ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

Privacy Preserving Random Decision Trees over Randomly Partitioned Dataset

Jintu Ann John, Neethu Maria John

M. Tech Student, Dept. of CSE, Mangalam College of Engineering, Ettumanoor, Kottayam, Kerala, India

Associate Professor, Dept. of CSE, Mangalam College of Engineering, Ettumanoor, Kottayam, Kerala, India

ABSTRACT: Nowadays, privacy preservation in data mining has become an important issue. Privacy preservation means protecting the sensitive information on data that are collected from different sources.The data collected have to collaborate together by maintaining the information on the data. Distributed data mining is concerned with the computation of data that is distributed among multiple participants. Privacy preserving distributed data mining allows the cooperative computation of data without parties revealing their individual data. Though there are some organizations which individualy collect a lot of data on their own, often large correlated data is collected over many sites.It is possible that several organizations collect similar data about different people.Thus in horizontal partitioning of dataset, parties collect data for different entities, but have data for all of the attributes.On the other hand different organizations may collect different information about the same set of people. Thus, in vertically partitioned data, all parties collect data for the same set of entities. In all of these cases, mining on the local data is simply not as accurate as mining on the global data. It may lead to inaccurate, even improper results. Privacy concerns restrict the free flow of information. Organizations do not want to reveal their private databases for various legal and commercial reasons. Neither do individuals want their data to be revealed to parties other than those they give permission to. Both horizontal and vertical partitioning of datasets are not so accurate. The time depends on the slowest machine so time delay occurs. The server sends data to multiple clients after doing dataset partitioning. In this paper the dataset partitioning is done by, Random partitioning of dataset. This method is more accurate than the existing horizontal and vertical partitioning of dataset.Random Partitioning of dataset provides more security and load balancing.

KEYWORDS: Data mining; privacy preserving data mining; horizontal and vertical partitioning; random partitioning; Random Decision Tree; homomorphic encryption; security.

I. INTRODUCTION

Data mining is a practice of examining large databases in order to look for hidden patterns in a group of data that can be used to predict future behaviour. True data mining techniques doesn't just change the presentation, but actually discovers previously unknown relationships among the data[15]. So basically data mining is “Knowledge Discovery in databases”. Privacy and security factors may limit the sharing or centralization of data. Privacy-preserving data mining has emerged as an effective method to solve this problem [2]. Privacy preservation means, protecting specific individual values, breaking the link between values and the individual they apply to, protecting source, etc.

The server sends data to multiple clients. The distributed data have to be protected. The data should be send only after the partitioning of dataset. The dataset partitioning is done in randomly. In Random split the dataset is partitioned equally and is given to each node according to its capacity. So no time delay takes place. But in horizontal and vertical partitioning the output depends on the slowest machines.So time delay occurs. After the partitioning of dataset construct Random decision tree.The Random Decision Tree [1] approach is more accurate and efficient than existing cryptographic techniques.

Random decision tree algorithm constructs multiple decision trees randomly.In Random partitioning, dataset is partitioned equally and sends data to different clients. Each client has different capability.Some may be slow whereas some may be faster.So according to the capability the data sends.The remaining data stored on a pool.

The RDTs algorithm builds multiple RDTs. One important aspect of RDTs is that the structure of a random tree is constructed completely independent of the training data. The RDT algorithm can be broken into two stages, training

DOI: 10.15680/IJIRCCE.2015. 0308072

and classification. The training phase consists of building the trees and populating the nodes.It is assumed that the number of attributes is known based on the training data set. The depth of each tree is decided based on a heuristic—Fan et al. [3] show that when the depth of the tree is equal to half of the total number of features present in the data, the most diversity is achieved, preserving the advantage of random modeling.

The process for generating a tree is as follows. First, start with a list of features or attributes from the data set. Generate a tree by randomly choosing one of the features without using any training data. The tree stops growing once the height limit is reached. Then, use the training data to update the statistics of each node. Note that only the leaf nodes need to record the number of examples of different classes that are classified through the nodes in the tree. The training data is scanned exactly once to update the statistics in multiple random trees.

Then apply Homomorphic encryption to the leaf nodes of RDT.This is done for providing more privacy and security.Homomorphic encryption is a form of encryption that allows computations to be carried out on ciphertext, thus generating an encrypted result ,which,when decrypted matches the result of operations performed on the plain text.
The rest of the paper is organized as follows: in Section 2,we describe the related works. In Section 3, we present the proposed method. In Section 4, reports the algorithms used. In section 5 reports, extensive experimental results to support the proposed method . Finally, in Section 6, we summarize the present study and draw some conclusions.

