

# Enhanced Security Based CART Algorithm for Vertically Partitioned Database in Multi-Party Environment

Priyanka Kaurav<sup>1</sup> Dr. C.S. Satsangi<sup>2</sup>

Medi-Caps Institute of Technology and Management, Indore<sup>1,2</sup>

Department of Information Technology<sup>1,2</sup>

Email id- [kaurav.priyanka31@gmail.com](mailto:kaurav.priyanka31@gmail.com)<sup>1</sup>; [cssatsangi@gmail.com](mailto:cssatsangi@gmail.com)<sup>2</sup>

**Abstract**—Distributed applications are increases rapidly these applications can serve a number of clients. Thus data distribution and management is key area of concern. For distribution of data, mining techniques are used for enhancing the server capabilities. Data storage and accessing need to convert data in horizontally or vertically partitioned. In this paper vertically partitioned data is considered. When the vertically partitioned data is arrived on server, it combined entire data and stored in a centralized storage. The data accessed by end clients need sharing between all parties. Thus the privacy and security is also a key area in such storage. To improve privacy and security in multiparty data access environment a cryptographic technique is proposed and implemented. The implementation of the proposed privacy preserving data mining technique is performed in JAVA technology. Finally the performance of the proposed classification technique is evaluated in terms of time and space complexity, accuracy and error rate. The results demonstrate effective performance and security during data access.

**Keywords**— Privacy preserving data mining, Security, Centralized data, vertically partitioned data.

## I. INTRODUCTION

The internet based applications and their uses are increase therefore to satisfying need of increasing demands efficient and secure manner of data storage and access are required. In addition of that there are some applications are also available that are consuming more than one database same time. In these applications storage is available in remote place and connectivity between all databases concurrently required. In these conditions data mining approaches are applied to databases for efficiently data access. Thus data model is responsible for data distribution.

Basically, data mining is a tool for analysing data from data source. During multiparty data access privacy and security is a major concern in distributed computing environment. For example, two parties, A and B, each having a private data denoted by a and b, want to collaboratively conduct classification on datasets. Because they are concerned about privacy, not any party disclose its dataset to other. Thus required make following assumptions on the data sets a and b,

and pre-processing doesnot require one to send data to other party:

- Dataset a and b have similar instances of records.
- Data set a and b can have different number of attributes.
- Both parties share the class labels of all records and names of attributes.

The key requirements of the presented work is given in this section the next section includes the proposed work and their working.

## II. ROUTING PROTOCOLS

The presented data model first include distributed data base which is divided into the classes and all the data distributed over multiple classes. There are different accessing parties that accesses data concurrently as available on server. In this context when data is accessed from individual parties then required securing and privacy on data. To manage this cryptographic technique is used. Only the data is recovered when the actual data owner making a request using private key. The working of proposed secure and privacy preserving data model are described as given in Figure 1:

1. **Server initialization:** To design a server that can serve multiple clients a multi-threaded program is designed that running on a fixed Port number. Server uses this port to listen client's request.
2. **Client initialization:** When client program initialized a request on server port number is made. If server is running then client program get connected to server and can communicate.
3. **Connection:** The required system is a network based modelling thus client-server architecture is prepared.
4. **Key generation:** There multiple clients are connected with server therefore a single key can create privacy issues thus N number of random keys

is generated according to connected parties. The key generation is performed on server and that is valid for a single session.

In data mining and machine learning applications the amount of input samples are correctly recognized is known as the accuracy of the classifier or algorithm. The accuracy can be estimated using the given formula.

