Privacy-Preserving Data Mining using RDT Framework and C4.5

Komal N. Chouragade[1], Trupti H. Gurav[2]

[1]P.G. Student, Department of Computer Engineering, SKN College of Engineering, Sinhgad, Vadgaon,Pune, India

[2]Associate Professor, Department of Computer Engineering, BVM Smt. Kashibai Navale College of Engineering, Sinhgad, Vadgaon, Pune, India

Abstract: ***In later year's privacy preservation in data mining has turned into main problem and needs to be solved. In this paper, to deal with this advancement in privacy preserving data mining technology using highlighted approach of Random Decision Tree (RDT) Random Decision Tree gives better efficiency and information protection over Cryptographic procedure. Cryptography technique is extremely slow and in plausible to enable truly huge scale investigation to manage time of enormous data. Random Decision Tree is utilized for different data mining task like classification, multiple classifications. Privacy-preserving RDT used for both cryptographic technique and randomization which provide data privacy for some decision tree algorithm. In this algorithm we are using ID3 and C4.5 Decision tree algorithms. By using C4.5 to improve Random Decision tree is the main contribution of our work.***

Keywords: Privacy-preserving, data mining, Classifier, decision Tree, ID3, C4.5.

1. Introduction

Data analysis and machine learning have led to the ability to improve customer service, simplify business processes, distribute scarce resources more efficiently, and. At the same time, due to the extensive availability and use of data, there is significant (and growing) worry about individual privacy. The difficulty of individual privacy is compounded by the availability of auxiliary information, which renders straightforward Approaches based on data masking is unsuitable. Data Mining has advanced in the field of distributed Environment and tactics of finding interesting examples and learning from expansive dataset. It allows data examination while preserving data privacy. Privacy preserving is to predict individual secret or private information from unnecessarily distributed publicly known or not be misused by third person or by adversary. In privacy preserving data Mining, interesting and valuable data is shared with security of confidential data has been protected. There are two stages into the privacy preserving data Mining first one is data collection and second one is data publishing. Information holder stores information which is collected by information owner. In information distributed, data can be discharged to data recipient by data holder and data recipient mines published secured data. Cryptographic methods are regularly too slow to be practical and can get to be computationally expensive as the rise in size of the data set and communications between various parties increase

[1]. Cryptographic techniques fails to handle big data. In this paper, we are using privacy preserving RDT [3] is Random Decision Tree with privacy preserving data mining which is developed by Fan et al. [2]. Privacy preserving RDT is the combination of cryptography method and randomization. This solution gives an order of magnitude improvement in efficiency over existing solutions while providing more data privacy and data utility. This is an effective solution to privacy preserving data mining for the big data challenge. Random Decision Tree gives better efficiency and data privacy than Cryptographic technique. RDT provides a structural property, more specifically, the fact that only specific nodes (the leaves) in the classification tree need to be encrypted/ decrypted, and secure token passing avoids adversary from using counting techniques to decipher instance classifications, as the branch structure of the tree is hidden from all parties. RDT used to generate trees that are random in structure, providing us with a similar end effect as perturbation without the associated pitfalls. A random structure provides security against leveraging a priori information to discover the entire classification model or instances.

2. Decision Tree Classifier

A Decision tree is one of the popular methods and is able to handle both categorical and numerical data and perform classification with less computation. Decision trees are often easier to interpret. Decision tree is a classifier which is a Coordinated tree with a node having no approaching edges called root. All the nodes except root have exactly one incoming edge. Each non-leaf node called internal node or splitting node contains a decision and most appropriate target value assigned to one class is represented by leaf node. ID3 was one of the first Decision tree algorithms. It works on wide variety of problems in both academia and industry and has been modified improved and borrowed from many times over. ID3 picks splitting value and predicators on the basis of gain in information that the split or splits provide. Gain represents difference between the amount of information that is required to correctly make a prediction both before and after the split has been made. Information gain is defined as the difference between the entropy of original segment and accumulated entropies of the resulting split segment.

