Privacy Preserving Data Mining Using Random Decision Tree Over Partition Data: Survey

Nashwan Adnan Othman[1*], *Mustafa Zuhaer Nayef* Al-Dabagh[1]

[1]College of Science, Knowledge University, Erbil, Kurdistan Region, Iraq

Abstract. The development of data mining with data protection and data utility can manage distributed data efficiently. This paper revisits the concepts and techniques of privacy-preserving Random Decision Tree (RDT). In existing systems, cryptography-based techniques are effective at managing distributed information. Privacy-preserving RDT handles distributed information efficiently. Privacy-preserving RDT gives better precision data mining while preserving information and reducing the calculation time. This paper deals with this headway in privacy-preserving data mining technology utilizing emphasized approach of RDT. RDT gives preferable productivity and information privacy than cryptographic technique. Various data mining tasks utilize RDT, like classification, relapse, ranking, and different classifications. Privacy-preserving RDT utilizes both randomization and the cryptographic method, giving information privacy for some decision tree-based learning tasks; this is an effective technique for data mining with privacy-preserving distributed information. Thus, in horizontal partitioning of the dataset, parties gather information for various entities but have data for all attributes. On the other hand, various associations may gather different data about a similar set of people. Thus, in vertically partitioned data, all parties gather data for the same collection of items. In all of these cases, both horizontal and vertical partitioning of datasets is somewhat inaccurate.

1 Introduction

Data Mining finds exciting data patterns, and insights from extensive databases. There are two phases in privacy-preserving data mining, the first is information collection, and the second is information publishing. In information collection, the holder stores information, which the proprietor gathers. In publishing, information could discharge to the receiver via the holder and the recipient mines published secured information. RDT algorithm builds numerous decision trees randomly. One essential structural part of RDTs is that they are entirely independent of the training information. Therefore, let us consider the two phases of the RDTs algorithm, training and classification. The training phase comprises building the trees and populating the nodes, which assumes the number of attributes is known based on the training dataset [1-3].

* Corresponding author: nashwan.adnan92@gmail.com

The procedure for producing a tree is as per the following. First, consider a features listing or note the attributes from the data set. Then, define a tree by randomly selecting a feature without utilizing any training information. The tree completes its development by reaching its uppermost limit. Then, it uses training data to refresh the statistics of each node. Note that only the leaf nodes need to record the number of cases of various classes grouped through the nodes in the tree. Finally, it scans the training information to upgrade the statistics in different random trees [1, 4].

RDT gives a better answer for the distributed data mining in concepts of privacy-preserving because of these reasons; random formation of the tree gives more security because to get prior information, one should find the entire classification model and cases. Since simple cryptographic strategy is slow and very little proficient concerning RDT because the branch of the tree is hidden to the outsider, the structure of RDT and its characteristics in which only the leaves of the tree are encrypted or decrypted [5, 6]. Four sorts of data mining tasks can utilize the same RDT code. RDT has another advantage because it can be made differentially private without losing data precision [7].

This paper proposes a proficient survey about privacy-preserving RDT. Its proficiency is its ability to maintain privacy and accuracy yet lessen computation time compared to existing algorithms. It uses an Iterative Dichotomiser 3 (ID3) and Boosting algorithm within an RDT, including a privacy-preserving algorithm. The algorithm developed safely builds an RDT for vertically and horizontally partitioned data. The Privacy-preserving RDT combines both randomization and cryptography techniques. This work executes a classification rule for data mining with privacy-preserving. Classification can be defined as storing information with similar features in the same class.

2 Random Decision Tree

2.1 Random Decision Trees Definition

A decision tree is a flowchart tree-like formation, which is utilized for decision analysis. In the decision tree, each internal node demonstrates the test of the attribute, the result of this test is represented by the branch of the tree, and the leaves demonstrate the class label or last result. The decision tree likewise demonstrates the alternative result for better comparison. A few classification algorithms for information discovery include neural networks, logistic regression, and decision trees. These decision tree classification techniques are variedly utilized with classification models, such as ID3, C4.5, CART, SLIQ, SPRINT, whereby every model heuristically employs splitting measures [6]. An RDT can be defined as the attribute selection done randomly since an arbitrary decision produces random information trees. RDT performs better than different models concerning calculation speed, in order to of the characteristics of random partitioning used as a section of tree development [8].

