ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

Adaptive Random Decision Tree: A New Approach for Data Mining with Privacy Preserving

Hemlata B. Deorukhakar[1], Prof. Pradnya Kasture[2]

M.E. Student, Department of Computer Engineering, R. M. D. Sinhgad School of Engineering, Pune, India.[1]

Assistant Professor, Department of Computer Engineering, R. M. D. Sinhgad School of Engineering, Pune, India.[2]

ABSTRACT: Now a day's fastest growing field data mining with privacy preserving is essential for fast development of high dimensional data and to manage that data efficiently while preserving privacy. In this paper, to deal with distributed data in privacy preserving data mining technology using classification approach of Adaptive Random Decision Tree. Privacy preserving ARDT uses ID3 and Boosting within RDT with privacy preserving framework to provide better performance than existing system. In existing system, cryptography based technique is still too slow to be effective for managing distributed data. Random Decision Tree with data privacy is generating equivalent and accurate model but it also slow in computational time when distributed data grows. Privacy preserving ARDT handles distributed data efficiently. Privacy preserving ARDT provides better accuracy with data mining while preserving data privacy and reduces the computation time as compared to RDT with privacy preserving framework.

KEYWORDS: Data mining with privacy preserving, Classification, Random Decision Tree, Boosting, ID3.

I. INTRODUCTION

Data Mining is the process of discovering interesting patterns, information and knowledge from large database. This also called as KDD process i.e. Knowledge Discovery from huge amount of Database. It allows data analysis while preserving data privacy and data utility. Data privacy is nothing but prevent personal confidential or sensitive data from unnecessarily distributed or publicly known or not be misused by third person. In data mining with privacy preserving, interesting and useful data is publish with privacy of sensitive information has been preserved. In data mining with privacy preserving, to deal with distributed data while preserving data privacy the perturbation based work do not provide satisfactory privacy [1], cryptographic technique are too inefficient and infeasible to enable large scale analytics of big data [1]. In Existing RDT with privacy preserving, uses both features randomization and cryptographic techniques but it still slow than proposed privacy preserving ARDT. This is an effective technique for data mining with privacy preserving of distributed data.

In propose system, we are use a simple classification approach for data mining with privacy preserving. Classification is data mining concept where it accurately predicts class labels for given data. Classifies data based on the training dataset and values (class labels) of classifying attribute and uses it in classifying new given data. Classification is two processes first is model construction (classifier): In this, describing a dataset of predetermined classes. Each tuple of training dataset is assumed to belong to a predefined class, and that determined by the class label attribute. The set of tuples of dataset used for model construction is training dataset. The model is represented as classification rules or classifier. Second is constructed model usage: In this constructed model used for classifying future or unknown objects and Estimate accuracy of the model by using test dataset. The known label of test dataset is compared with the classified result from the model. Accuracy rate of constructed model is the percentage of test dataset that are correctly classified by the model. Test dataset is independent of training dataset, otherwise over-fitting will occur. If the accuracy is satisfactory, use that model to classify given data tuples whose class labels are unknown.

The proposed work based on Random Decision Tree one of the classification techniques in data mining. Actually, RDT is implemented by W. Fan, H. Wang, P.S. Yu, and S. Ma [1], [2]. Important characteristics of RDT is that the

DOI: 10.15680/ijircce.2015. 0307004

same code can be used for multiple data mining tasks i.e. classification, regression, multiple classification, ranking [1], [2], [3], [4]. It can naturally fit into a parallel and fully distributed architecture. RDT also develop protocol to implement privacy-preserving multiple random decision tree that enable to general and efficient distributed privacy-preserving knowledge discovery [1].

In this paper, proposed work is Adaptive Random Decision Tree technique for data mining while preserving data privacy. Our contribution is provide better accuracy with data mining while preserving data privacy and reduces the computation time than RDT with privacy preserving and cryptographic technique by using ID3 and Boosting algorithm within RDT with privacy preserving algorithm.

II. RELATED WORK

In this First, Up to this point there are vast amount works in data mining with privacy preserving. The approach to data mining with protecting privacy of distributed data sources using cryptographic techniques using construction of decision trees and ID3 by Y. Lindell and B. Pinkas [5]. This work establishing system to secure multiparty computation for a preserving privacy owned data and resulting tree. But this work is still too inefficient for practical usage and it cannot manage distributed data sources efficiently.

K. Wang, Y. Xu, R. She and P.S. Yu [6] proposed solution is Distributed construction of decision trees using secure join classification for integrate private database that provide more secure and efficient system based on passing the transaction identifiers between site; but it does not reveal specific values of attribute, parties obtain knowledge about which transactions follow which path of the tree, that is one site say these two individuals have the same values of attribute.

W. Du and Z. Zhan [7] demonstrating solution system for constructing a decision tree classifier while preserving privacy for vertically partitioned data. This system more secure and efficient for two parties use semi trusted commodity server. But it is limited to two parties and is not implemented.

