PPARM: Privacy Preserving Association Rule Mining Technique for Vertical Partitioning Database

[1]Ankit Jain, [2]Hitesh Kag
M.Tech Student, Assistant Professor
Computer Science Department,
MIST, Indore, India.

Abstract : With the development and penetration of data mining within different fields and disciplines, security and privacy concerns have emerged. The aim of privacy-preserving data mining is to find the right balance between maximizing analysis results and keeping the inferences that disclose private information about organizations or individuals at a minimum. In this paper, we proposed the Privacy Preserving Association Rule mining i.e. "PPARM" technique for multiparty computation of privacy-preserving data model for aggregation, cryptographic security, and association rule mining concept. In this process, the data is secured using the cryptographic techniques and for providing the more secure mining technique the server generated random keys are used. Using the proposed technique the data is mined in a similar manner as the association rule mining do, but for securing the data sensitivity the cryptographic technique is used at the client end. After mining of data, the association rules are recoverable at client end also by the similar keys as produced by the server**.**

Keywords - Data Mining, Cloud Computing, Association Rule Mining, Apriori, PPDM, PPARM.

I. INTRODUCTION

Recent advances in data mining and knowledge discovery have generated controversial impact in both scientific and technological arenas. On the one hand, data mining is capable of analyzing a vast amount of information within a minimum amount of time. On the other hand, the excessive processing power of intelligent algorithms puts the sensitive and confidential information that resides in large and distributed data stores at risk. Providing solutions to database security problems combines several techniques and mechanisms [1] [2].

1.1 Privacy Preserving

Privacy is a matter of individual perception, an infallible and universal solution to this dichotomy is infeasible. The common term of privacy in the general limits the information that is leaked by the distributed computation to be the information that can be learned from the designated output of the computation. The current state-of-the-art paradigm for privacy-preserving data analysis is differential privacy, which allows un-trusted parties to access private data through aggregate queries [3]

Privacy-preserving [4] has originated as an important concern with reference to the success of data mining. Privacy-preserving data mining (PPDM) deals with protecting the privacy of individual data or sensitive knowledge without sacrificing the utility of the data. People have become well aware of the privacy intrusions on their personal data and are very reluctant to share their sensitive information. This may lead to the inadvertent results of data mining. Within the constraints of privacy, several methods have been proposed but still, this branch of research is in its early life.

1.2 Association Rule Mining

Association rule mining has been an active research area in data mining, for which many algorithms have been developed. In data mining, association rule learning is a popular and well-accepted method for discovering interesting relations between variables in large databases. Association rules are employed today in many areas including web usage mining, intrusion detection and bioinformatics [24].

In general, the association rule is an expression of the form $X \rightarrow Y$ where X is antecedent and Y is consequent. An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent. The main aim is extracting important correlation among data items in the database. Basically, it extracts the pattern from the data based on the two measures such as minimum confidence and minimum support. Support it indicates how frequently the items appear in the database. The confidence indicates the number of times the if/then statement have been found to be true. Support it is the probability of item or item sets given transactional database [23].

1.3 Categories of Privacy Breach

A security rupture happens when private and secret data about the client is unveiled to a foe. Along these lines, protecting security of people while distributing client's gathered information is a vital research zone. The security breaks in informal communities can be sorted into three kinds [20]:

- **Identity Disclosure** - Identity Disclosure happens when a person behind a record is uncovered. This kind of rupture prompts the disclosure of data of a client and relationship he/she imparts to different people in the system.
- **Sensitive Link Disclosure** - Sensitive Link Disclosure happens when the relationship between two people are uncovered. Social exercises create this kind of data when online life administrations are used by clients.
- **Sensitive Attribute Disclosure** – Sensitive Attribute Disclosure happens when an assailant acquires the data of a touchy and classified client characteristic. Delicate qualities might be connected with a substance and connection relationship.

All these referenced protection breaks present serious dangers like stalking, extorting and burglary since clients expect security of their information from the specialist organization end. Other than that it harms the picture and notoriety of a person. There are numerous instances of unplanned divulgence of private data of clients' information that makes associations be a traditionalist in discharging the system information, for example, the AOL seek information model and assaults on Netflix information.

According to the guarantees of informal communities, there is a need to address these issues. Thusly, information should be discharged to outsiders so that guarantees the protection of the clients. In this way, information ought to be anonymized before discharging or distributing to outsiders.

