Indian Journal of Emerging Electronics in Computer Communications
Vol.2, Issue 1, 2015, Page.306-310, ISSN: 2393-8366

Privacy Preserving Data Mining using Random Decision Tree

C.Rajesh[1], S.Hari[2], U.Selvi[3],
UG Scholar[1, 2], Assistant Professor[3],
Department of Computer Science and Engineering[1, 2, 3]
Professional Group of Institutions, Tirupur, India. [1, 2, 3]

Abstract:

Data processing with information privacy and information utility has been emerged to manage distributed information expeditiously. In this paper, to deal with this advancement in privacy protective data processing technology victimization intensify approach of Random Decision Tree (RDT). Random Decision Tree provides higher potency and information privacy than Privacy secured Data mining Techniques. Privacy Preserving Data mining is simply too slow and impracticable to modify really massive scale analytics to manage era of huge information. Random Decision Tree is employed for multiple data processing tasks like classification, regression, ranking, 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.

Keywords: *Privacy Preserving, Data Mining, Random Decision Tree*.

1 INTRODUCTION

Data Mining is quick growing field of distributed atmosphere and method of discovering fascinating patterns and knowledge from giant information. It's additionally known as KDD process i.e. information Discovery from knowledge. It permits knowledge analysis whereas conserving knowledge privacy. Data privacy conserving is forestalling personal secret or non-public data from unnecessarily distributed or in public identified or not be put-upon by person or by oppose. In privacy preserving data processing, fascinating and helpful data is distributed with privacy of guidance has been preserved. There square measure 2 stages in privacy conserving knowledge mining initial is knowledge assortment and second knowledge commercial enterprise. In data assortment, knowledge holder stores knowledge that is gathered by data owner. In knowledge commercial enterprise, knowledge may be free to knowledge recipient by knowledge holder and knowledge recipient mines printed secured knowledge. Cryptographic techniques square measure typically too slow to be sensible and can become computationally expensive because the rise in size of the info set and communications between numerous parties increase [1]. Crypto graphical techniques cannot handle huge data. During this paper, we tend to square measure victimization privacy conserving RDT is Random decision Tree with privacy conserving data processing which is developed by Fan et al. [3]. Privacy conserving RDT is combination of randomization and cryptography technique.

This resolution provides Associate in nursing order of magnitude improvement inefficiency over existing solutions whereas providing a lot of knowledge privacy and knowledge utility. This can be an efficient resolution to privacy-preserving data processing for the massive knowledge challenge. Random decision Tree provides higher potency and knowledge privacy than crypto graphical technique. RDT provides a structural property, a lot of specifically, the very fact that solely specific nodes (the leaves) within the classification tree have to be compelled to be encrypted /decrypted, and secure token passing prevents adversary from utilizing count techniques to decipher instance classifications, because the branch structure of the tree is hidden from all parties. RDT to get trees. That square measure random in structure, providing USA with an analogous finish result as perturbation while not the associated pitfalls. A random structure provides security against investing priority information to get the whole classification model or instances.

2. RANDOM DECISION TREE

Random decision tree algorithmic rule constructs multiple iso-depth decision trees haphazardly. RDT relies on 2 stages, training and classification and s structure of a random tree is constructed utterly freelance of the coaching knowledge. When constructing every tree, first, begin with a listing of attributes from the info set. Generate a tree by haphazardly choosing one in every of the attributes while not victimization any coaching knowledge. The tree stops growing once the peak limit is reached. Then, use the coaching knowledge to.

Update the statistics of every node. During this solely the leaf nodes have to be compelled to record the amount of values of various categories that square measure classified through the nodes within the tree. The coaching knowledge is scanned precisely once to update the statistics in multiple random trees. When a brand new instance j, the likelihood outputs from multiple trees square measure averaged to estimate the a posterior probability.

