Privacy Preserving Data Mining in a Shard Database: Architectural Aspect

Mona Shah[1], Dr. Hiren D. Joshi[2]
[1](Research Scholar, RK University, Rajkot, India & Assistant Professor, JG College of Computer Applications, Ahmedabad, India)
[2](Associate Professor and Director Incharge, School of Computer Science, Dr. BabaSaheb Ambedkar Open University, Ahmedabad, India)

ABSTRACT : *Data mining as defined generally is a journey of discovering the underlying unusual, unnoticed and undetected patterns of data .It is not merely an area of interest for the research community but it has a share of inquisitiveness also – inquisitiveness in terms of finding something new, unusual, expecting something of interest and need both. This slice of curiosity in data mining adds that extra care by being meticulous while handling such data. The concept of decentralization of data introduced the need of extra care to be taken. It features parameters like prevention of misuse of data, security of data and unambiguousness of data so that it yields more meaningful, interpretable and applicable results. Scattered data over a group of sites can be analysed to find the hidden patterns which can be useful for all the involved parties. This inculcates scope for areas like secured data mining viz. Privacy preserving data mining, collaborative data mining, cooperative data mining and a few more to name. This paper is an endeavour towards proposing framework for one the focal requirements of collaborative data mining: privacy preserving data mining. A number of solutions in term of algorithm have been suggested so far to achieve Privacy Preserving Data Miming (PPDM), each with its own dynamics. This paradigm aims towards achieving accuracy while maintaining vital level of confidentiality among the participants involved in group data mining. The solution proposed suggests the use of a randomisation in selection and the use of an intermediate party also. This paper also covers the comparison between a few similar solutions in the same neighbourhood.*

Keywords– *Architecture, Data Mining, Distributed Database, Privacy Preserving.*

I. INTRODUCTION

Data mining in its classic definition says that it is the non-trivial extraction of implicit, previously unknown and potentially useful information from data[1]. The data mining term has now been quite customary and has an inter-disciplinary reach over for its findings. Since it merges well with other domains, it has a strong ability of evolving with a number of pertinent solutions in each. To name a few having major applicability, we have medicine, security, education, human resource management, financial sector and web services. This paper has been divided in to three sections: 1) Introduction 2) Related work 3) Proposed architecture for privacy preserving in a fully distributed environment over homogeneous database for multiple parties.

A number of ideas have been conceived so far in terms of proposing and devising logical design for the solution of a privacy preserving data mining in the case of a distributed database. Each solution is different in its own perspective with challenges addressed and confinements involved. The basic purpose being the same to make mining happen over the data among the parties without letting any party learn the data of any other. The distributed computing has gained momentum in the past few years with different companies joining their hands to reach out the market with data mining results

Also, for an organisation with expansion, distributed environment will be the ideal choice in terms of making data storing and data access efficient. For organisations that work on global platform need distributed data miming and require cohesive and integrated knowledge from the data.

II. RELATED WORKS

A general feature about distributed data mining is that the data under study is spread across different geographical locations. Wherein a distributed environment, every node or user in the system has partial amount of the total data, this data can be distributed homogeneously or heterogeneously. Homogeneous distribution is also referred to as horizontal partition. When it is referred as horizontal partition of a data set, it denotes different set of records with exactly the same set of attributes. Similarly, a heterogeneous distribution is called as vertical partition. It is a set of same records but with different set of attributes distributed at different locations. When such horizontal / vertical partition are combined, the complete database becomes available for further tasks. Data mining components in its simplest terms includes data, users, hardware, data mining software and a few supporting software. In case of a distributed database, parameters like pattern of data distribution among users, mining software,

communication resources and other hardware play a major role in deciding the feasibility of the architecture to be deployed.

The idea of agents is a new technique where agents can be hardware or software or a combination of both which does a part of work for the database users with certain level of independence. It is inbuilt with a kind of intelligence to decide on behalf of the participating users to an extent. Every intelligent agent works at local level to generate a local model from the data. All local models can be sent to one location to form a global model.

