An Enhanced CART Algorithm for Preserving Privacy of Distributed Data and Provide Access Control over Tree Data

Monika Gupta
Department of Computer Science & Engineering,
Acropolis Institute of Technology and Research
Indore,India
guptamonica12@gmail.com

Abstract **—** Now in these days the utilization of distributed applications are increases rapidly because these applications are serve more than one client at a time. In the use of distributed database data distribution and management is a key area of attraction. Because of privacy of private data organizations are unwilling to participate for data mining due to the data leakage. So it is required to collect data from different parties in a secured way. This paper represents how CART algorithm can be used for multi parties in vertically partitioned environment. In order to solve the privacy and security issues the proposed model incorporates the server side random key generation and key distribution. Finally the performance of proposed classification technique is evaluated in terms of memory consumption, training time, search time, accuracy and there error rate.

Keywords **—***Privacy Preserving Data Mining, Classification, CART,*

*****

I. INTRODUCTION

Data mining [1] often called knowledge discovery of data is a useful process of extracting useful data and hidden predictive information from large databases. Data mining takes relational database, data warehouse, transactional database, flat files, data streams and world wide web as input.

Classification rule mining [2] is one of the most popular algorithms used for data mining. Classification is a data mining algorithm that assigns data or transactions in a group or classes. The goal of classification is to accurately forecast the class attribute for each transaction of the data.

Decision tree classification is one of the most popular classification technique used for classifying the class attributes. It is called supervised learning because first decision tree is build on training dataset then test attributes are introduced for classification of test data. In decision tree internal nodes are called test and leaf nodes are called class labels and the arrows shows the path between tests and leads to class label. Different decision tree uses different attribute selection measure for best splitting node. Most popular decision trees ID3 [2], C4.5 [3], CART [4] that uses different attribute selection measure information gain, gain ratio and gini index respectively.

Advancement of technology has naturally evolved distributed database. Distributed database [5] is a database in which database are partitioned and stored in different systems which logically belongs to the same system. Database is distributed either vertically or horizontally.

The proposed work is simulated using vertically partitioned data. From different sources, and applied CART algorithm for decision tree making. For security scheme key generation and distribution is used.

II. PROPOSED WORK

The key objective of the proposed work is to find a suitable method of classification which works efficiently with the vertically partitioned data and also provides the secure access of data among multi-party access environment. The entire work is sub divided in the following modules

.

Study of different privacy preserving data mining approaches: In this phase the different techniques which providing the security during the transactional data base is studied.

Find efficient and accurate classification technique: In this phase different decision tree algorithms are studied which are high efficient and accurate for data mining. Thus the CART algorithm is selected among ID3,C4.5 and CART algorithms.

Design a model for the effect of privacy preserving data mining: In this phase using the selected decision tree a new privacy preserving technique is developed and implemented using JAVA environment.

Performance study of the proposed classification scheme: After implementation the performance of designed data model is evaluated for justifying the performance of the proposed privacy preserving data model.

A. Problem Domain **:**

Classification schemes are the supervised learning process of data mining, where the attributes and class labels are exist in order to learn about the data. The classification is used to find the pattern in data to recognize the organization of the data, in order to recognize the similar data patterns in the data which is provided in the real time. The main issues in distributed computing environment where a single data model is available for more than one client, and they access the required data from same source. To manage the privacy and access control on the data, when the data model is shared between more than one clients. There for to handle the data integrity and privacy management a new kind of data model is required to develop.

B. Solution Domain **:**

The presented model first includes the distributed database which is divided into the clusters and all the data distributed over multiple places. In addition of that there are three accessing parties which are consuming data concurrently as the data updated on server. Each having a set of attributes with a class label, all parties from the system need to update data on server and then required to process using a centralized algorithm which process data and produce the decision tree. In this context when the data is accessed from the individual parties then required securing and preserving the privacy on data. Therefore a cryptographic technique is utilized which is used to encrypt and decrypt the data when the data accessed on the client side. Only the data is recovered when the actual data owner making a request using private key.

