PRIVACY PRESERVING RANDOM DECISION TREE OVER PARTITION DATA

Miss. Pratiksha D Kale [1], Miss. Archana R Panhalkar[2]

[1] *Student, M.E, Information Technology, AVCOE, Sangamner, Maharashtra, India*

[2] *Assistant Professor, Information Technology, AVCOE, Sangamner, Maharashtra, India*

ABSTRACT

In recent years distributed data is present everywhere in current information driven approach. For the various sources of data, the inherent challenge is how to decide to merge effectively across organizational border line while maximizing the benefit of information collection. Privacy-preserving knowledge discovery techniques must be developed because local data is used suboptimal utility. Previous privacy-preserving cryptography work is too slow to be used for huge data sets to face difficulties for large data. The past work on Random Decision Trees (RDT) introduce that to possible to generate identical and accurate models with smaller cost .In this paper to utilize the fact that RDTs can particularly fit into a distributed architecture such as fully and parallel , and originate some protocols to execute RDTs that authorize distributed knowledge discovery for privacy-preserving.

Keywords : - *Distributed data , RDT, data mining, and classification*

1. INTRODUCTION:

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. When implemented on high performance client/server or parallel processing computers, data mining tools can analyze massive databases to deliver answers to questions such as, "Which clients are most likely to respond to my next promotional mailing, and why?"

This white system provides an introduction to the basic technologies of data mining. Examples of profitable applications illustrate its relevance to today's business environment as well as a basic description of how data warehouse architectures can evolve to deliver the value of data mining to end users.

A decision tree is a graph that uses a branching method to illustrate every possible outcome of a decision. A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm.

Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.

Random decision tree are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set.

2. LITERATURE SURVEY:

R. Agrawal and R. Srikant [2] studied and then examine the technical possibility of privacy-preserving data mining.

D. Agrawal and C.C. Aggarwal [4] survey the privacy-preserving data mining algorithms for configuration and quantification. They assume that the maximization algorithm which is proves maximum probability evaluation for original distribution of data.

Notwithstanding, H. Kargupta et. al [5] indicated a some of the difficulties in the data privacy preserving. It indicates the specific conditions to break the privacy security.

The distributed sources for cryptographic methods were applied in data mining to development of decision trees by Lindell and Pinkas [3].

Jagannathan et al. [9] proposed the method to create private RDT classifier from the concentrated data set. Hence, the data is distributed, it cannot be used.

Wang et al. focused on the transaction identifiers between sites [7]; while this does not display attribute values, parties exchanges the value one by one the way is downwards to the tree, then one site to said to two particulars have the value for same attributes.

Du and Zhan [8] can present a method to create vertically partitioned data of decision tree classifier by using privacy-preserving.

3. PROBLEM DEFINATION

The Problem is defined as Follows:

1) In distributed classification the basic problem is to instruct a classifier from the distributed data and then categorize new instance.

2) The main objective is the distributed data is used to create a decision tree classifier.

4. PROPOSED SYSTEM

In this paper to develop methods to securely construct RDTs for both horizontally and vertically partitioned data sets. To implement the proposed protocols and analyze the computation and communication cost, and security. And also compare the performance of the proposed protocols with the existing ID3-based protocols.

5. CONTRIBUTION:

The main contribution is to realize that RDTs can provide good security with very high efficiency. In this system to address the issue of privacy preserving data mining. Specifically, to consider a scenario in which two parties (Admin and Provider) owning confidential databases wish to run a data mining algorithm on the databases, without revealing any unnecessary information. Our work is motivated by the need to both protect privileged information and enable its use for research or other purposes. For that we use public-key cryptosystem, to enhance the securing data we use onion layer encryption. Means we use multi encryption techniques for different types of data. At end we have two types of data i.e. Yes & No Outputs. We use two encryptions for both different outcomes. We would take Damgard-Jurik encryption for one type and Advanced Encryption Standard for another one. For this enhancement we increase the security **of data.**

6. IMPLEMENTATION DETAILS:

A. System Overview

User module can perform following operation.

1) Registration

In this each user register his/her user details for using files. Only registered user can able to login and proceed on data.

2) Data Upload

In this user upload a weather data set in to the database and it is to be protected from unauthorized user.

I] Provider module:

In this module the provider can register first .After doing registration then the provider can login. After login then the provider can insert the weather data into the data base. After inserting data into the database.

II] Admin module:

In the admin module admin can login first. After login an admin can view this data which is inserted by providers. Then an admin can create a random decision tree and share the key to providers to see the vertical partition data.