The trio Chow, Lee and Subramanian suggested the design [2] for two parties which has the use of four entities: the randomizer, the computing engine, the query front engine and the database. Each user sends queries to the query front engine which then forwards it to the randomizer. The query front engine sends coded queries to the computing engine which coordinates with the individual database to compute the query result. There is an assumption of strong privacy guarantee from each participating entity. When a query is received, the randomizer sends it to each database along with an asset of randomization parameters. It also makes sure to give them a set of derandomization parameters to the query front end. Every database computes the local query response and then presents the results in no so simple and unclear way with the help of randomisation parameters provided earlier. The computing engine now combines all so "not clear" results from all databases and ends up with one solution. Now, it is the job of the front end query engine to use de-randomisation parameters to decode the hazy result produced by the computing engine. For this model, it is understood that all the databases belong to the same schema and also there is no communication between the randomizer and the computing engine. The model has shown a linear line when comparing the number of records against the time taken to process.

In the work given by Kapoor, Poncelet, Trousset and Teisseire [3], a PriPSep Architecture, an extension of SPAM algorithm has been suggested with the components: Data Miner site DM, non colluding (NC) sites NC1 and NC2 and processing sites PS. The data miner DM is a randomly selected collaborator among the databases whose role is to perform mining. It may also be considered to choose the data miner other than any other databases involved in mining. The sites NC1 and NC2 sites collect data from each database including the data miner and perform a number of operations. The processing site PS cannot learn anything and is used by NC1 and NC2 to perform various functions. The solution asserts that the participating entities cannot learn anything above they have been entitled to during the entire process. It has the constraint that each party has the same number of records.

Baik, Bala and Rhee [4] brain waved the idea of an agent based approach so that data located at each site is analyzed to produce the results. It makes use of the decision tree approach to solve the problem of privacy preserving while mining. The work given by Vaidya and Kantarcioglu in [5] has three entities namely original site OS, no-colluding storage sites NSS and processing sites PS. PS helps the task of mining to be efficient while OS has all the data from different sites. NSS stores the user information which is shareable. This approach requires that the entire database may be transmitted fully once, which might be allowed in certain situations. The solution is such that bare minimum data gets revealed to PS and NSS. The original site OS does not learn anything and even if it does it cannot differentiate between the information of two users.

The DARM (Distributed Rules Association Mining) concept is given here [6], where the notion of three party types is taken into account. They are data providers, data consumers and master miner. It is based on service oriented architecture. The data providers is one of the parties participating in mining by making its data available. The data consumer is the user who is interested in getting the results. The master miner is the broker trusted by both data provider and data consumer. Whenever there is any need for the result, the miner performs mining with the data providers maintaining privacy and provides the result to the data consumer. It is assured that with the increase in number of attributes and records the solution remains scalable with the same performance. Alka and Ravindra[7] conveyed the approach based on SMC and is called 3LLPPVPNBC(Three layered privacy preserving vertically partitioned Naive Bayes Classifier) .The model comprises of three layers.- the input layer, the intermediate layer and finally the output layer. In the input layer all the participating users calculate their individual probabilities. The intermediate layer will calculate the total probability from all the individual probabilities. Eventually, the output layer will classify the new tuple and send this class value to all the parties. The execution time for this solution is less than the remaining solutions of its genre. In DiCrescenzo's contribution [8], the architecture comprises of clients, server and a distributed database. The data resides on each server and it is masked using a masking function whenever any query is generated by any client. This model strongly addresses issues of privacy, utility and performance: the key parameters for privacy preserving in a distributed environment .It is termed as zero knowledge collection of databases for the reason that on query from the client, each database produces each data in a masked and

randomized versions so that zero information is revealed to the client.

Alex and Ehud[9]came up with a framework which would have a miner, a calculator and the participating datasets. The miner decides what computation will be performed and the calculator performs the same without knowing which itemsets taking part. During the process, only the miner and the datasets get the mining results. The calculator merely performs the suggested calculations. The model has been constructed on the assumptions that neither the miner nor the calculator have any part of the database, the miner reports the results to the other participants ,the calculator performs all the calculations and there is no external knowledge present. The authors have presented three models viz: Horizontal, vertical and general. In case of horizontally partitioned data, the authors assume the dataset to be binary and the use of one more calculator.