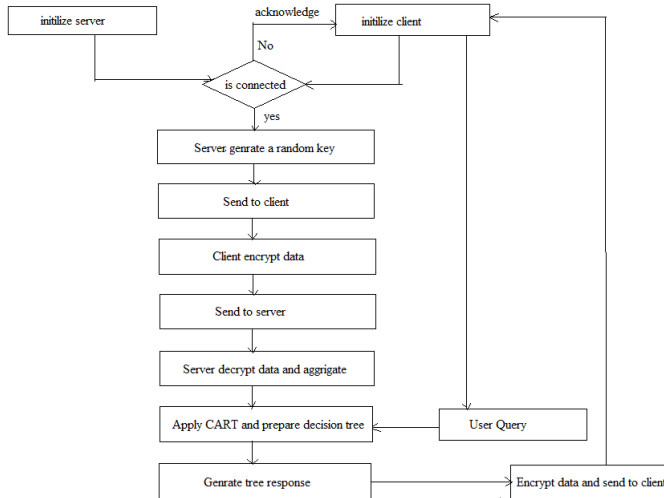


Figure 1: Proposed System

5. **Client data encryption:** Client having a part of entire data which vertically partitioned thus client usage the obtained key to encrypt data and send data to server.
6. **Server data decryption:** Server has a set of keys for each client. The encrypted data from the connected client is sent by the server. After recovery of data server aggregate it on a data table. The data comes from different users are decrypted and aggregated to the server.
7. **CART algorithm:** CART algorithm is applied to centralized data on server. Using this decision tree data is queried is satisfied.
8. **Client query:** The trained decision tree on server can be used to perform classification, thus a user query can be applied on decision tree through client.
9. **Server data encryption:** Server again encrypts decision tree rules and sends classification of data to client.
10. **Client data decryption:** The attribute wise data encryption to server allows a user to recover data only if user has the part of data which is encrypted before.

### III. RESULTS ANALYSIS

The performance of the algorithm is evaluated in order to reflect the performance and security achieved using proposed CART algorithm when it works on secure environment.

#### Accuracy

$$accuracy = \frac{\text{total correctly identified samples}}{\text{total samples to classify}} \times 100$$

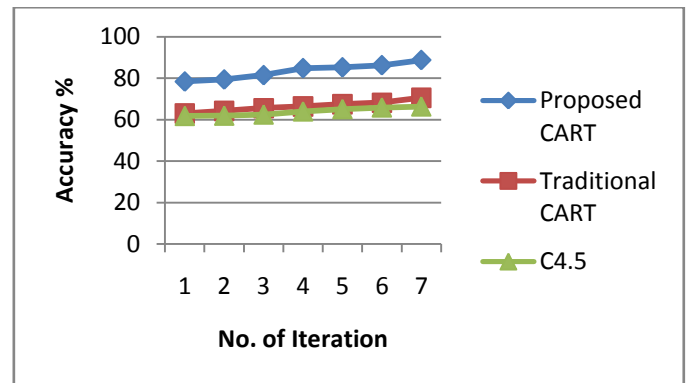


Figure 2: Accuracy

The performance of the implemented algorithms namely proposed CART algorithm, traditional CART and C4.5 decision tree over encrypted is given using Figure 2. In this diagram the X axis shows the number of different iteration performed with the algorithm is given. During iteration the amount of data among all the parties are increased and then their performance is evaluated. On the other hand the percentage accuracy of algorithm is given in Y axis. According to the given results the performance of the proposed algorithm is increases as the amount of Data during learning is increases as compared to the both traditional algorithms namely C4.5 and CART.

#### Error Rate

The error rate of the algorithm demonstrates the amount of data which is not correctly identified during classification. The error rate of an algorithm can be evaluated using the below given formula.

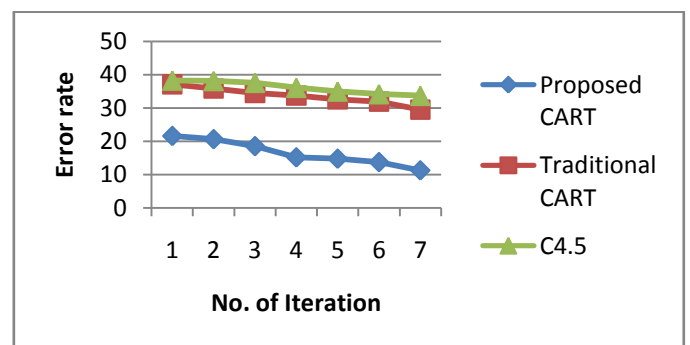


Figure 3: Error Rate

$$error\ rate = \frac{\text{total incorrectly identified samples}}{\text{total samples as input}} \times 100$$

Or

$$\text{error rate} = 100 - \text{accuracy}$$

The given figure 3 shows the comparative percentage error rate observed during classification with the implemented algorithms. The X axis simulates the different iteration performed with the system and the Y axis shows the percentage error rate. According to the obtained results the performance of the algorithm improved by decreasing the error rate of classification additionally only those parties are able to view the outcomes which are providing the correct key input.

### Memory Used

The amount of main memory required to successfully execute the algorithm is known as the memory consumption or space complexity of the algorithm. The amount of memory consumed during different iteration is reported using Figure 4.

