamper the data communication and drop of data packet. So privacy preservation is mostly very essential in order to maintain a good network topology as well as good network atmosphere. The main purpose of secure multi party computation is to provide security and privacy. This actually comes to success when at the end of communication the destination node only knows the result. One way is to use trusted third party and all calculation and propagation should be over this third party and other party should have no knowledge about this.

In this paper we have used multiple linier regressions to calculate the trusted nodes over a big network. These trusted nodes can be called as the

„Allegiant Nodes“ that are to be calculated. In order to calculate the allegiant nodes first we have to study the behavior of the nodes. In order to study the behavior we have focuses upon three parameters. The following algorithm provides the procedure to calculate the allegiant node.

Algorithm:

1. Calculate the average response time(X) of nodes for a data packet that is propagated over the network.

2. Find out the actual response time(R) over a node upon a particular span of time.

3. Observe the number of data packets(Z) that are being propagated over the node over a particular span of time.

4. If the average response time is „X“, actual response time over the time is „R“ and number of data packets that are propagated over the time is „Z“ than the allegiant value „A“ can be calculated as

$$A = \frac{X-R}{Z} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

In this paper we have focuses upon only the response time and the number of packets reached over the nodes upon the particular time. According to Eqn(1) if the actual response time is more than the calculate allegiant value(A) will be negative. If we consider the network over another network than the response time can be taken from the several nodes. If the number of nodes upon the small network is „n“ than „R“ can be calculated as

$$R = \sum R_1 + R_2 + R_3 + \cdots \ldots\ldots . + R_n \quad \ldots\ldots\ldots (2)$$

Here (R_1, R_2, R_3….R_n) may be taken as the actual response time of the nodes (1,2,3….n) respectively. Similarly the average response time „X“ can be calculated as

$$X = (X_1 + X_2 + X_3 + \cdots \ldots\ldots\ldots X_n)/n \ldots\ldots\ldots\ldots (3)$$

Here (X_1, X_2, X_3….X_n) may be taken as the average response time of the nodes (1,2,3….n) respectively.

V. SIMULATION SETUP AND RESULT

In this paper our main intention is to find out the allegiant node which should be considered for the secure data transmission. Linier regression model such as simple and multiple regression models are very much popular in the current days for the Big data analytics. We have taken „RGui“ for simulating and to study the behavior of nodes. „RGui“ is a very good simulator for the processing of huge amount of data with considering several predictors. We have used multiple regression models to calculate the allegiant node that are to be participating for the data communication.

Generally in the twenty first century internet and networking is one the most useful and used communication medium in between organization and peoples. As these types of nodes are very high in number and very big amount of data packets are being propagated in these days so the multiple linier regression method can be useful to capture the optimal results. For an experimental purpose we have taken 15 nodes with average response time 20 sec. In each node the in time response time should be calculated. If the number of data packet reached at the particular time span is 5 than the allegiant node can be calculated as in fig.1.

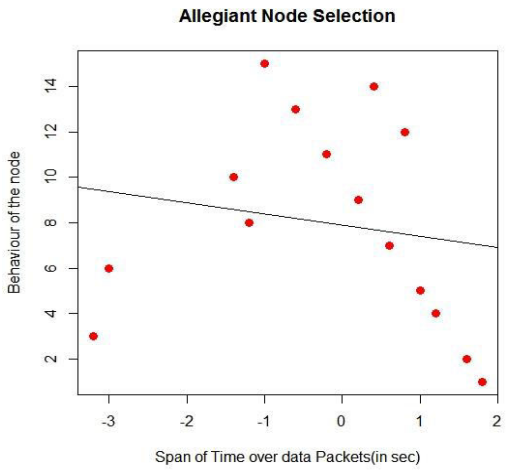

Fig.1. allegiant node selection

In this experiment we have taken fifteen nodes. Upon a natural flow in data packets over a network of 100 nodes we have chosen 15 nodes for a parent node. Our intention is to capture the allegiant nodes by taking the response time as only the parameter to calculate the allegiant node. In this natural data flow speed and with good working nodes we observe the response time upon the flow of data packets. Upon each node we have propagated 5 packets upon the network. For the set up

we have taken the average response time of 20 second. Upon the span of one minute and during the propagation of data packets we have studied the actual response time of the nodes. According to equation(1) if the response time is high so the corresponding value for „E“ will be negative. So in this aspect the value that are greater than or equal to zero can be considered as the allegiant node as these nodes are having less or equal response time.

In this experiment we found Eight nodes that are having allegiant value(A) greater than or equal to one. If the allegiant value is greater than or equal to zero than we can call these as allegiant nodes. These eight nodes can be considered as the allegiant node for the parent node. If a multiparty computation applied for this network and if the parent nodes needs to communicate over a secure and private communication than the parent node can consider these eight nodes as reliable node for the data communication.

For an experimental purpose we have taken only fifteen nodes. So regression analysis which is used to process the big data applications will be helpful for process several nodes. This process should be observed upon the network over time to time. The data packets counter should work for all the nodes that are being associated for the whole network. Most of the time the neighbor nodes of the desired node may behave like an allegiant node according to their response time, But that may be the malicious node. To overcome this problem we have to take several parameters instead of a single parameter like the response time. Similarly if we consider several parameters to calculate an allegiant node than first we have to calculate the similar property of a good behaving node than we have to apply over the network. The nodes that wants to y participate over a secure network have to solely depend upon the allegiant nodes as data lost or data stolen can be avoided in case of these nodes.