3. Related Work

C.Rajesh, S.Hari, U.Selvi Privacy Preserving Data Mining using Random Decision Tree explains [1] Data processing with information privacy and information utility has been emerged to manage distributed information expeditiously. In

Licensed Under Creative Commons Attribution CC BY

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.

b) T. Satya Narayana Murthy [5] An Efficient Privacy Preserving Approach Using Id3 Decision Tree Learning Algorithm explains Privacy-preserving is an important issue in the areas of data mining and security. The aim of privacy preserving data mining is to develop algorithms to modify the original dataset so that the privacy of confidential information remains preserved and as such, no confidential information could be revealed as a result of applying data mining tasks. In existing system they introduced a new privacy preserving approach via data set complementation which confirms the utility of training data sets for decision tree learning.

c) L. Sweeney et al. in [6], K-anonymity: A Model for protect Privacy is an information change approach that expects to protect private data of the examples by generalizing attributes. K-obscurity exchanges security for utility.[7] Introduces a privacy preserving approach which can be applied for decision tree learning, without loss of accuracy. It depicts a approach to the protection of the privacy of collected data tests in situations where data from the sample database has been part of the way lost. This methodology changes over the first sample data sets into a gathering of changed or unreal data sets, from which the original samples can’t be recreated without the whole group of unreal data sets. This new approach can be grouped particularly to the information capacity when first sample is collected. The methodology is decently matched with other privacy preserving methodologies, for example, cryptography, for extra protections). Privacy-Preserving Decision Trees over Vertically Partitioned Data: Murat Kantarcioglu, Jaideep Vaidyaand A. Scott Patterson [8] introduce a Privacy-Preserving Decision Trees over Vertically Partitioned Data, generalized privacy-preserving variant of the ID3 algorithm for vertically partitioned data .which is distributed over two or more parties.. This paper presents a new protocol to construct a decision tree on vertically partitioned data with an arbitrary number of parties where only one party has the class attribute. It shows a general framework for constructing a system in which distributed classification would work. It serves to show that the methods can actually be built and are feasible. This work provides an upper bound on the complexity of building privacy preserving decision trees. Significant work is required to find a tight upper bound on the complexity. Privacy Preserving Decision Tree Mining from Perturbed Data: In this paper, Li Liu, Murat Kantarcioglu and Bhavani Thuraisingham [9] propose a new perturbation based technique to modify the data mining algorithms. They build a classifier for the original data set from the perturbed training data set. This paper proposed a modified C4.5 [10] decision tree classifier which is suitable for privacy preserving data mining can be used to classify both the original and the perturbed data. This technique considers the splitting point of the attribute as well as the bias of the noise data set as well. It calculates the bias whenever try to find the best attribute, the best split point and partition the training data. This algorithm is focused around the Perturbation plan, however avoids the steps of reconstructing the first data distribution. The proposed method has expanded the privacy protection with less computation time. Privacy as a security issue in data mining region is still remains a challenge.

4. Implementation Details

a) System Overview

A. Random Decision Tree:

The utilization of RDTs may appear unreasonable; there are many advantages as far as performance and accuracy that are taken up by utilizing this system versus conventional algorithms. Find that for classification, utilization of a random model can match, as far as solutions, other inductive learning models in discovering an optimal theory. In the RDT performs much better than different models with respect to computation speed, because of the characteristics of random partitioning utilized as a part of tree development.

The RDTs algorithm constructs multiple (or m) iso-depth RDTs. One important aspect of RDTs is that the structure of a random tree is built totally independent of the training data. The RDT algorithm can be divided into two steps, those are training and classification. The training step comprises of building the trees (Build Tree Structure) and populating the nodes with training example data (Update Statistics). It is considered that the quantity of attributes is known depended on the training data set. The depth of every tree is selected based on a heuristic a Fan et al. [11] Demonstrate that the most diversity is gifted, when the depth of the tree is equal to a half of the total number of features present in the data, preserving the benefits of random modeling. The procedure for generating a tree is as per the following.

1) Begin with a list of features that is attributes from the dataset. Create a tree by randomly selecting one of the features without utilizing any training data. The tree halts growing as Height limit is reached.
2) Utilize the training data to upgrade the statistics of every node. Note that just the leaf nodes. Require storing the number of illustrations of diverse classes that are classified through the nodes in the tree. The training data set is analyze once to redesign the statistics in different random trees. At the point when classifying another example x, the probability outputs from numerous trees are averaged to calculate the a posterior probability.