II. Related work

In [1] ,the privacy preserving data mining is done by generating Random Decision Tree framework.Here,the dataset partitioning is done by vertical and horizontal partitioning of dataset.In both cases,according to the slowest machine the time delay takes place.So when a new slowest machine came the time increases. Privacy and security concerns can prevent sharing of data, derailing data mining projects. Distributed knowledge discovery, if done correctly, can alleviate this problem. Here, introduced a generalized privacy preserving variant of the ID3 algorithm for vertically partitioned data [6] distributed over two or more parties. A completely random decision tree algorithm [2] that achieves much higher accuracy than the single best hypothesis and is comparable to boosted or bagged multiple best hypotheses. The advantage of multiple random trees is its training efficiency as well as minimal memory requirement. Data is distributed in various sites that need to be mined in a secure manner without revealing anything except the results of mining. Privacy-preserving horizontal distributed classification techniques[8] where multiple sites collaborate and broadcast the mining results. However in the process, no information about either the data maintained in the sites or data obtained during computation is divulged. Two protocols are presented to construct a Privacy Preserving Naïve Bayesian classifier using the Pailler's homomorphic encryption techniques.

The problem of association rule mining where transactions are distributed across sources are explained in [7]. Each site holds some attributes of each transaction, and the sites wish to collaborate to identify globally valid association rules.However, the sites must not reveal individual transaction data. Here, presents a two-party algorithm for efficiently discovering frequent itemsets with minimum support levels, without either site revealing individual transaction values.Privacy and security concerns can prevent sharing of data, derailing data mining projects. Distributed knowledge discovery, if done correctly, can alleviate this problem. The key is to obtain valid results, while providing guarantees on the disclosure of data. Here, present a method for *k*-means clustering when different sites contain different attributes for a common set of entities. Each site learns the cluster of each entity, but learns nothing about the attributes at other sites[4].

Advances in computer networking and database technologies have enabled the collection and storage of large quantities of data, also the freedom and transparency of information flow on the Internet has heightened concerns of privacy. Nowadays the scenario of one centralized database that maintains all the data is difficult to achieve due to different reasons including physical, geographical restrictions and size of the data itself. The data is normally maintained by more than one organization, each of which aims at keeping its information stored in the databases privately, thus, privacy-preserving techniques and protocols are designed to perform data mining on distributed data when privacy is highly concerned. Cluster analysis [14] is a frequently used data mining task which aims at decomposing or partitioning a usually multivariate data set into groups such that the data objects in one group are most similar to each other. It focuses on arbitrarily partitioned data [15] which is a generalization of horizontally partitioned data and vertically partitioned data along with Shamir's Secret Sharing Schemes which was designed with the goal of achieving complete privacy for secure computation and communication between different parties**.**

.

III. PROPOSED ALGORITHM

A. Random Partitioning of dataset

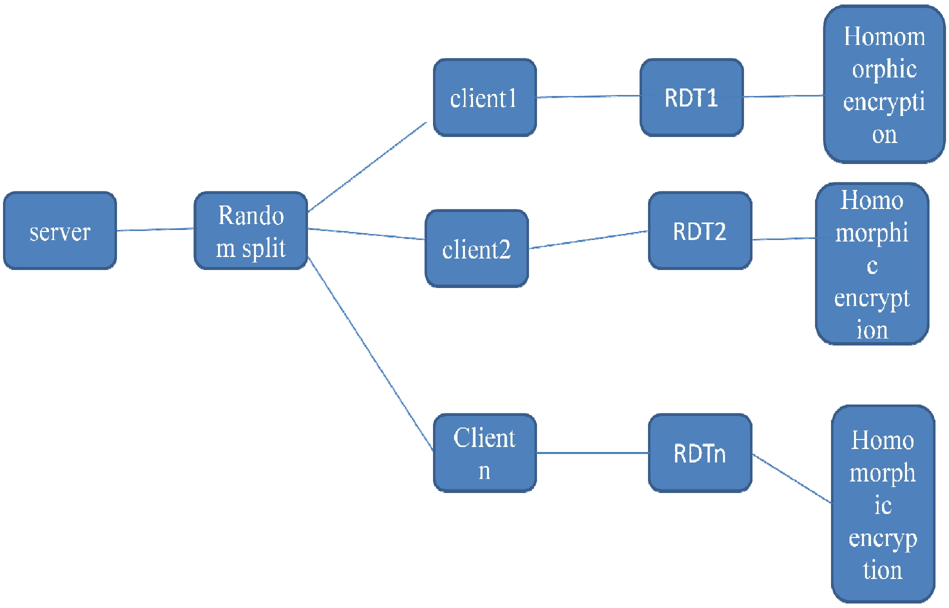

Fig 1: Framework for Random Split

In Random split, the dataset is partitioned equally and according to the capacity of each machine the datas are send.The server sends the data to the clients after the dataset partitioning.Heterogeneous nodes are used here.So there will be no time delay and provides high security and load balancing.