2 PROPOSED WORK

This section includes the solution process which satisfies identified problem or the established goal. Therefore first the straightforward solution is provided and then the entire modeling of the system is defined.

2.1 Methodology

The proposed system can be understood by the figure 1. In this diagram, the more than one party wants to store their vertically partitioned data in a centralized data base. Additionally, the centralized database is accessed by the application for further use.

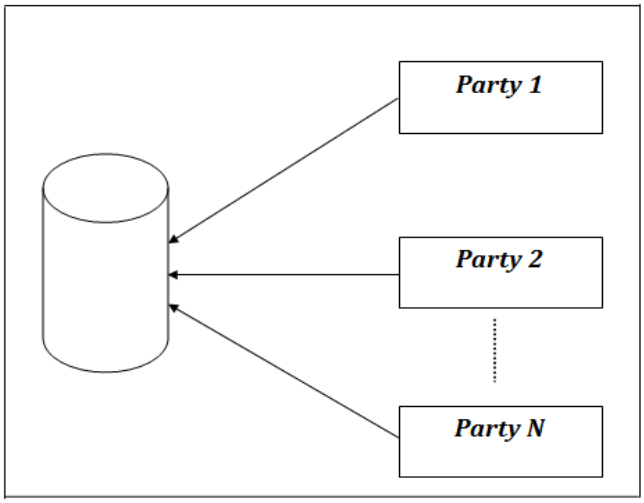

Figure 1: Database Organization

In client server model, it is more important to know how the data is transferred to the server end and which format data is delivering to the server. Let consider by an example of a list of some transaction on which basis we have to understand the scenario. If we have some itemsets

$$I = \{a, b, c, d\}$$

On the basis of this itemsets, we make list of random transaction for multiple clients. Here we consider an example of two clients transaction list i.e. $client$ 1 and $client$ 2.

Table 1: Clients Transaction List

$Client$ 1	$Client$ 2
a, b	d
b, c	a
c	a, d
a	b

Table 1 shows the client transaction on the basis of Item list for two party's databases. Each client's transaction may be same of different fir every item set. After make the transaction of the item sets, this is needs to encode on client side for data privacy preservation. Here table 2 and table 3 show the encode table transaction list which map the on the basis of transaction list for both individual client. This data will be encoded in binary format at the client end. After encoding, the data client sends it to server end and server mapping this binary data in other format of item list.

Table 2: Encoding of Client 1 Data

Client 1 Data Encode			
a	***b***	***c***	***d***
1	1	0	0
0	1	1	0
0	0	1	0
1	0	0	0

Table 3: Encoding of client 2 data

Client 2 Data Encode			
a	***b***	***c***	***d***
0	0	0	1
1	0	0	0
1	0	0	1
0	1	0	0

Here, table 2 and 3 list the encoding of the client's transactions. The item will be mapped in another encrypted form e.g. $\alpha, \beta, \gamma, \delta$ for item seta, b, c, d. Hence, newly mapped data will be shown in table 4. This mapped data list with the help of generated above-encoded data. Here we list the combine list the client1 and client 2 data.

Table 4: Mapping Table

Client 1				*Client 2*				
α	***β***	***γ***	***δ***	***α***	***β***	***γ***	***δ***	*Mapping*
1	1	0	0	0	0	0	1	α, β, δ
0	1	1	0	1	0	0	0	β, γ, α
0	0	1	0	1	0	0	1	γ, α, δ
1	0	0	0	0	1	0	0	α, β

Finally, this database transfer to the server side in each transmission and on the server side this data are combined. For privacy preserving data mining, the combined database resides on the server side which is shown in mapping table 4. Whenever the client receives the generated encrypted rules it will decrypt this by using the reverse mapping table and generated client key. The user decrypts only that part of the rule which he wants to access information. The proposed system ensures the strong privacy of the user confidential information using mining association rule.

This data is transmitted in binary format to the server side. Hence, the transmission of data has happened through 'memory stream'. Memory stream is a region between server and client where server and client read and write data respectively. Firstly clients write the data into memory stream in byte format and server read this byte data and converts it into its original form. The internal process of client and server- based of socket programming. Therefore, step by step process is run and generates intermediate output. The server will not do anything until the client requests it to the server, as soon as the server starts the process of Request Accept. Hence process is run continuously and privacy-preserving data mining approach is successfully applied using the proposed methodology.