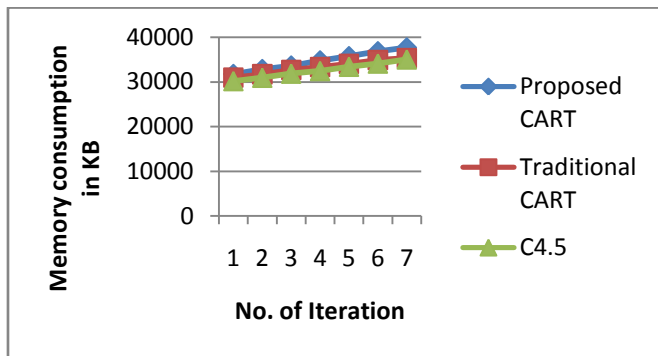


Figure 4: Memory Usage

The amount of memory consumed during different iterations is reported using Figure 4: In order to show the performance of algorithm X axis of the diagram shows the different iteration performed and the Y axis shows the amount of main memory consumed during iteration in terms of KB (kilobytes). According to obtained comparative results with traditional CART and C4.5 algorithms as given in figure 4. The performance of the proposed algorithm in terms of memory usage is higher as compared to both algorithms.

### Training Time

The amount of time required to perform training using the algorithms is known as the training time. The comparative result contains the c4.5 algorithm, traditional CART algorithm and proposed secure CART algorithm. In this Figure 5 the amount of time consumed is given using Y axis in terms of milliseconds and the X axis shows the different iteration performed. According to the obtained results the amount of time for training is increases as the amount of data for learning is increases.

Additionally the comparative results show the performance of the proposed algorithm is week in terms of training time with respect to the traditional algorithms CART and C4.5.

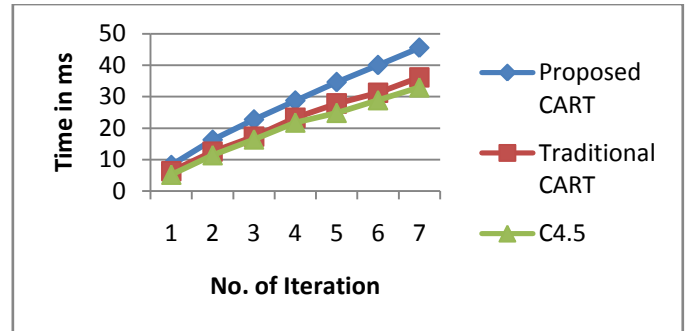


Figure 5: Training Time

### Search Time

The amount of time required to traverse the tree on server side and produce the class label for input attributes is known as the search time of the algorithm. The comparative search time of proposed CART algorithm and traditionally implemented CART and C4.5 algorithms is demonstrated using Figure 6.

The search time of CART algorithm during different iteration is reported using figure 5.5. In this diagram the Y axis shows the amount of time consumed in terms of milliseconds and the X axis shows the different Iteration performed with the algorithm. After estimating the performance of the algorithm that is observed the average time of computing the class labels is not affected on the size of data.

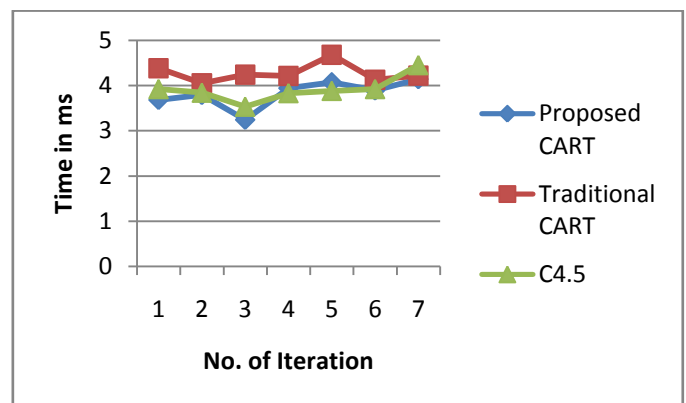


Figure 6: Search Time

## IV. CONCLUSION AND FUTURE WORK

Data mining is a tool for analysing huge amount of data using the intelligent algorithms. Therefore data mining algorithms compute significant patterns from the input data and preserve them for utilizing them in future data pattern evaluation. This process in data mining techniques is called the training of algorithms. After training of algorithms when the new data

arrived on the algorithm then these data are identified on the basis of extracted patterns from data. In this presented work the data mining algorithms and their classification functioning is investigated. Additionally that is also investigated how the data is secured in a distributed environment when a number of parties are concurrently communicating with the database servers.