B. Horizontally Partitioned Data:

At the point when data is horizontally partitioned, parties gather data for distinctive entities, yet have data for the each of the attributes. We now need to evaluate how the RDTs can be developed and how classification is obtained. As all the parties share the schema, a simple solution is for all parties to separately make some random trees. Altogether these will create the ensemble of random trees. Besides that, every party can freely make the structure of the tree. All parties must Co-operatively and safely calculate the parameters that are estimations of every leaf node, over the universal data set. Dissimilar to the basic RDT technique,

Licensed Under Creative Commons Attribution CC BY

there is no inducing reason to keep the class distribution at every non-leaf node this data is just needed at the leaf nodes.

Presently, there are two conceivable outcomes:
1) Every member is known of the structure of the tree.
2) Every member is unknown of the structure of the tree.

Inside first possibility, there are three further possibilities:
1) All parties will be known of the global class distribution vector for every leaf node.
2) Only the party owning the tree is known of the global class distribution vector for every leaf node.
3) No party is known of the global class distribution vector for every leaf node.

Inside second possibility, there are two further possibilities:
1) The tree owning party is known of the values for every leaf node.
2) No party is known of the values for every leaf node.

C. Vertically Partitioned Data

With vertically partitioned data, data for same set of entities is collected by all parties. Not with standing, every party gathers data for a different set of attributes. Presently the party cannot separately make even the structure of a random tree, unless they share the attribute data between them. Subsequently, there are two possible outcomes:
1) All parties share fundamental attribute data that is metadata. currently they can freely generate random.
2) There is no sharing of data. Presently, the parties require to work together to generate the random trees. These trees could themselves exist in a distributed structure. In the horizontal partitioning case, the structure of the tree does uncover conceivably sensitive data, since the parties do not comprehend what are the attributes possessed by other parties.

Homomorphic Encryption and Decryption Scheme:

A cryptosystem is homomorphism [12] with respect to some operation on the message space if there is a corresponding operation on the cipher text space such that
e(m) * e(m) = e(m * m).
1) The algorithm uses a large number N, such that
N = P * Q, where P and Q are large security prime Numbers.
2) Given X, be the a plaintext data , the encrypted?
Value is computed:
Y = Ep(X) = mod((X + P * R);N)
Where mod() is a common modulo N-operation, R is an irregular number inside the uniform distribution (1;Q).
Given y, which is a cipher text message, we use the security key p to recover plain text
X = E -1(Y) = () = (Y; P);
Y = mod((X + P * R);N)
Note that: for any X although E1
X 6= E2; (E1(X)) = (E2(X)) which means there
is one to many relationship between plaintext X and cipher textE(X).

B. Mathematical Model for Proposed Work
1) Set Theory: The system S is represented as: S={HD,VD, TG, ED, DD}
1) Input Horizontal Dataset
Let HD is the set of input HD={hd1, hd2,...,hdn} Where, hd1,hd2,...hdn are the set of inputs.
2) Input Vertical Dataset
Let VD is the set of input VD={vd1, vd2,....dn} Where, vd1,vd2,...vdn are the set of inputs.
3) Tree Generation at each party.
Let TG is the set of tree TG={tg1, tg2,...tgn} Where, tg1,tg2,...tgn are the set of trees at pi party.
4) Fix the tree structure of tree.
5) Encryption of tree for security purpose before sharing
of tree.
Let ED is the set of dataset which is to encrypt tree.
ED= {ed1, ed2,...edn} Where, ed1, ed2,...edn are the set of encrypted dataset.
6) Formation of Global tree by aggregation of encrypted tree.
7) Decryption of Dataset for trusted party and use the result.
Let DD is the set of dataset which is to be decrypted.
DD={dd1,dd2,...ddn} Where, dd1,dd2,...ddn are the set of decrypted dataset.
8) Test the data to find its class.

5. Algorithm

Algorithm 1 ID3 algorithm

```
1: Input: A data set, S
2: Output: A decision tree
3: If the all the instances have the similar value for target attribute then return a decision tree that is simply this value
Else
4: Compute Gain values for all attributes and select an attribute with the highest value and create a node for that attribute.
5: Make a branch from this node for every value of the attribute
6: Assign all possible values of the attribute to branches.
7: Follow each branch by partitioning the dataset to be only instances whereby the value of the branch is present and then go back to 1.
```

The entropy of a datasets, with respect to one attribute, in that phase the target attribute, with the following calculation:

$$Entropy(s) = \sum_{i=1}^{c} p_1 \, log_2 p_1$$

calculates the reduction in entropy (Gain in information)that would result on splitting the data on an attribute, A.