The coaching part consists of making the trees (Build Tree Structure) and populating the nodes with coaching instance knowledge (Update Statistics). It's assumed that the number of attributes is thought to all or any parties supported the training knowledge set. The depth of every tree is determined supported a heuristic Fan et al. [3] show that once the depth of the tree is equal to 1/2 the entire range of attributes gift within the data, the foremost diversity is achieved, conserving the advantage of random modelling.

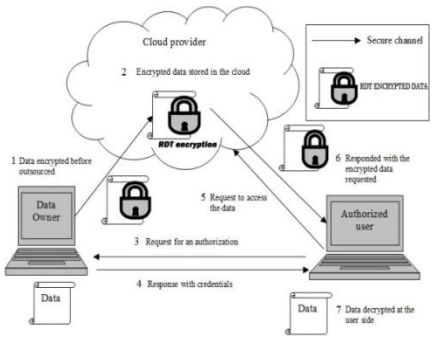

Fig:1 Architecture Design

Figure 1 shows the Architecture design of generating Random Decision trees. Data Owners stores the encrypted data in Cloud servers. Homomorphic Encryption is applied to nursery data. Only authorized users can access the data from Cloud servers. Hence the privacy of the data can be preserved.

In RDT, tree stops growing any deeper if one in every of the following conditions are met:

• A node becomes empty or there aren't any a lot of examples to split within the current node.

• The depth of tree exceeds some limit

Large-Scale distributed applications are subject to frequent disruptions attributable to resource rivalry and failure. Such disruptions are unpredictable and so hardiness may be a designable property for the distributed operated atmosphere .Describe and judge a sturdy topology for applications them to operate on a Random tree overlay network. This system is used for raising the hardiness of a distributed system. Random trees are utilized in communication network to disseminate data from one node to any or all alternative nodes and/or to gather data at one selected node. The most common Random trees are shortest path and minimum Random tree Is learned from the result (and the native input). Presumptuous the global knowledge set D ≡ (T, R), wherever T represents the world setoff transactions, and R represents the world schema, the general downside will be developed as follows;

Definition 1: (Privacy-Preserving RDT). Given a knowledge set D ≡ (T, R) distributed among k parties P1; . . . ; Pk Securely build a random decision tree classifier RDT, and provide a privacy- protective distributed classification mechanism to classify a replacement instance. There are two dependent steps to partition data: Horizontal and Vertical partitions. In this, once knowledge is horizontally divided between Pk parties, every party holds different instances, however collects a similar section of information. All parties share the schema, although the particular transactions in their native databases are distinctive. Clearly, since the schema is shared by all parties, the category attribute C is also renowned to any or all parties.

3. EXAMPLE

Table 1 shows the weather knowledge set distributed between two different parties. During this initial, knowledge set is horizontally partitioned; instances 1-5 are in hand by Party 1, while 6-10 are in hand by Party 2. If it's vertically divided, Party 1owns the parent nominal and has-nurs nominal attributes whereas Party a pair of owns the housing nominal, financial nominal, and social nominal. Suppose a replacement instance is to be classified. Then, as per the primary random tree, the prediction is (2, 0) without standardization. The prediction as per the second random tree is (1, 2). Therefore, the non-normalized overall class distribution vector provided by RDT is (1.5, 1).

4. DOWNSIDE STATEMENT

RDT code will be used for multiple data processing tasks; basic problem in distributed classification is to coach a classifier from the distributed knowledge so classify every new instance. For distributed random decision tree classification, the objective is to form a random decision tree classifier from the distributed knowledge. Within the privacy-preserving case, the additional constraint is that the method of building the classifier or of classifying AN instance mustn't leak any additional data or personal data.