VI. CONCLUSION AND FUTURE WORK

Privacy preservation is very much essential in each and every type of network starting from wired to wireless, Ad-hoc to distributed network or sharing a small data over a network to sharing a big data analysis over the internet. Data mining is also associated with each and every type of network. Privacy preservation in distributed data mining over a multi party computation is very much essential for the recent times. Security of data and information is required in every step for building a healthy and trustful communication. Many times several security issues like stolen of data and lost of data crates a big problems for the data mining and data sharing. In this topic our proposed techniques and algorithm is basically designed for a multiparty computation over a distributed data mining. This technique will be helpful for securing the network and securing the data by providing a better security. In our upcoming research we want to focus upon Big Data Analytics and Security over the Social Networking site as well as over the multi party computation.

REFERENCES

[1] L. A. Dunning, K. Ray, „Privacy Preserving Data Sharing With Anonymous ID Assignment“, IEEE Transaction on Information Forensics and Security, Vol. 8(2), 2013.

[2] S. Lohiya, L. Ragha, „Privacy Preserving in Data Mining Using Hybrid Approach“, Fourth International Conference on Computational Intelligence and Communication Networks, IEEE, 2012.

[3] D. K. Swain, S. N. Mishra, S. Mishra, „Privacy Preserving in Data Mining to Protect Privileged information in Adhoc Network“ International Journal Of Engineering and Computer science, Vol. 5(10), 2016.

[4] O. Gornerup, N. Dokoohaki, A. Hess, „Privacy-preserving mining of frequent routes in cellular network data“, IEEE Trustcom/ BigDataSE/ISPA, 2015.

[5] A. Monreale, W. H. Wang, Privacy-Preserving Outsourcing of Data Mining“ 40th Annual Computer Software and Applications Conference, IEEE, 2016.

[6] S. B. Rout, S.N. Mishra, B.S. P. Mishra, „Mapping of Genes using Cloud Technologies“ International Journal of research in Engineering and Technology, Vol. 2(2), 2013.

[7] M. Bashir Malik, M. Asger, R. Ali, A. Sarvar, „A model for Privacy Preserving in Data Mining using Soft Computing Techniques“ Proceedings of IEEE Conference, 2015.

[8] S. B. Rout, S. Mishra, S. N. Mishra, „A Review on Application of Artificial Neural Network(ANN) on Protein Secondary Structure Prediction“ Proceedings of

International Conference on Electrical, Computer and Communication Technologies, IEEE, 2017.

[9] A. Kaur, S. Sofat „A proposed hybrid approach for Privacy Preserving Data Mining“ Proceedings of IEEE Conference, 2015.

[10] Sin G Teo, V. Lee, S. Han, „A Study of Efficiency and Accuracy of Secure Multiparty Protocol in Privacy-Preserving Data Mining“, 26th International Conference on Advanced Information Networking and Applications Workshops, IEEE, 2012.

[11] J. Wang, Y. Luo, Y. Zhao, J. Le, „A Survey on Privacy Preserving Data Mining“ Proceedings of First International Workshop on Database Technology and Applications, 2009.

[12] N. Pathak, S. Pandey „An Efficeient method for privacy preserving data mining in secured multiparty computation“ International Conference on Engineering, 2013.

[13] B. Pinkas, „Cryptographic techniques for privacy preserving data mining“, SIGKDD Explorations, HP Lab, Vol. 4(2), 2004.

[14] H. Kargupta, S. Datta, Q. Wang, S. K. Krishnamoorthy, „On the Privacy Preserving Properties of Random Data Perturbation Techniques“ Proceedings of the Third IEEE International Conference on Data Mining , 2003.

[15] S. Gomathi, N. G. Bhuvaneswari Amma, „Privacy Preserving Data Mining Approach for Extracting Fuzzy Rules“ International Conference on Green Engineering and Technologies, IEEE, 2015.

[16] M. Narwaria, S. Arya, „Privacy Preserving Data Mining - A State of the Art“ IEEE, 2016.

[17] H. Kargupta, S. Datta, Q. Wang, S. K. Krishnamoorthy, „Random-data perturbation techniques and privacy-preserving data mining“ Knowledge and Information Systems, Springer, 2005.

[18] Z. Shaikh, P. Garg, „Secure Multiparty Computation during Privacy Preserving Data Mining: Inscrutability Aided Protocol for Indian Healthcare Sector, 2012.

[19] Y. Lindell, „Secure Multiparty Computation for Privacy Preserving Data Mining“ Secure Multiparty Computation for Privacy Preserving Data Mining, 2016.

[20] Y. Shen, J. Han, H. Shan, „The Research of Privacy-preserving Clustering Algorithm“, Third International Symposium on Intelligent Information Technology and Security Informatics, IEEE, 2010.

[21] C. Clifton, M. Kantarcioglu, , X. Lin, M. Y. Zhu , „Tools for Privacy Preserving Distributed Data Mining“, SIGKDD Explorations. Vol.4(2).

[22] J. Vaidya, C. Clifton, „Privacy Preserving Association Rule Mining in Vertically Partitioned Data“, SIGKDD, 2002.

[23] Y. Lindell, B. Pinkas, „Secure Multiparty Computation for Privacy-Preserving Data Mining“, The Journal of Privacy and Confidentiality Vol. 1(1), 2009.

[24] F. Meskine, S. N. Bahloul, „Privacy Preserving K-Means Clustering: A Survey Research“ The International Arab Journal of Information Technology, Vol. 9(2), 2012.

[25] Y. Lindell, B. Pinkas, „Secure Multiparty Computation for Privacy-Preserving Data Mining“ Cryptology ePrint, 2008.