The key issue in this given model all the parties have some private, confidential information which is not disclose-able to others. But how the decisions are made is required to distribute. Thus between the server and different parties, the following operations are needed to utilize for effective data processing and knowledge discovery as given in figure 2. In this figure, we show the step-by-step process of entire work methodology. The process has happened between client and server. Client and server are the two basic entities by request transmit for a different purpose. Initially, clients who want to privacy preserving data mining technique need to prepare a connection between client and server. Therefore, firstly client or party sends the request to the server for establishing a connection. After that server accepts the request and response of enabling connection in terms of acknowledgment.

After establishing a successful connection between both, the server generates the random key for linked parties of a client. And this key sends to the client for further use. After receiving this random key of client-side and use this for transforming data into ciphertext. Thereafter, clients upload this data to the a server for utilization by server and combine database. Therefore, the server calls the apriori algorithm for processing the data and extracts the association rule for the form data. Finally, the recovered rules from the server are provided to the end client. The end client applies the key again to the received data for recovering the part of the information which is owned by the self. The process shows the data privacy preservation for multiple clients. For data encryption and decryption we have used AES algorithm. Apriori algorithm has been implemented for both scenario i.e. proposed and base method and results generated on the basis of time complexity, memory utilization and the number of rule generated by algorithms.

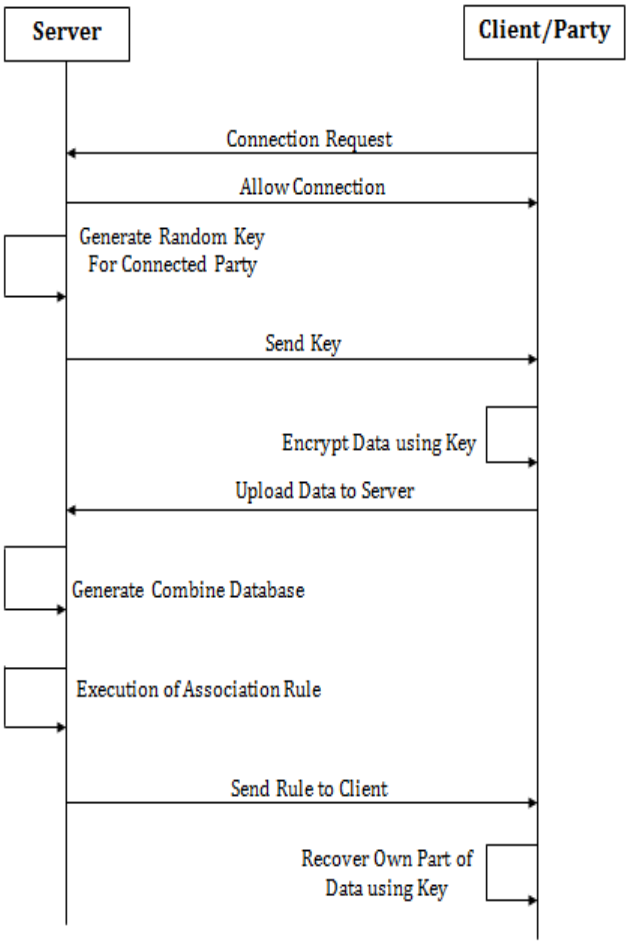

Figure 2: System Processing Architecture

2.2 Proposed Algorithm

This section provides the algorithm in table 5 steps for the security and privacy-preserving data mining technique. Let the N number of parties $C = \{C_1, C_2, \dots, C_n\}$ are want to associate their data to find the common decision using different provided attributes. The data from different clients can be defined in the following manner.

$$D = \{d_{C^1}, d_{C^2}, d_{C^3} \dots\dots, d_{C^n}\} \quad ----- eq\ (1)$$

In the above eq (1) show the aggregated data from the different participating clients and d_{C^n} is the data provided by the n^{th} user of the system. Using the provided symbols the following function is used to prepare the privacy-preserving model.