During investigation there are various kinds of data models are observed. These data models not only provide the exact learning of patterns it also reduces the effort of data evaluation and query processing time. But the architecture of data mining is a prepared for centralized database architecture thus whole data which is required to classify is combined in single place. Then the training operation over data is performed. The training process results a decision tree which is used during data access different parties. Therefore security and privacy management is required to adopt over centralized data.

Thus securing the privacy and confidentiality on the data during data mining a new kind of technique is required to design. Therefore a multiparty data submission and accessing in secured manner is simulated using encryption techniques. For data submission the private key encryption technique is utilized on the other hand for providing data mining algorithm the CART algorithm is implemented with the secure data access mechanism. After designing the secure algorithm that is implemented using the JAVA technology and their performance is estimated. According to the obtained results the performance of the proposed secure and privacy preserving technique not disclosing the attributes during data access among the participating parties and also providing much efficient results. In order to justify the presented work the performance of proposed algorithm is compared with the traditional CART algorithm and C4.5. The comparative performance summary is given in the table below.

Table 6.1 performance summary

S. No.	Parameters	Proposed CART	Traditional CART	C4.5
1	Accuracy	High	Low	Low
2	Error rate	Low	High	High
3	Memory usage	High	Low	Low
4	Training time	High	Low	Low
5	Search time	Low	High	High

According to the obtained results the performance of the proposed technique is found adoptable due to improved classification rate over the encrypted data additionally the low error rate and stable prediction time. On the other hand the performance of algorithm is affected in terms of memory consumption and training time. That is increases as compared to the traditional CART and the C4.5 due to additional computational overheads of the encryption and decryption during data storage and access with multiple parties. Thus the proposed algorithm is suitable when the high accurate results are required with security, but that is not much effective when the need of low resource consumption is required.

## V. FUTURE WORK

The proposed algorithm is adoptable and provides the security as well as the efficient performance during learning and classification. The presented model is not yet implemented on the real time data thus in near future the given model is implemented to secure the real time transactions and other kind of horizontally partitioned data. Additionally that is also required to improve the performance of the proposed system in terms of the memory utilization and the training time.

## REFERENCES

- [1] Jayanti Dansana, Debadutta Dey, Raghvendra Kumar, "A Novel Approach: CART Algorithm for Vertically Partitioned Database in Multi-Party Environment", Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013)
- [2] Distributed Databases, Learning Objectives, chapter 12 and 13 [http://wps.prenhall.com/wps/media/objects/3310/3390076/hoffer\\_ch13.pdf](http://wps.prenhall.com/wps/media/objects/3310/3390076/hoffer_ch13.pdf)
- [3] Mahesh Manchanda, Dr. Neena Gupta, "Make Web Page Instant: By Integrating Web-Cache and Web-Prefetching", Conference on Advances in Communication and Control Systems 2013 (CAC2S 2013)
- [4] Mavridis I., Pangalos G., "Security Issues in a Mobile Computing Paradigm", Informatics Laboratory, Computers Division, Faculty of Technology Aristotle University of Thessaloniki Thessaloniki 540 06, Greece
- [5] Lukasz A. Kurgan and Petrusilek, "A survey of Knowledge Discovery and Data Mining process models", The Knowledge Engineering Review, Vol. 21:1, 1-24. 2006, Cambridge University Press
- [6] Data Mining - Cluster Analysis, [http://www.tutorialspoint.com/data\\_mining/dm\\_cluster\\_analysis.htm](http://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm)
- [7] "Data Mining - Classification & Prediction Introduction", [http://www.idc-online.com/technical\\_references/pdfs/data\\_communications/Data\\_Mining\\_Classification\\_Prediction.pdf](http://www.idc-online.com/technical_references/pdfs/data_communications/Data_Mining_Classification_Prediction.pdf)
- [8] Arya Apoorva, Mr. Parikshit Singla, "A Review of Information Sharing Through Shared Key Cryptography", international Journal of Research in Engineering Technology and Management ISSN 2347 - 7539
- [9] Ramesh S, K N Haribhat, R Murali, "On Linear Complexity of Binary Sequences Generated Using Matrix Recurrence Relation Defined Over Z4", International Journal of Distributed and Parallel Systems (IJ DPS) Vol.1, No.2, November 2010
- [10] Guillermo Navarro-Arribas, "User k-anonymity for privacy preserving data mining of query logs", 2011 Elsevier Ltd. All rights reserved. .