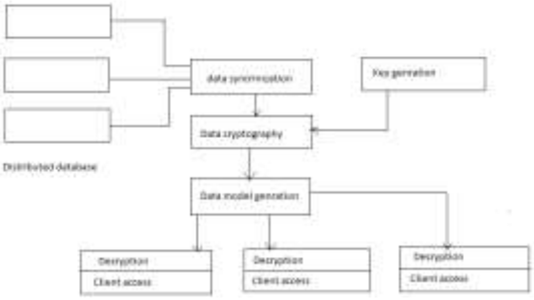

Fig. 1 System Architecture

C. Simulation Architecture:

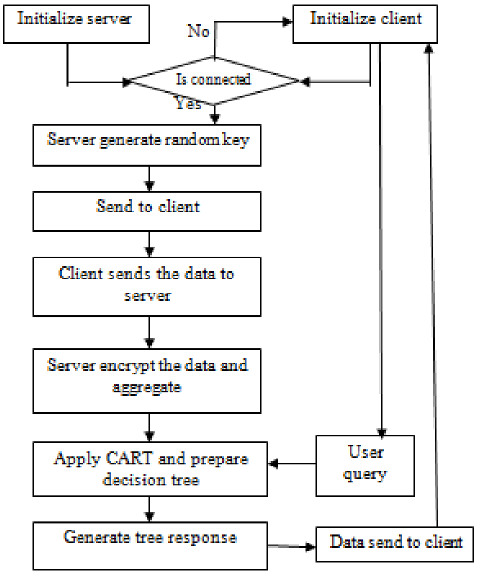
Initialize server
No
Initialize client
Is connected
Yes
Server generate random key
Send to client
Client sends the data to server
Server encrypt the data and aggregate
Apply CART and prepare decision tree
User query
Generate tree response
Data send to client

Fig 2 Simulation Architecture

The given figure 2 shows the simulation architecture of proposed system in this context the entire system is divided into a set of modules. The entire modules are described as:

1) *Server Initialization:* in order to design a server that serves more than one client required to design a multi-threaded server. Thus a server program is designed that works on a fixed port number. When the server initiated then the server can listen the clients request using this fixed port number.

2) *Client Initialization:* when client program is initialized then a request on the same application port number is made. If server program is initialized then the client program is connected through the server port and data communication can initialized.

3) *Connection:* the given system is a network based data mining technique thus the client server architecture is prepared. The communication among client and server is performed using TCP ports. Thus for connection the socket programming is utilized. In this context the socket needs two parameters first target IP address and a fixed communication port. On the basis of this the connectivity among client and server is prepared.

4) *Key Generation:* there are a number of clients can connect with the server therefore a single key can create problem for the privacy of data, thus for data distribution and processing the N number of random keys are generated. The key generation process is taken place on the server end and this key is only valid for current session. After key generation for a client that is distributed to all the connected clients.

5) *Client Data send:* client always having a part of data for the entire dataset which vertically partitioned. Client sends the data to the server end.

6) *Server Data Encryption:* server having a set of keys for each client thus server gather the data from the connected client and encrypts the data using client's key. After encryption of the data server aggregate the data on an existing data table. Additionally all the data which is comes from the different users are encrypted and aggregated to the server end's table.

7) *CART Algorithm:* A classical CART algorithm is applied to the aggregated client's data which is centralized process of data mining. After applying CART algorithm server generates a decision tree which is trained on aggregated data. Thus the decision tree is ready to resolve the user generated query.

8) *Client Query:* When the decision tree is prepared on the server side the user can send the query to server for finding the classification label, thus a user query can apply on decision tree through the client end.

9) *Server Data Encryption:* Server again encrypts the decision tree generated rules and sends to the client end. In order to encrypt the rules the attribute wise data which belongs to the target user is encrypted.

10) *Client Data Decryption:* The attribute wise data encryption from the server side allows a user to recover only those part of data which is encrypted by the client's key.