Table 5: Proposed PPARM Technique Algorithms

Input: Number of Clients (C), Clients Data d_{C^n} **Output:** Rules (R)
Process: **1:** Client Send Connection Request to Server **2:** connection = client. connect (server) **3:** Count number of active connection for single process **4:** client send data to server for each active client **5:** Server read data of individual client **6:** newAttributeList = mapping (merge, AES) **7:**newTransection = generateTransection (merge, newAttributeList) **8:**Rules = Apriori. getrules (newTransection, min support, min confidenc **9:** Return R

III. RESULT ANALYSIS

This section provides details about the performance evaluation of the proposed privacy preserving association rule mining technique. Therefore the different factors and their observations are listed in this section.

3.1 Memory Usage

Memory usages term indicate here the amount of main memory utilized for data analysis using the proposed privacy preserving association rule mining context. The memory usages of the system are also known as space the complexity of algorithms. The comparative memory usages of the proposed and traditional algorithm are given using figure 3

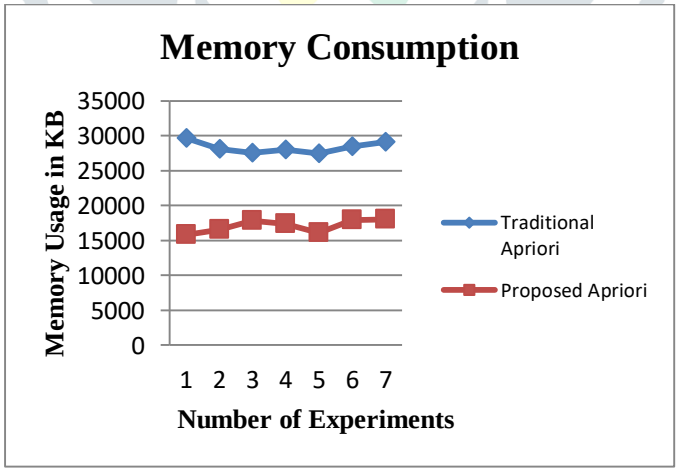

Figure 3: Memory Consumption

The memory usages of the implemented privacy-preserving algorithms are reported using the figure 3. In this diagram, the X axis contains different experiments conducted with the proposed PPDM system and Y-axis shows the amount of main memory consumed. The measurement of memory is given here in terms of KB (kilobytes). In order to indicate the performance of algorithms red and blue lines used. Where the red line shows the performance of proposed PPDM based association rule mining technique and the traditional algorithm is demonstrated using a blue line. According to the observed experimental performance, the performance of both algorithms is fluctuating but it remains within the range. Thus it depends upon the size of data, so when the size of data increases their memory consumption is also increases. Additionally, if the number of candidate set generation is large then the memory requirement is higher.

3.2 Time Consumption

The time consumed for the generation of privacy-preserving Apriori algorithm based association rules the difference of time between algorithm initialization and finishing the algorithm process is computed. In this presented work the total difference of time is denoted as time consumption or time complexity of algorithm. The time consumption or complexity of proposed privacy preserving association rule mining algorithms is given using figure 4. In this diagram, X-axis shows experiments conducted with the system and the Y-axis shows the amount of time consumed. The measurement of time is given here as far as seconds. As indicated by the given execution proposed PPDM based affiliation rule mining calculation expends less measure of time when contrasted with the conventional PPDM based framework. But the time consumption is increased as the amount of data for association rule mining is increases. Based on the experimental observations the proposed PPDM based Apriori algorithm is time preserving system and reduces the amount of time consumption and running time of the algorithm.

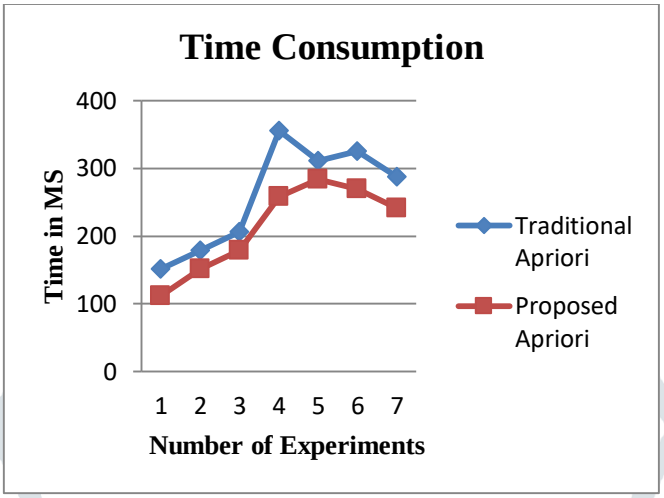

Figure 4: Time Consumption

3.3 Transaction vs. Rules

To know which algorithm is working more effectively an additional parameter is computed. that parameter shows the amount of rule generated for the fixed amount of transaction sets input.