Gain(S, A)= $\frac{Sv \vee \frac{}{S}}{\sum_{v€A}}$ Entropy(Sv)

Algorithm 2 C4.5 algorithm:

```
1: Input: Example, Target Attribute, Attribute
2: Classified Instances
3: Check for the base case
4: Construct a DT using training data
5: Find the attribute having the highest info gain
6: (ABest) ABest is assigned with Entropy minimization
7: divide S into S1,S2 ,S3...
8: according to the value of ABest
9: Repeat the steps for S1, S2, S3
10: For each D, apply the DT Base cases are the following:
```

www.ijsr.net ISSN (Online): 2319-7064

1) All the examples from the training which is belong to the same Class
2) The training set is empty
3) The attribute list is empty (returns a leaf labelled with the most frequent class or the disjunction of all the classes)

6. Results and Discussion

In our result shows C4.5 decision tree algorithm gives the more accuracy than RDT and ID3. Accuracy: The measurements of a quantity to that quantity's factual value to the degree of familiarity are known as accuracy. The Table 1 presents a comparison of ID3, C4.5and RDT accuracy with different data set size, this comparison is presented graphically in Figure 1.

Dataset size	ID3(%)	C4.5(%)	RDT(%)
14	94.15	96.2	91
24	78.47	83.52	72.33
35	82.2	84.12	77.65

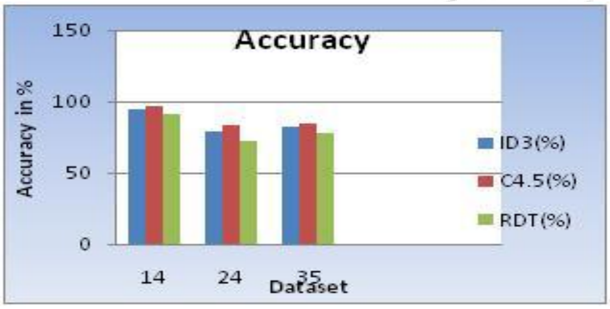

Figure 1: Comparison Graph

The 2nd is compared between ID3 and C4.5 is the execution time, Table 5 present the comparison. This comparison is presented graphically in Figure.