A RDT depends on two phases, training and classification, and the formation of a random tree is built entirely independent of the training information. While constructing every tree, first, begin with a list of attributes from the data set. Then, produce a tree by randomly choosing one of the attributes without utilizing any training information. Only the leaf nodes require to register the number of values of various subjects classified through the nodes in the tree. The training information is scanned exactly once to append the statistics in various random trees. While classifying a new example, the likelihood of various tree results is averaged to gauge a posteriori likelihood. The training stage consists of making the trees, defined as "BuildTreeStructure", and populating the nodes with the training example of information, defined as "UpdateStatistics". The assumption, all parties know the number of attributes depends on the training data set [1, 8].

In RDT, tree quits growing any deeper if a node becomes empty or there are no more instances of splitting in the current node, and the profundity of the tree exceeds a few limits. Random trees are utilized in communication networks to disseminate information from one node to another or gather information at a single assigned node.

2.2 Random Decision Trees Architecture

The structure of the privacy-preserving RDT appears below in fig 1. Distributed data acquisition depends on an issued query, which is a test dataset given as information tuples. That data alludes to obscure class names used to construct a privacy-preserving RDT. After ID3, the RDT classification happens, which utilizes a training dataset to delete noise and update the training dataset; then applies a random key on each classified part. Finally, the ID3 with RDT information produces a pattern dataset; again, used to classify the test dataset after boosting; this precisely reclassifies the training dataset information using a boosting algorithm. The verification and prediction phases classify information, foresee the class name, and prove the execution considering computational duration and accuracy [9-11].

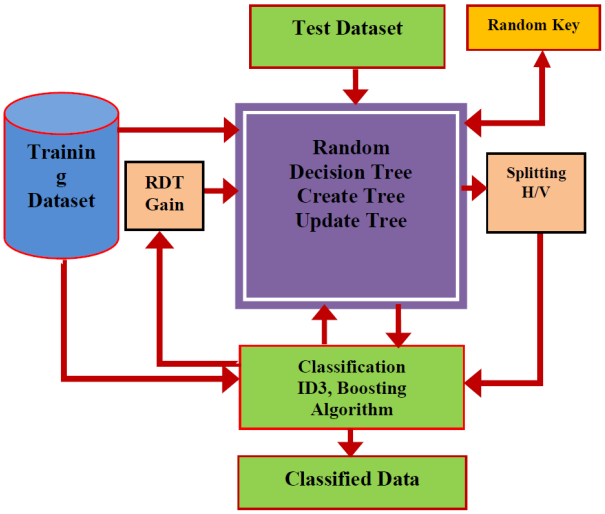

Fig. 1. Random Decision Tree Architecture.

3 Privacy-Preserving Data Mining

Privacy-preserving data classification aims to construct accurate classifiers without uncovering private data in the mined data. As of late, data mining has become noteworthy due to the proliferation of vast amounts of digital information [12]. Data mining includes analysing massive datasets, requiring computational procedures to ascertain patterns in those vast datasets. Furthermore, in the data mining process, data is shared between numerous clients while sharing information supplier needs to secure his sensitive information; hence, there is a method developed called privacy preserved data mining. Data mining privacy preservation is paramount today. Privacy preservation implies protecting gathered sensitive data gathered from various sources. The information patterns require collaborative discussions to refine handling methods employed in vast databases. However, these patterns can uncover sensitive data about individuals whose data alludes to the discussed patterns [13]. The idea of privacy-preserving data mining is to distinguish and forbid such disclosures, as evident in the sorts of patterns learned utilizing conventional data mining methods. In

addition, privacy-preserving permitted the linkage of databases to associations by protecting privacy.

