Survey Paper on Privacy Preserving Random Decision Tree over Partition Data

Miss. Pratiksha D Kale[1] Miss. Archana R Panhalkar[2]
[1,2]Student [4]Assistant Professor
[1,2]Department of Information Technology
[1,2]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.

Key words: Distributed data, RDT, data mining, and classification

I. INTRODUCTION

Data mining sometimes called as data or knowledge discovery is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both.

Data mining is the process of collecting, searching through, and analyzing a large scale data in database, as to discover pattern or relationship. Data mining software is one of a number of analytical tools for analyzing data .It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. It is a subfield of computer science. The main purpose of data mining process is to select data from a different data set and modify it into an understand\able structure for further use. Data mining tools predict future trends and behavior, allowing businesses to make proactive, knowledge-driven decisions. Data mining is a process used by company to turn raw data into useful information. Data mining is the analysis step of the KDD.

Random decision tree algorithm constructs multiple decision trees randomly. In RDT same code can be used in different data mining tasks such as: regression, classification, ranking and multiple classifications. Fan et. al. can originate the RDT. When constructing each tree, the algorithm picks a "remaining" feature randomly at each node expansion without any purity function check (such as information gain, gini index, etc.). A categorical feature (such as gender) is considered "remaining" if the same categorical feature has not been chosen previously in a particular decision path starting from the root of tree to the current node. Once a categorical feature is chosen, it is useless to pick it again on the same decision path because every example in the same path will have the same value (either male or female). However, a continuous feature (such as income) can be chosen more than once in the same decision path. Each time the continuous feature is chosen, a random threshold is selected.

RDT performs better than various models regarding the speed of computation. RDT have two advantages that is accuracy and performance. The multiple (or m) iso-depth RDTs can be developed by RDTs algorithm. To build fully independent training data is one of the main feature of RDT. The RDT algorithm can be distributed into two sections, such as classification and training. The training step encompasses building the trees and update statistics. The training sets of data examine the quantity of attributes. In existing system, the Bayes optimal classifier (BOC), has efficient implementation of RDT for non-parametric density and high order statistics.

II. LITERATURE REVIEW

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.

III. CONCLUSION

The privacy and security suggestions are assumes that when to manage distributed data that is partitioned either on vertically or horizontally for multiple sites.

REFERENCES

[1] List and number all bibliographical references in 9- [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.