III. PROPOSED ALGORITHM

The given cryptographic data model can be summarized using steps, the entire process of data aggregation and their distribution is given as:

```
1.  initialize server
2.  initialize clients
3.  if( is_server_connected)
4.  server generate a random key
5.  send to connected client
6.  end if
7.  client send the data to server
8.  server collect all the streamed data
9.  Encrypt data and aggregate into a single data unit
10. apply CART algorithm on data
11. prepare the decision tree
12. wait for client request
13. when a client query arrived
14. find the class label and traversal
15. send to client
16. client decrypt data using allotted key
17. recover only the part of data decrypted through the
    allotted key
18. end
```

IV. RESULT ANALYSIS

This section provides the understanding about the outcomes and the analysis of the performance of the implemented algorithms.

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

$$accuracy = \frac{total\ correctly\ identified\ samples}{total\ samples\ to\ classify} X100$$

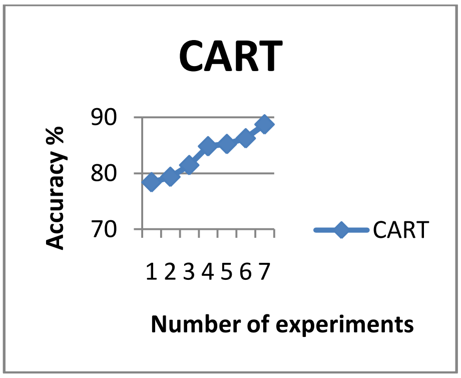

Fig 3 Accuracy

The performance of CART algorithm in secure environment is evaluated and reported in fig 3, in this diagram the X axis shows the number of different experiments performed with the algorithm is given. During experimentation the amount of data among all the parties are increased and then their performance is evaluated. On the other hand accuracy of algorithms given in Y axis according to the given result the performance of algorithm is increases as the amount of data during learning increases.

2. *Error rate:* The error rate of 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.

$$error\ rate = \frac{total\ incorrectly\ identified\ samples}{total\ samples\ as\ input} X100$$

Or

$$error\ rate = 100 - accuracy$$

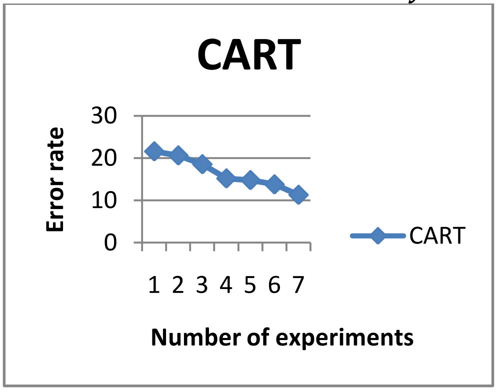

Fig 4 Error Rate

The given fig 4 shows the percentage error rate observed during classification, the X axis shows the different experiments performed with the system and the Y axis shows the percentage error rate. According to the obtained results the performance of 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.

3. *Memory used:* The amount of main memory required to successfully execute the algorithm is known as the memory consumption..

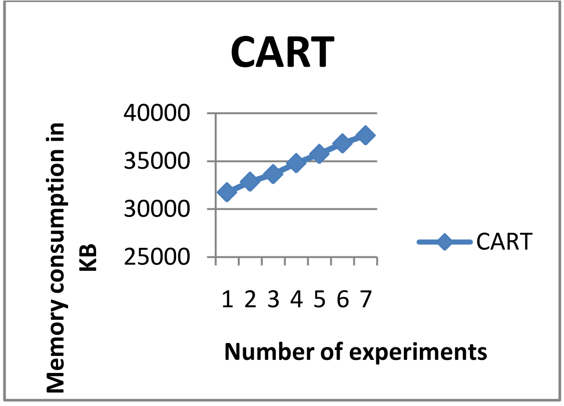

Fig 5 Memory Usage

The amount of memory consumed during different experiments is reported using fig 5. In order to show the performance of algorithm X axis shows the different experiments performed and Y axis shows the amount of main memory consumed during experimentations. According to the results the performance of the algorithm in terms of memory usage is decreases as the amount of data in experiments are increases.