Dataset size	ID3(%)	C4.5(%)
14	0.215	0.0015
25	0.32	0.17
35	0.39	0.23

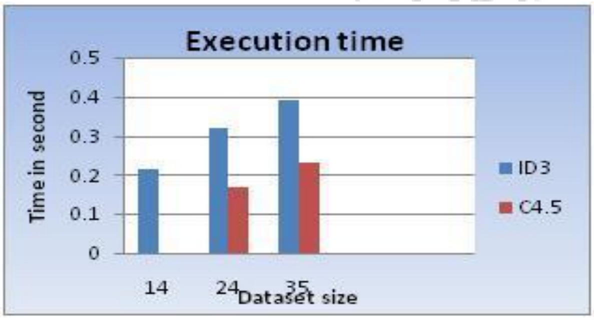

Figure 2: Comparison of Execution Time for ID3 C4.5 Algorithm

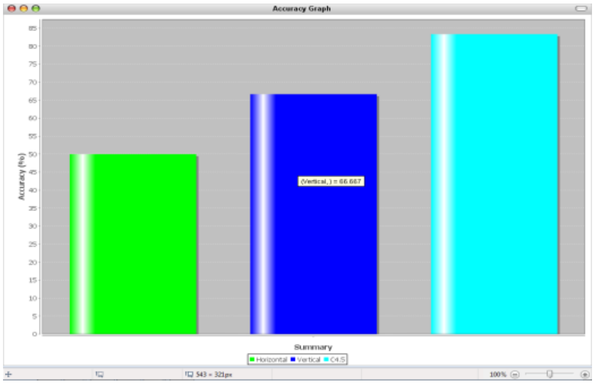

Figure 3: Accuracy Graph

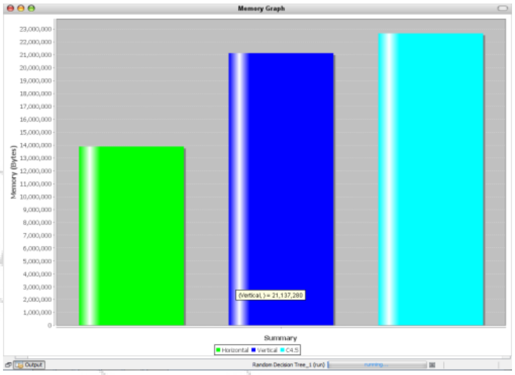

Figure 4: Memory graph

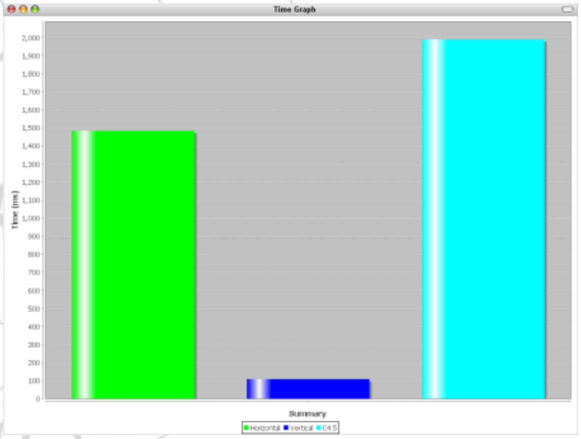

Figure 5: Time Graph

7. Conclusion

The security and privacy suggestions are considered when managing distributed data that is partitioned either on horizontally or vertically across multiple sites, and the difficulties of obtaining data mining tasks on such data. Since RDTs can be used to create identical, exact and off and better models with much less cost, distributed privacy-protecting RDTs is presented. This methodology powers the way that randomness in structure can give solid privacy with low computation. We also use C4.5 algorithm to generate decision for privacy preserving tree which have better performance than ID3 on the basis of accuracy, speed and memory storage.

8. Acknowledgment

The authors would like to thank the researchers as well as publishers for making their resources available and teachers for their guidance. We are thankful to the authorities of Savitribai Phule University of Pune and concern members of cPGCON2015 conference, organized by, for their constant guidelines and support. We are also thankful to the reviewer for their valuable suggestions. We also thank the college authorities for providing the required infrastructure and support. Finally, we would like to extend a heartfelt gratitude to friends and family members.

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] 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

[3] Jaideep Vaidya, Senior Member, IEEE, Basit Shafiq, Member, IEEE, Wei Fan, Member, IEEE, Danish Mehmood, and David Lorenzi "A Random Decision Tree Frameworkfor Privacy-Preserving Data Mining"

[4] C. Aggarwal and P. Yu, "Privacy-Preserving Data Mining:", Models and Algorithms. Springer, 2008.

[5] T. Satya Narayana Murthy [5] An Efficient Privacy Preserving Approach Volume 5, Issue 3, March 2015 [6] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," Int.l J. Uncertainty, Fuzziness and Knowledge-based Systems, vol. 10, pp. 557- 570, May 2002

[6] Pui K. Fong And Jens H. Weber-Jahnke, Privacy Preserving Decision Tree Learning Using Unrealized Data Sets Proc. IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 2, February 2012

[7] J. Vaidya and C. Clifton. Privacy-preserving decision trees over vertically partitioned data. In Proceedings of the 19th Annual IFIP WG 11.3 Working Conference on Data and Applications Security, Storrs,Connecticut, 2005. Springer. L.

[8] Liu, M. Kantarcioglu, and B. Thuraisingham, "Privacy Preserving Decision Tree Mining from Perturbed Data," Proc. 42nd Hawaii Int'l Conf. System Sciences (HICSS '09), 2009.

[9] Pui K. Fong and Jens H. Weber-Jahnke, "Privacy Preserving Decision Tree Learning Using Unrealized Data Sets". IEEE Transl. on knowledge and data engineering, vol. 24, no. 2, February2012

[10] J. R. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993

[11] A Random Decision Tree Framework for Privacy-Preserving Data Mining Jaideep Vaidya, Senior Member, IEEE, Basit Shafiq, Member, IEEE, Wei Fan, Member, IEEE, Danish Mehmood, and David Lorenzi

[12] R.Rivest, L. Adleman, and M. Dertouzos. "On databanks And privacy homomorphisms". In Foundations of Secure Computation, eds. R. A. DeMilloetal., AcademicPress, pp. 169-179.,1978.

www.ijsr.net
ISSN (Online): 2319-7064