PARTY		P1				P2			
	S.NO	PARENTS NOMINAL	HAS_NURS NOMINAL	FORM NOMINAL	CHILDREN NOMINAL	HOUSING NOMINAL	FINANCE NOMINAL	SOCIAL NOMINAL	HEALTH NOMINAL
	1	usual	proper	complete	1	convenient	nonprob	recommended	recommend
	2	usual	proper	complete	1	convenient	nonprob	priority	priority
	3	usual	proper	complete	1	convenient	nonprob	not_recom	not_recom
	4	usual	proper	complete	1	convenient	slightly_prob	recommended	recommend
p1	5	usual	proper	complete	1	convenient	slightly_prob	priority	priority
	6	usual	proper	complete	1	convenient	slightly_prob	not_recom	not_recom
	7	usual	proper	complete	1	convenient	problematic	recommended	priority
	8	usual	proper	complete	1	convenient	problematic	priority	priority
	9	usual	proper	complete	1	convenient	problematic	not_recom	not_recom
P2	10	usual	proper	complete	1	inconv	nonprob	recommended	very_recom

Table.1 Sample Nursery Database

5. HORIZONTAL PARTITION DATA

In this paper, once knowledge is horizontally divided between parties, every party holds completely different instances, however collects the same section of information. Every party will solely severally create the structure of the tree. All parties should co-operatively and firmly figure the parameters over world dataset. There are two options: 1) the structure of the tree is known to every party. 2) The structure of the tree is unknown to each party. Formula one offers the main points. When a replacement instance has to be classified, the party owning the instance identifies all of the leaf nodes that it reaches, and multiplies the encrypted category vector parts along to get the encrypted ads of the category distribution vectors as preach tree. This can be currently collaboratively decrypted and averaged to urge the particular category distribution vector. Note that, getting the add doesn't reveal quite obtaining the typical since the add will continually be retrieved from the typical and the total variety of trees. In algorithm 1, building a random tree is explained for horizontal partitioned data. The data is horizontally partitioned between sites. When there is a disagreement between parties in generating trees, secure electronic voting protocols is used. Otherwise, generated random tree is communicated to all parties. For each tree, each party encrypts the class distribution vector for all leaf nodes and hence privacy is preserved.

Algorithm 1: Build Tree (l,d)

Prerequisites: Transaction set T partitioned horizontally, n – the number of random trees such that$\sum_i n_i = m$, the total number of random trees.

Step 1:loop while condition is not satisfied for all random trees

a. Secure electronic voting protocol is used to hide information of corresponding party from other party of tree.
b. A random tree is created by each party
c. Structure of each created tree is alone communicated to all parties.

Step 2: loop for each tree created as

a. Class distribution is locally created for each leaf node.
b. Each party encrypts the class distribution vectors of all leaf nodes.

6. VERTICAL PARTITIONING DATA

Vertically partitioned data, all parties collect knowledge for the same set of values. Each party collects knowledge for a unique setoff attributes. Parties cannot severally produce even the structure of a random tree, unless they share the attribute information among one another. Thus, there are a unit 2 options: All parties share basic attribute info in order that they will independently produce random trees (at least the structure).There is no sharing of knowledge. Now, the parties have to be compelled to collaborate to form the random trees. These trees might themselves exist during a distributed kind.

6.1 Fully Distributed Tree

Unlike the horizontal partitioning case, the structure of the tree will reveal probably sensitive info, since the parties don't understand what the attributes area unit owned by the other parties. Therefore, we have a tendency to directly address the case of totally distributed trees. Each website is aware of the entire variety of random trees, denoted by m. The algorithms for making random trees area unit given in rule three.

Algorithm 2: Classify_Horizontal(instance_id)

Prerequisites: m- the number of random trees built(node id_j is the root node of jth tree)

Step 1: loop for all values of i- choose values of random numbers for the cryptographic system

$lv_i = encrypt(0, r)$

Step 2: for all values of j update the current status as encrypted class distribution vector for which $lv_i = lv_i * current\_status_i$

Step 3: decrypt each lv_i *and divide by m to get the actual statistics*

Algorithm 2 explains classify instance horizontally, with node_id as root node, it generate the decision tree. Each leaf node generated by the parties is encrypted and update the current statistics.