Consider n clients and each one's capacity is C1,C2,C3,…..Cn.Let,total capacity being C.Consider a pool of data where remaining datas are storing.Initially,the dataset partitioned equally and is given to each clients.The dataset size is 100.The number of clients is 5.The dataset partitioned equally, 20 each.Each clients capacity is 10,5,20,5,10.The server sends data to each clients.The remaining data stored on the pool.The client having highest capacity will complete its processing first and takes the remaining data from pool.So load balancing occurs and there will be no time delay.

IV. PSEUDO CODE

A. Random Split

```
Step 1:Partition the data equally.
Step 2:Server send the data to clients.
Step 3:The remaining data are stored on the pool.
Step 4:The client having highest capacity will take the remaining data and completes the processing.
```

B. RDT construction

Input: Transaction set T partitioned randomly between different partities,P_1,….P_n.
Output: The total number of random trees created.

```
Step 1: while Every party does not agree to all of the random trees do
Step 2: Each party generates its random trees
Step 3: Every tree is communicated  to all of the parties.
Step 4: end while
Step 5: for each tree $T_k$ do
Step 6: Each party $P_K$ encrypts the leaf nodes in $T_k$ using threshold additively homomorphic encryption system and
        sends to all other parties.
Step 7:  Every party then multiply corresponding encrypted elements they receive for each leaf node to get encrypted
        global value for that node.
Step 8: end for
```

V. SIMULATION RESULTS

The proposed method is more accurate than the existing method. In vertical and horizontal partitioning the dataset is partitioned vertically and horizontally. But in the proposed system the dataset is partitioned equally in a random manner. Random split is more faster than horizontal and vertical splits.When a new slowest machine arrives,no time delay occurs because the fastest machine will complete the remaining data.

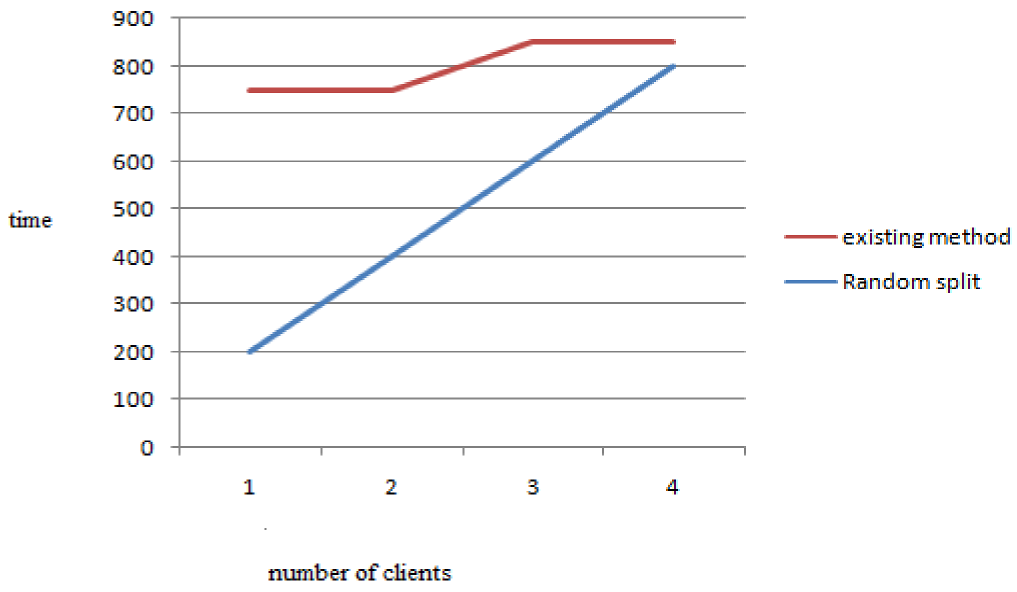

Fig 2: Performance evaluation

In horizontal and vertical partitioning the time delay depends on the slowest machine.Suppose if a new slowest machine arrives the time for complete the process depends on that slowest machine.Therefore,the graph for existing method as stepwise manner.But in the proposed method there is no time delay.The dataset is partitioned equally and is given to all clients.The remaining data are stored in a pool.The fastest machine after completing the process will take the remaining data and completes the task.So the graph plotted as straight line.