4. *Training Time:* The amount of time required to perform training using the algorithm is known as the training time.

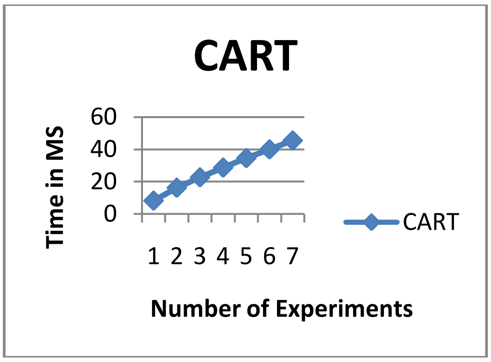

Fig 6 Training Time

The CART algorithm's training time for secure environment is given using fig 6. Here the amount of time consumed is given using Y axis in terms of milliseconds and the X axis shows the different experiments with increasing dataset size performed. According to the obtained results the amount of time for training is increases as the amount of data for learning is increases.

5. 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 algorithm.

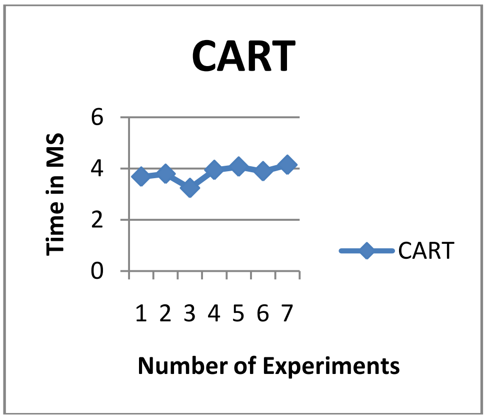

Fig 7 Search Time

The search time of CART algorithm during different experiments is reported using fig 7. In this the Y axis shows the amount of time consumed in terms of milliseconds and the X axis shows the different experiment performed with the algorithm. After estimating the performance of the algorithm it is observed that the average time of computing the class labels is not affected on the size of data.

V. CONCLUSION

Data mining is a tool for analysing huge amount of data using the intelligent algorithms. There for 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 are 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 working is investigated.

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. But the architecture of data mining is a centralized architecture thus whole data is combined in single place and then the computation is performed on that place. And during data access different parties those are preserve the data on the centralized data base are accessed accordingly. But due to access of data sometime sensitive and private data of a party can be given to another party and the issues of security and privacy can arises. Thus securing the privacy and confidentiality on the data during data mining a new kind of technique is required to investigate and design. Therefore in this presented work a multiparty data submission and accessing in secured manner is simulated. 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 after applying security on algorithm the performance is summarized on the table given 1

S. No.	**Parameters**	**Remark**
1	Accuracy	The accuracy of the implemented data model increases as the amount of data for learning increases. the increment of accuracy is about 2-5%
2	Error rate	The error rate of the algorithm decreases as the learning set of algorithm increases the reduction on error rate about 2-5%
3	Memory usage	The memory consumption of the system is increases as the amount of data increases for processing
4	Training time	The training time is also depends on the size of data

		provided as input
5	Search time	The search time of algorithm is not much affected due to increasing size of data

Table 1 performance summary

VI. 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.

VII. REFERENCES

[1] J. Han and M. Kamber, Data Mining: Concept and Techniques, 2nd ed., Morgan Kaufmann, New York, Elsevier, 2009.

[2] S.Archana et al,"Survey of Classification Technique in Data Mining", International Journal of Computer Science and Mobile Apllications, vol.2 no. 2, February- 2014.

[3] Tejaswini pawar, Prof Snehal Kamalapur "A Survey on Privacy Preserving Decision Tree Classifier", International Journal of Engineering Research and Application (IJERA) ISSN:2248-9622 www.ijera.com vol.2, no.6, pp.843-847,November- December 2012

[4] V.S. Verykios, E. Bertino, I.N. Fovino, L.P. Provenza, Y. Saygin, Y.Theodoridis, State-of-the-art in privacy preserving data mining, ACM SIGMOD, 2004.