J. Vaidya, C. Clifton, M. Kantarcioglu, and A.S. Patterson [8] system solution has been implemented and extend to this in the multi-party case, assumption made up that the all parties must be known value of class attribute but for large scale data, its computational complexity is too high.

G. Jagannathan, K. Pillaipakkamnatt, and R.N. Wright [9] propose system building differentially private random decision tree classifier from a centralized data set that handles data updates. It returns without significant loss of accuracy of differentially private result. But there are limited Computational and communication resources.

There has been work on horizontally partitioned data [10] and vertically partitioned data [11] using association rule. The perturbation based technique in data mining with privacy preserving required additional technique to remove noise from original data [12], [13], and [14].

Jaideep Vaidya, Danish Mehmood, Wei Fan, Basit Shafiq, and David Lorenzi [1] used features of both randomization and cryptography to provide secure approach to performing classification but still high its computational time and communication cost.

In this paper, propose work is Adaptive Random Decision Tree for data mining with privacy preserving to overcome disadvantage of RDT with privacy preserving and cryptographic approach. Privacy preserving ARDT provide better accuracy with data utility and data privacy than cryptography Technique and RDT with privacy preserving. It provides better performance than RDT with privacy preserving. It reduces computational time as compared to RDT with privacy preserving.

III. PROPOSED WORK

Privacy preserving Adaptive Random Decision Tree algorithm building multiple decision trees randomly. Constructing a tree is as following. Start first with making a list of attributes using training dataset. Construct tree by randomly choosing attribute using list of attributes. The tree stops developing once when limit of depth level of tree is reached then create pattern dataset using ID3 instead of resulting classifier and update statistics at each leaf node using training dataset. To classify test datasets again using boosting algorithm with help of pattern for prediction of class label and apply random key on each classified data. When constructing each tree, the algorithm from list of attribute select a ”remaining” attribute randomly at each node and attribute is considered ”remaining” if the same attribute has not been select in the past in a particular adaptive random decision tree node starting from the root of adaptive random decision tree to the current node. However, a continuous attribute can be selected once in the same decision path of adaptive random decision tree. Each time the class attribute is chosen, a random threshold is selected. To improve the robustness of a distributed system for applications system that operates on a Random decision tree overlay network. Random decision trees are usually used for communication networks as a means to distribute information from one intermediate node to all other nodes and collect information at a single designated node while preserving privacy of information.

A. Our contribution

The main aim of this proposed solution to provide better accuracy while preserving privacy of data with reducing computational time and enhanced the performance of RDT with privacy preserving framework. We have to create pattern dataset using ID3 within RDT instead of resulting classifier. To classify test data again using boosting algorithm with help of pattern for prediction of class label.

B. Problem statement

In Data mining with privacy preserving ARDT, a dataset distributed among different parties securely build an adaptive random decision tree classifier using ID3 and Boosting within RDT algorithm, and provide a privacy preserving distributed classification mechanism to classify a new instance as tree grow as well as reducing computational time with better accuracy.

C. Architecture of proposed work

The architecture of privacy preserving Adaptive Random Decision Tree is shown below in fig 1. Distributed data collection based on given query. In this take input test datasets as given data tuples whose class labels are unknown for building privacy preserving adaptive random decision tree after ID3 within RDT classification using training dataset for removing noise from dataset and update training dataset then apply random key on each classified data. By using threshold function that is gain and entropy to get decision for creating random classier. Finally, training data and ID3 with RDT data create pattern dataset for again classify test dataset after Boosting to speed up the classification. In this reclassified training dataset for precise data using boosting algorithm. Boosting tree stage is used when dataset bigger. Verification and Predication stage classified data predict the class label and verify the performance with computational time and accuracy of the system.