4 Partitioning of Data

4.1 Vertically Partition Data

Jaideep Vaidya et.al [14] introduce privacy-preserving decision trees over vertically partitioned data, summed up privacy-preserving variation of the ID3 algorithm for vertically partitioned data disseminated over at least two parties. With vertically partitioned data, every party will include similar entities and gather information for a heterogeneous collection of attributes [10]. Presently, the parties cannot autonomously make a random tree except if they disclose the attribute data; along these lines, all parties share essential attribute data (i.e., metadata). Presently, the parties can autonomously form random trees without sharing data. Subsequently, the parties must work together to make the random trees; these trees could be in a distributed shape [15]. As opposed to cases using horizontal partitioning, the formation of random trees does expose potentially sensitive data because the parties do not know the attributes owned by the alternate parties; this way facilitates a direct address of instances within entirely distributed trees.

4.2 Horizontally Partition Data

However, gathering and partitioning data from various entities for every attribute alludes to how to build and classify RDTs. Since each party shares the diagram, a straightforward solution is that they make a few random trees autonomously [16]. Nonetheless, every party can independently make the formation of the tree. All parties must cooperatively and safely calculate the parameters over the worldwide data set. Dissimilar to the fundamental RDT approach, there is no requirement to keep the class circulation at every non-leaf node; only some leaf nodes require this data. Presently, there are two possibilities. First, the tree formation is known to every member, and second, the formation of the tree is obscure to every member. The worldwide class distribution vector for every leaf node is known to none of the parties.

4.3 Data Partitioning Example

Let us consider a situation for predicting cars in a car exhibition; this requires defining a set of attributes, such as car name, model, colour, price, and quantity. For an example, these are the attributes:

Table 1. Distributed exhibition cars dataset.

Car Name	Model	Colour
Hyundai	2012	White
BMW	2003	Black
Volkswagen	2014	White
Kia	2009	Black
Audi	2012	White
Camry	2003	Black

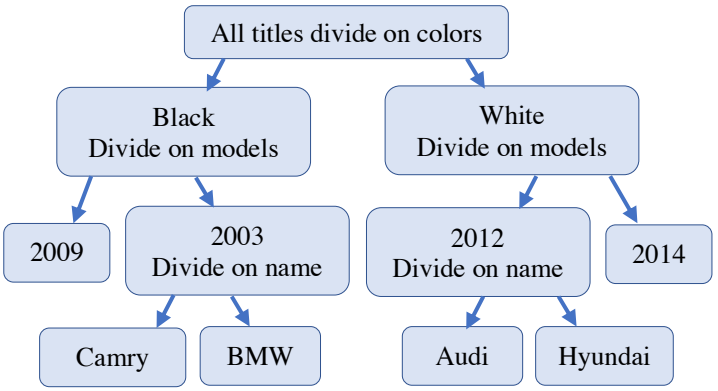

Fig. 2. A random decision tree with vertically partitioning data.

5 Privacy-preserving Random Decision Tree algorithm

The method outlined safely constructs RDTs for both horizontally and vertically partitioned datasets. Furthermore, the execution of the protocols provides a secure calculation at a low communication cost. Thus, RDTs can give excellent security with high profitability. Privacy-preserving RDT algorithm constructs different decision trees randomly. Building a tree is as follows: firstly, make a list of attributes utilizing the training dataset. Then, build a tree by randomly choosing listed attributes. Finally, the tree quits growing when it achieves the tree's limit of profundity. Thus, utilizing the ID3 makes a pattern dataset instead of a resulting classifier, which upgrades statistics at every leaf node utilizing the training dataset. To classify test datasets again requires the boosting algorithm with the help of a pattern. The algorithm builds every tree from the list of attributes and elects a "remaining" attribute randomly at every node. The consideration of an attribute as being "remaining" is if a similar attribute has not been selected in the past in a specific RDT node beginning from the root of the RDT to the present node. However, a continuous attribute can be chosen once in a similar decision path to the RDT. Thus, every time the class attribute is picked, a random threshold is chosen to enhance the conveyed of a distributed system for applications that work on a RDT overlay network.