Algorithm 3: Creation of Random trees

Prerequisites: Transaction set T, P_i *that holds* n_i *attributes, p class values to hold class attributes, m- the number of random trees to build.*

Step 1: compute $n = \sum_i n_i$ *using the secure sum protocol.*

Step 2: depth is calculated as $^{n}/_{2}$

Step 3: loop i for all values of m, the node is built as Build_tree(level, depth)

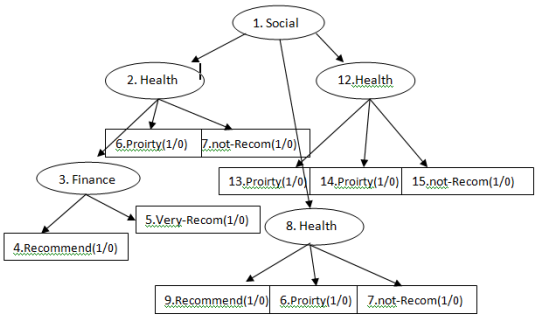

Fig.2 Random Decision Tree Generated

Depth of random tree is half the entire variety of attributes and is computed by all sites along victimization the secure add protocol[2]. Additively homomorphic encoding is employed for computation of the addition of encrypted values for change the class statistics. Secure Sum protocols are used by each party, where the overall average statistics is known but not the individual statistics and hence the privacy is preserved.

6.1.1 *Update Statistics*

For change the statistics, RDT formula, for every random tree and instance within the coaching dataset, the tree is traversed to the acceptable leaf node. At the leaf node, the part in the class distribution vector equivalent to the category label of the instance is incremented by one. This method is repeated for all the random trees. Within the vertically partitioned data, the nodes and therefore the attributes of the dataset square measure distributed across multiple sites. Whereas change statistics no site ought to learn the attributes and attribute values.

6.1.2 *Classification*

Instance classification additionally takes during a distributed fashion similar to update statistics. as an example classification, the prediction of the given instance is computed by taking AN average of the likelihood outputs from multiple RDTs. The distributed formula as an example classification is given in Algorithm seven. Then these class/ distribution as an example the instance diagrammatical by inst ID in tree given by node ID.

7. CONCLUSION AND FUTURE WORKS

In this paper, we tend to study the technical feasibleness of realizing privacy-preserving data processing. RDTs are often accustomed generate equivalent, correct and typically higher models with a lot of smaller cost; we tend to area unit exploitation distributed privacy-preserving RDTs. Our approach leverages the actual fact that randomness in structure will offer sturdy privacy with less computation. Within the future, we tend to conceive to develop general solutions that may work for indiscriminately partitioned off knowledge and overlapping group action.

REFERENCES

[1] Jaideep Vaidya, Senior Member, IEEE, Basit Shafiq, Member, IEEE, Wei Fan, Member, IEEE, Danish Mehmood, And David Lorenzi "A Random Decision Tree Framework Or Privacy-Preserving Data Mining" Proc. 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.ser. 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] W. Fan, J. McCloskey, and P. S. Yu, "A General Framework for Accurate and Fast Regression by Data Summarization in Random Decision Trees," Proc. 12th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '06), pp. 136-146, 2006.

[5] X. Zhang, Q. Yuan, S. Zhao, W. Fan, W. Zheng, and Z. Wang,"Multi-Label Classification without the Multi-Label Cost," Proc. SIAM Int'l Conf. Data Mining (SDM '10), pp. 778-789, 2010.

[6] A. Dhurandhar and A. Dobra, "Probabilistic Characterization of Random Decision Trees," J. Machine Learning Research, vol. 9, pp. 2321-2348, 2008.

[7] 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.

[8] 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.

[9] O. Goldreich, "General Cryptographic Protocols," The Foundations of Cryptography, vol. 2, pp. 599-764, Cambridge Univ. Press, 2004.