[5] T. Pawar , Prof S. Kamlapur "Decision Tree Classifier for Privacy preservation", International Journal of Emerging Technologies in Computational and Applied Sciences, pp. 309-314,Dec. 12- Feb.13.

[6] J.Dansana, D.Dey, R. Kumar, "A Novel Approach: CART Algorithm for Vertically Partitioned Database in Multi-Party Environment", Proceedings of 2013 IEEE Conference on Information and Communication Technologies,2013.

[7] Distributed Databases, Learning Objectives, chapter 12 and 13 http://wps.prenhall.com/wps/media/objects/3310/3390076/hoffer_ch13.pdf.

[8] M. Manchanda, Dr.N.Gupta, "Make Web Page Instant: By Integrating Web-Cache and Web-Prefetching", Conference on Advances in Communication and Control Systems 2013.

[9] I. Mavridis, G. Pangalos., "Security Issues in a Mobile Computing Paradigm", Informatics Laboratory, Computers Division, Faculty of Technology Aristotle University of Thessaloniki Thessaloniki 540 06, Greece

[10] A. Lukasz . Kurgan and Petrmusilek, "A survey of Knowledge Discovery and Data Mining process models", The Knowledge Engineering Review, Cambridge University Press,vol. 21, pp 1–24, 2006,

[11] Data Mining - Cluster Analysis, http://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm.

[12] "Data Mining - Classification & Prediction Introduction", http://www.idc-online.com/technical_references/pdfs/data_communications/Data_Mining_Classification_Prediction.pdf

[13] A. Apoorva, Mr. P. Singla, "A Review of Information Sharing Through Shared Key Cryptography", international Journal of Research in Engineering Technology and Management ISSN 2347 – 7539

[14] S.Ramesh, K N Haribhat, R Murali, "On Linear Complexity of Binary SequencesGenerated Using Matrix Recurrence RelationDefined Over Z4",International Journal of Distributed and Parallel Systems (IJDPS) vol.1, no.2, November 2010

[15] G. Navarro-Arribas, V. Torra, A. Erola, J. Castellà-Roca, "User k-anonymity for privacy preserving data mining of query logs", Contents lists available at ScienceDirect Information Processing and Management, Elsevier Ltd. All rights reserved,2011

[16] W. Jue, Y. Lei, P. Lingxi, and L. Feng, "Privacy-Preserving Data Mining Algorithm Quantum Ant Colony Optimization", Applied Mathematics & Information Sciences An International Journal, no. 3, 1129-1135, 2013

[17] F. Giannotti, V.S. Lakshmanan, A. Monreale, D. Pedreschi, and H.Wang, "Privacy-preserving Mining of Association Rules from Outsourced Transaction Databases", IEEE Transactions on Knowledge And Data Engineering vol.7 no.3, 2013

[18] K. Patel, H. Patel, P. Patel, "Privacy Preserving in Data stream classification using different proposed Perturbation Methods", IJEDR vol 2, no. 2,ISSN: 2321-9939,2014

[19] N. Pelekis, A.G. Divanis, M. Vodas, A.Plemenos, D. Kopanaki, Y.Theodoridis, "Private-HERMES: A Benchmark Framework for Privacy Preserving Mobility Data Querying and Mining Methods", EDBT'12, March 26–30, 2012, Berlin, Germany. Copyright 2012 ACM 978-1-4503-0790-1/12/03

[20] D. Saahas Reddy, V. Uma Rani, Dr. M. SreenivasaRao, "Dynamic Non-Cooperative Computation For Privacy Preserving Data Analysis", International Journal of Computer Science and Mobile Computing, vol. 3, no. 10, pg.337 – 343 October 2014.

[21] A. Badanidiyuru, B. Mirzasoleiman, A. Karbasi, A. Krause, "Streaming Sub-modular Maximization: Massive Data Summarization on the Fly", KDD'14, New York, NY, USA, Copyright is held by the owner/author(s). Publication rights licensed to ACM, ACM