7. ALGORITHM

Algorithm for creation random decision tree:

Require: Tranaction set T1 partitioned vertically between sites S1,…..,Sk**.**

Require: Si holds mi attributes

Require: s class values,d1,….,dp, with Sk holding the class attributes.

Require: n, the number of random trees to build

```
All parties together compute $m=\sum I\ mi$ using the secure some protocol[10].
Depth← m/2{The depth of random trees.}
for i=1…n {Build the ith tree}do
level←1
$nodeId_i$←Build Tree{level ,depth}
end for
```

8. RESULT ANALYSIS

RDT utilize for various data processing tasks such as, ranking, regression, classification and multiple classifications. Privacy protective RDT uses each randomization and cryptographic technique which offers information privacy for a few Decision trees primarily based learning task. The proposed system tends to study the technical feasibleness of realizing privacy-preserving data processing. RDTs are often familiar to generate equivalent, correct and typically higher models with a lot of smaller cost; the proposed system tends to area unit exploitation distributed privacy - preserving RDTs. Our approach controls the actual fact that randomness in structure will offer well-built privacy with less computation. The distributed RDT algorithms and implementation presented in this paper are a significant step forward in creating usable, distributed, privacy-preserving, data mining algorithms. The running time of the algorithms, is comparatively much faster than the existing implementations, and is usable on everyday computing hardware. As compared to the standard, non-privacy-preserving version, the accuracy of the privacy-preserving solution is exactly the same, though the computational overhead is significant.

9. EXPERIMENTAL SETUP

In this the System consists of technology like Advance JAVA, HTML, CSS and JavaScript. For back end SQL Server is used. Also, Hence before experimental set up Software like Eclipse, Tomcat is projected to be installed on server. User should have basic windows family, good browser to view the results.

10 CONCLUSIONS:

The privacy and security suggestions are assumes that when to manage distributed data that is partitioned either on vertically or horizontally for multiple sites. This system wants to develop general solutions for an arbitrarily partitioned data.

11. ACKNOWLEDGEMENT

It gives me an immense pleasure to express my honest and heartiest gratitude towards my guide Prof. Archana R. Panhalkar and Prof. Bharat S. Borkar for their guidance, encouragement, moral support and affection during the course of my work. I am especially appreciative of their willingness to listen and guide me to find the best solution, regardless of the challenge. This work is also the result of the blessing guidance and support of my parents and family members and friends. I am also thankful to all who have contributed indirectly and materially
in words and deeds for completion of this work.

12. REFERENCES

[1] C.Rajesh, S.Hari, U.Selvi "On the Privacy Preserving Properties of Random Data Perturbation Techniques," Proc. Third IEEE Int'l Conf. Data Mining (ICDM '03), Nov. 2003 .

[2] R. Agrawal and R. Srikant, "Privacy-Preserving Data Mining," Proc. ACM SIGMOD Conf. Management of Data, pp. 439-450, May 2000.

[3] Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining," J. Cryptology, vol. 15, no. 3, pp. 177-206, 2002.

[4] D. Agrawal and C.C. Aggarwal, "On the Design and Quantification of Privacy Preserving Data Mining Algorithms," Proc. 20th ACM SIGACT-SIGMOD-SIGART Symp. Principles of Database Systems, pp. 247-255, May 2001.

[5] 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.

[6] G. Jaideep Vaidya, Senior Member, IEEE, Basit Shafiq, Member, IEEE, Wei Fan, Member, IEEE, Danish Mehmood, And David Lorenzi "A Random Decision Tree Framework for Privacy-Preserving Data Mining," Proc. IEEE Transactions On Dependable And Secure Computing, Vol. 11, No. 5, pp. 399-411, September/October 2014

[7] 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.

[8] M. Kantarcioglu and C. Clifton, "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data,"Proc. ACM SIGMOD Workshop Research Issues on Data Mining and Knowledge Discovery (DMKD '02), pp. 24-31, June 2002.

[9] G. Jagannathan, K. Pillaipakkamnatt, and R.N. Wright, “A Practical Differentially Private Random Decision Tree Classifier,” Proc. IEEE Intl Conf. Data Mining Workshops (ICDMW), pp. 114-121, 2009.

[10] Z. Huang, W. Du, and B. Chen, “Deriving Private Information from Randomized Data,” Proc. ACM SIGMOD Int’l Conf. Management of Data, June 2005.

[11]R. Wright and Z. Yang, “Privacy-Preserving Bayesian Network Structure Computation on Distributed Heterogeneous,” Proc. 10th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining, Aug. 2004.