The framework suggested by Karthikeswarant, Sudha , Suresh and Sultan[10] has six components namely, the original database, the modified database, the sensitive and non-sensitive rule table, the transaction table, the template table and the output table. The idea begins with converting each distinct item with a unique prime number. The information rule also is recorded in sensitive and non-sensitive rule table. The transaction rule index is constructed. The main crux lies in selecting the items and transactions to modify. Each modified class is presented as a template, which are selected one by one. The templates are stored in the template table and the selected templates in the action table. Each time a template gets selected, all the components are updated.

III. THE PROPOSED ARCHITECTURE

Three-Layered Proposed Architecture For Secured Multi-Party Computation

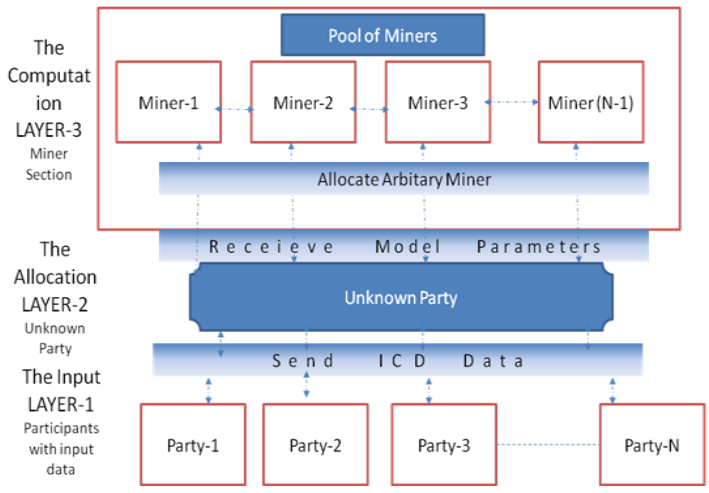

Figure 1

In this segment, we propose a three layered architecture for privacy preserving data mining in a horizontally partitioned distributed database. This solution takes into account multiple parties. We will be using Naive Bayes Classifier. The three layers are namely: 1) The input layer 2) The allocation layer 3) The computation layer. The components in the model are then participants , multiple miners N-1 and the unknown party. Before we divulge into the details of the model, we hereby present the assumptions under which the model will be functioning.

Assumptions of the model:

- Each database has data from homogeneous schema.
- Each data participant may play dual role viz. role of the miner as well as of the participant.
- A miner who participates once will not be used any more.
- The final result is sent to each participant on whose basis they can classify the new record.
- The maximum amount of release of information at any stage is that they can know the information is of any two participants of the N participants. They cannot figure out which two of the N participants.
- There is no interruption or data insecurity during the data being transferred from the participants to the miner.

The model shown in figure 1 is designed to work in the following fashion. There are N participants who take part in data mining. Each participant has the same set of attributes of data but with different set of records. The number of records available with each participant may also be different. Hence, the data are horizontally partitioned. All N participants are located in the form of a distributed database. The other role of each participant is to play as miner also. Any two participants are arbitrarily selected by the unknown party. Let us call them N1 and N2.The data of N1 and N2 is masked partially only to protect their identity. Let us call this masked data as Identity Coded Data (ICD).These data collectively is sent to the unknown party. The middle unknown party also decides upon the random miner, who will be one of among the N participants but other than N1 and N2. This unknown middle party has the job of allocating the miner to the data of N1 and N2. This miner is one of the participants as mentioned earlier but other than N1 and N2 generated by the system. Let us call this participant who is working as miner now as M1. So, now miner M1 has the model parameters from the mining results of data of participant N1 and N2. We go over the process again. We select two other participants namely N3 and N4, mask the identity data and select miner

M2. We are watchful about the point that the participant playing the role of a miner will never be playing that role again. This miner M2 will be other than N3 and N4 and obviously the miner M1. At this stage, we are ready with results of 4 participants belonging with two different participants in the role of miners. We will go ahead with the same process after combining the results of first four parties into one set of model parameters. In the end the model parameters of all the participants will be available with one of the participants. They can be made available to the rest of the participants. With the model parameters available to all the participants, any query for classification of a new record can be generated by any participant. The idea behind adding the new unknown party is that the participating entities will never come to know to whom this data was sent and the receiving miner will never know from whom this data has come.