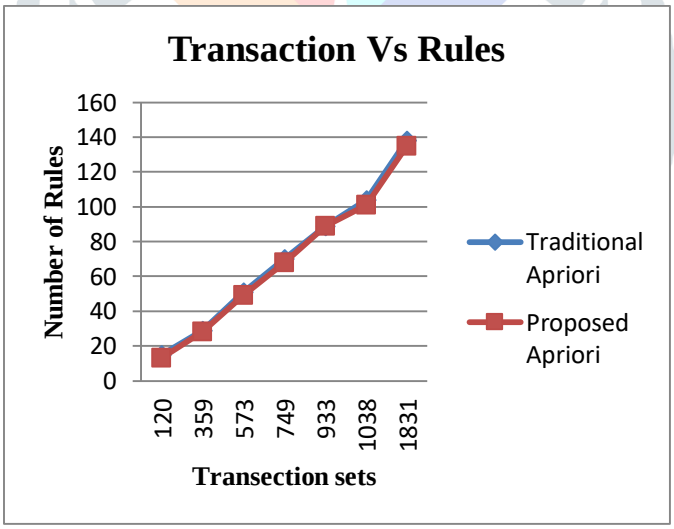

Figure 5: Transactions vs. Rule

Therefore proposed PPDM based association rule mining algorithm and traditional technique both are compared on the same parameter. The observed experimental performances are provided in figure 5 for both the algorithms. In this diagram, the input transactions are given in X-axis with their amount and generated rules are associated in Y-axis. Here the orange line demonstrates the performance of proposed PPDM based association rule mining technique and performance of traditional technique is given using the blue line. According to experimental observations both the algorithm, most of the time generate similar numbers of rules. but sometimes the proposed algorithm generates less number of rules as compared to the traditional approach.

IV. CONCLUSION

The tremendous growth in the IT field and the problems addressed during the storage of the huge data is the major problem. Data mining aims to discover secret information from large database although secret data is kept safe when data is allowed to access by a single person. The proposed work is intended to develop a privacy-preserving technique for extracting the association rules form the transaction database. Therefore a secure technique is developed using the cryptographic data manipulation and reorganization of the association rules. To simulate the proposed technique there are a client system is developed. This client system first connects to the centralized server. After a successful connection on the client, the request server transmits the secure key randomly generated. For the entire session, we follow the process for two parties. This key is works as a session key and only

one time used with the server. The obtained key by the server, the client manipulate their own data and upload to the server. The uploaded data is combined with the other parties' data and encoded in binary form and finally, the Apriori algorithm is implemented to extract the rules. These rules are also deliverable to the end client and the client can recover their own part of data using the obtained key.

REFERENCES

[1] Agrawal, Dakshi, and Charu C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms", Proceedings of the twentieth ACM SIGMOD- SIGACT-SIGART symposium on Principles of database systems, ACM, 2001.

[2] Oliveira, Stanley RM, and Osmar R. Zaiane, "Privacy preserving frequent itemset mining", Proceedings of the IEEE international conference on Privacy, security and data mining-Volume 14. Australian Computer Society, Inc., 2002.

[3] Deepa Tiwari and Raj Gaurang Tiwari, "A Survey on Privacy Preserving Data Mining Techniques", IOSR Journal of Computer Engineering (IOSR-JCE), Volume 17, Issue 5, Ver. III (Sep. – Oct. 2015), PP. 60-64

[4] Vishal Ravindra Redekar and Dr. K. N. Honwadkar, "Privacy-Preserving Mining of Association Rules in Cloud", International Journal of Science and Research (IJSR), Volume 3 Issue 11, November 2014

[5] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "The KDD process for extracting useful knowledge from volumes of data." Communications of the ACM 39.11 (1996): 27-34.

[6] Ahmed HajYasien, "Preserving Privacy In Association Rule Mining", Ph D., thesis, Griffith University, June 2007.