6 Conclusion

Today, information privacy for various associations is paramount to expand their business since almost all organizations must share data without compromising privacy. This paper looks at randomization and cryptographic methods applied to sensitive information. The techniques outlined pertain to managing circulated information partitioned either horizontally or vertically across various locations within a secure framework of non-disclosure. This paper looks at the challenges of data mining tasks considering this backdrop. This paper concentrates on the technical plausibility of realizing privacy-preserving data mining. Privacy-preserving RDTs framework gives better precision with security and proficiency than RDT with privacy safeguarding system. The utilization of RDT can produce precise and sometimes better models with fewer costs. It improved the execution of RDT with a privacy safeguarding system. The approach utilizes distributed privacy-preserving RDTs, which influence a random structure to give complete privacy with less calculation than other methods. It can decrease computational time than RDT with a privacy-preserving system. Privacy-preserving RDT produces a highly accurate classifier, and learning is quick. This paper provides a case study of the horizontal and vertical partitioning of data using a privacy-preserving RDT.

References

1. Hemlata B. Deorukhakar1, Prof. Pradnya Kasture2" Adaptive Random Decision Tree: A New Approach for Data Mining with Privacy-Preserving", Vol. 3, Issue 7, July 2015.
2. R. Agrawal and R. Srikant, "Privacy-Preserving Data Mining," Proc. ACM SIGMOD Conf. Management of Data, pp. 439 450, May 2000.
3. G. Jagannathan, K. Pillaipakkamnatt, and R.N. Wright, "A Practical Differentially Private Random Decision Tree Classifier," Proc. IEEE Int'l Conf. Data Mining Workshops (ICDMW '09), pp. 114-121,2009.
4. W. Du and Z. Zhan, "Building Decision Tree Classifier on Private Data,"Proc. IEEE Intl Conf. Data Mining Workshop on Privacy, Security and Data Mining, pp. 1-8, Dec. 2002.
5. 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.
6. A. Dhurandhar and A. Dobra, "Probabilistic Characterization of Random Decision Trees," J. Machine Learning Research, vol. 9, pp. 2321-2348, 2008.
7. Jintu Ann John, Neethu Maria John,"Privacy-Preserving Random Decision Trees over Randomly Partitioned Dataset", vol.3, Issue 8, pp. 7746- 7750,2015.
8. W. Fan, H. Wang, P.S. Yu, and S. Ma, "Is Random Model Better? On Its Accuracy and Efficiency," Proc. Third IEEE Intl Conf. Data Mining (ICDM 03), pp. 51-58, 2003.
9. Matthew N. Anyanwu and Sajjan G. Shiva, "Comparative Analysis of Serial Decision Tree Classification Algorithms," International Journal of Computer Science and Security, (IJCSS) Volume (3): Issue (3).
10. Ming-Jun Xiao, Liu-Sheng Huang, Hong Shen and Yong-Long Luo, "Privacy-Preserving ID3 Algorithm over Horizontally Partitioned Data," Sixth International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT'05).
11. Saeed Samet and Ali Miri, "Privacy-Preserving ID3 using Gini Index over Horizontally Partitioned Data," IEEE2008.
12. Priyank Jain, Neelam Pathak, Pratibha Tapashetti ,A.S. Umesh " Privacy-Preserving Processing of Data Decision Tree Based on Sample Selection and Singular Value Decomposition" In Proceedings the 9th International Conference on Information Assurance and Security, pp. 91- 95,2013.
13. J. Vaidya, C. Clifton, and M. Zhu, Privacy-Preserving Data Mining. Advances in Information Security first ed., vol. 19, Springer-Verlag, 2005.
14. 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.
15. Weiwei Fang and Bingru Yang, "Privacy-Preserving Decision Tree Learning Over Vertically Partitioned Data," International Conference on Computer Science and Software Engineering 2008.
16. M. Kantarcioglu and C. Clifton, "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data," IEEE Trans. Knowledge and Data Eng., vol. 16, no. 9, pp. 1026-1037, Sept. 2004.