The overall idea of the model hypothetically seems to give better results in terms of scalability. The more the number of participants, the more will be randomness and the more will be the factor of mysteriousness of the belonging of data. The model will generate more secrecy, since the process has been bifurcated at twofold selection every time, the accuracy will be promising. In the course of existing solutions, this solution can be treated at par in terms of secrecy preserving and accuracy achievable. Moreover the masking done at initial level does not required de-masking of the data.

IV. CONCLUSION

The idea of maintaining secrecy has been like whispering and sill keeping it not known to the people who observe it. Here, a novel paradigm has been proposed to achieve accuracy with secrecy. The whole idea also aims to maintain credible amount of security with not so elevated cost of communication. The transfer of data over the channel may not be so depleted but considering the amount of secrecy, it goes with the purpose. The idea proposed here will be found having a directly proportional relationship between number of participants and exactness of results with anonymity maintained. The authors intend to come up with theoretical proof for the same approach and establishing it experimentally.

REFERENCES

Journal Papers:

[1].Frawley, W., Piatetsky-Shapiro, G., Matheus, C., "Knowledge Discovery in Databases: *An Overview", AI Magazine, fall 1992, pp. 213-228, 1992*

[4]. Sung Wook Baik, Jerzy Bala, Daewoong Rhee, "An Agent Based Privacy Preserving Mining for Distributed Databases" *CIS, volume 3314 of Lecture Notes in Computer Science,page 910-915. Springer,* (*2004*)

[7]. Alka Gangrade and Ravindra Patel," Privacy Preserving Three-Layer Naïve Bayes Classifier for Vertically Partitioned Databases*", International Journal of Computer Applications © 2013 by IJCA Journal Volume 64 - Number 6 Year of Publication: 2013*

[8]. Giovanni DiCrescenzo, "Privacy architecture for distributed data mining based on zero-knowledge collections of databases".

Proceedings Papers:

[2].Sherman S.M. Chow Jie-Han Lee Lakshminarayanan Subramanian, "Two-Party Computation Model for Privacy-Preserving Queries over Distributed Databases"*_Proceedings of the Network and Distributed System Security Symposium, NDSS 2009, San Diego, California, USA, 8th February - 11th February 2009. The Internet Society 2009*

[3]. V. Kapoor,P. Poncelet,F. Trousset and M. Teisseire," Privacy Preserving Sequential Pattern Mining in Distributed Databases*", CIKM '06 Proceedings on 15th ACM international conference on Information and knowledge management. Pages 758 – 767*

[5]. Murat Kantarcioglu_ Jaideep Vaidya, "An Architecture for Privacy-preserving Mining of Client Information*", CRPIT '14 Proceedings of the IEEE international conference on Privacy, security and data mining – Volume 14 Pages 37-42.*

[6]. Omar Abdel Wahab, Moulay Omar Hachami, Arslan Zaffari, Mery Vivas, Gaby G. Dagher, DARM: A Privacy-preserving Approach for Distributed Association Rules Mining on Horizontally-partitioned Data*", IDEAS'14 Proceedings of the 18th In international Database Engineering and Applications Symposium* Pages 1-8

[9]Alex Gurevich, Ehud Gudes," Privacy preserving Data Mining Algorithms without the use of Secure Computation or Perturbation*", 10th International Database Engineering and Applications Symposium (IDEAS'06)*

[10]. D.Karthikeswarant, V.M.Sudha2 , V.M.Suresh, A.Javed Sultan4," A pattern based framework for privacy preservation through association rule mining". *IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM -2012) March 30, 31, 2012*