VI. CONCLUSION

In this paper, we have presented a novel random partition of dataset for high dimensional data. The algorithm involves 1) Partition the data equally, 2) Server sends the data to clients and 3) The client having highest capacity will take the remaining data and completes the processing.The dataset partitioning based on this approach is more efficient and accurate than the existing vertical and horizontal partitioning.Then construct Random Decision Tree.Since RDTs can be used to generate equivalent, accurate and better models with much smaller cost, we have proposed distributed privacy-preserving RDTs. Our approach leverages the fact that randomness in structure can provide strong privacy with less computation.Then apply homomorphic encryption to provide more security. The RDT algorithm scales linearly with data set size, and requires significantly less time than existing cryptographic approaches.Thus,the proposed method provides more security and load balancing.

REFERENCES

1. Jaideep Vaidya, Basit Shafiq, Wei Fan, Danish Mehmood, and David Lorenzi,"A Random Decision Tree Framework for Privacy Preserving Data Mining", IEEE Transactions On Dependable And Secure Computing, Vol. 11, No. 5, September/October 2014
2. J. Vaidya, C. Clifton, and M. Zhu, "Privacy-Preserving Data Mining", Advances in Information Security first ed., vol. 19,Springer-Verlag, 2005.
3. W. Fan, H. Wang, P.S. Yu, and S. Ma, "Is Random Model Better?On Its Accuracy and Efficiency," Proc. Third IEEE Int'l Conf. Data Mining (ICDM '03), pp. 51-58, 2003
4. G. Jagannathan and R.Wright, "Privacy-preserving distributed k-means clustering over Vertically partitioned data",In 11th KDD, pages 593–599, 2005.
5. W. Du and Z. Zhan, "Building Decision Tree Classifier on Private Data," Proc. IEEE Int'l Conf. Data Mining Workshop on Privacy, Security,and Data Mining, pp. 1-8, Dec. 2002.
6. J. Vaidya, C. Clifton, M. Kantarcioglu, and A.S. Patterson,"Privacy-Preserving Decision Trees over Vertically Partitioned Data," ACM Trans. Knowledge Discovery from Data, vol. 2, no. 3,pp. 1-27, 2008.
7. M. Kantarcioglu and C. Clifton, "Privacy-Preserving Distributed Mining of Association Rules on Vertically Partitioned Data," IEEE Trans. Knowledge and Data Eng., vol. 16, no. 9, pp. 1026-1037,Sept. 2004.
8. G. Jagannathan and R.N. Wright, "Privacy-Preserving Distributed Distributed Mining of Association Rules over Horizonatlly Partitioned Data," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining,pp. 593-599, Aug. 2005.
9. H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the Privacy Preserving Properties of Random Data Perturbation Techniques," Proc. Third IEEE Int'l Conf. Data Mining (ICDM '03),Nov. 2003.
10. J. Souza, "Feature Selection with a General Hybrid Algorithm,"PhD dissertation, Univ. of Ottawa, 2004.
11. L. Yu and H. Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy," J. Machine Learning Research, vol. 10,no. 5, pp. 1205-1224, 2004.
12. Sha, X. Qiu, and A. Zhou, "Feature Selection Based on a New Dependency Measure," Proc. Fifth Int'l Conf. Fuzzy Systems andKnowledge Discovery, vol. 1, pp. 266-270, 2008.
13. M. Dash and H. Liu, "Consistency-Based Search in Feature Selection," Artificial Intelligence, vol. 151, nos. 1/2, pp. 155-176, 2003.
14. X. Lin, C. Clifton, and M. Zhu, "Privacy Preserving Clustering with Distributed EM Mixture Modeling," J. Knowledge and Information Systems, vol. 8, no. 1, pp. 68-81, July 2005.
15. [15] G. Jagannathan and R,N Wright, "Privacy Preserving Using Distributed K-means Clustering for Arbitrarily Partitioned Data
16. Proc.ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining,pp. 593-599, Aug. 2005.

BIOGRAPHY

Jintu Ann John, M-Tech student in the Department of Computer Science and Engineering at Mangalam college of Engineering,Mahatma Gandhi University, Kerala, India. She received B-Tech from Saintgits college of Engineering,Mahatma Gandhi University,Kerala,India.Her area of interest is Data Mining.

Neethu Maria John, Associate Professor in the Department of Computer Science and Engineering, Mahatma Gandhi University,Mangalam college of Engineering, Ettumanoor, Kerala, India.