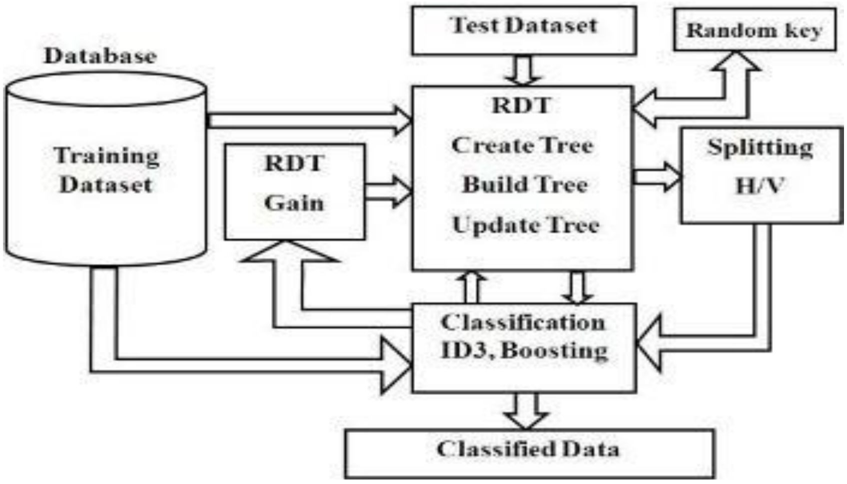

Figure 1: Architecture of Adaptive Random Decision Tree

DOI: 10.15680/ijircce.2015. 0307004

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

D. Algorithm of Adaptive Random Decision Tree

Algorithm 1: Adaptive Random Decision Tree

```
Input: D= {the Training Dataset}
       X= {the Test Dataset}
Output: Classified Data, ARDT i.e. R
     R= Building Tree Structure (UpdateSecrete(X));
        Update Statistics(R, D);
        Prune Sub tree with zero counts;
   Return R
```

Algorithm 2: Building Tree Structure(X)

```
If X= Ø    then
    Return a leaf node
Else
    Randomly choose on attribute F as testing attribute;
    Create on internal node r with F as the attribute;
    Assume F has m valid Training dataset value (D)
        r = Set Entropy ();
          Set Gain ();
          Set Random Union Key On RKey (RK*r);
          R = (RK ⊕ r);
            For i=1 to m do
              $C_i$= Build Tree Structure (X-{F});
              Add $C_i$ as a child of R(r);
            End For
End If
Return R(r)
```

Algorithm 3: Update Statistics (R(r), D)

```
For R (x) in D do
    Add instance (r, x);
End For
```

Algorithm 4: Add instance

```
If (R (r)) is not a leaf node then
    Let F be the attribute in r
    Let C represent the child of r that
    correspond to the value of F in x
    Add instance (c, x);
Else
    r is a leaf node;
    Let R be the Random Key
    R = (R key ⊕ r) + ∑ Boosting [t];
End If
```

Algorithm 5: Boosting tree

```
Input :{$R_1$................$R_N$ } and Keyed  ARDT & x, the row
      to be boost;
Output: Probabilities for all possible labels
   For a tree $R_i$ let $l_i$ be the leaf node reached by x
   Let α i[t] represent the count for label t in $l_i$
      $P_{Boosting}(t/x) = \sum_{i=1}^{N} \alpha\, i[t] / (\sum_r \sum_{i=1}^{N} \alpha\, i[t])$
   Return probability for all t
```

E. Mathematical Model

The Set theory:

R={S, A, C, M, P}

where, S = Input data set = {s_1, s_2, s_3, s_4,.., s_i}

A = No. of attributes= {a_1, a_2, a_3… a_n}

C = No. of data classes = {c_1, c_2, c_3… c_i}

M = Total no of trees = {m_1, m_2, m_3… m_i}

P = No. of Parties= {p_1, p_2, p_3… p_k}

To calculating gain information:

If a data set S contains examples from C classes, the Entropy(S):

$$\boldsymbol{Entropy(S) = -\sum_{i=1}^{c} p_i \log_2 (p_i)} \text{...} (1)$$

Based on the entropy in eq. (1) then we can compute the information gain if attribute A is used to partition the data set *S* shown in eq. (2):

$$\boldsymbol{Gain(S, A) = Entropy(S) - \sum_{v \in V(A)} \frac{|S_v|}{|S|} Entropy(S)} \text{..............} (2)$$

Where, v represents any possible values of attribute A; S_v is the subset of S for which attribute A has value v; $|S_v|$ is the number of elements in S_v; |S| is the number of elements in S.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

F. Feasibility Study

For feasibility study of proposed system used 3 SAT problem. The 3 SAT Problem is NP Complete, and system can be reduced to 3SAT problem.

- Boolean Formula is denoted by F and there are 3 Literals (a, b, c) in eq. (3)
- a: Transform training dataset in horizontally/ vertically.
- b: Classify each new instance and split the current node.
- ~b: A node becomes empty or there are no more examples to split in the current node.
- c: Build ARDT tree.
- ~ c:Calculate predication.
- F = (a ᵛ b ᵛ c)ᴧ (a ᵛ ~b ᵛ ~ c)ᴧ (a ᵛ b ᵛ ~ c)………………..(3)

F stands true for each and every case because a is true in all cases. So it proved that problem is NP-Hard. We know that all problems in P are also in NP. Therefore, since our problem is NP and NP hard (as proved above), problem is NP Complete.

IV. Results

The proposed ARDT privacy preserving framework using financial market dataset (share market dataset) is in .Net. In this use four company's financial market dataset as test dataset and user dataset i.e. how user invest in share market during that duration as training dataset to generate privacy preserving ARDT. It shows that ARDT privacy preserving provides better accuracy than RDT with privacy preserving framework as shown in fig. 2. It manages distributed dataset effectively by providing better accuracy while preserving privacy of data. It handles distributed data efficiently shown by using different formatted financial market dataset to generate privacy preserving ARDT.

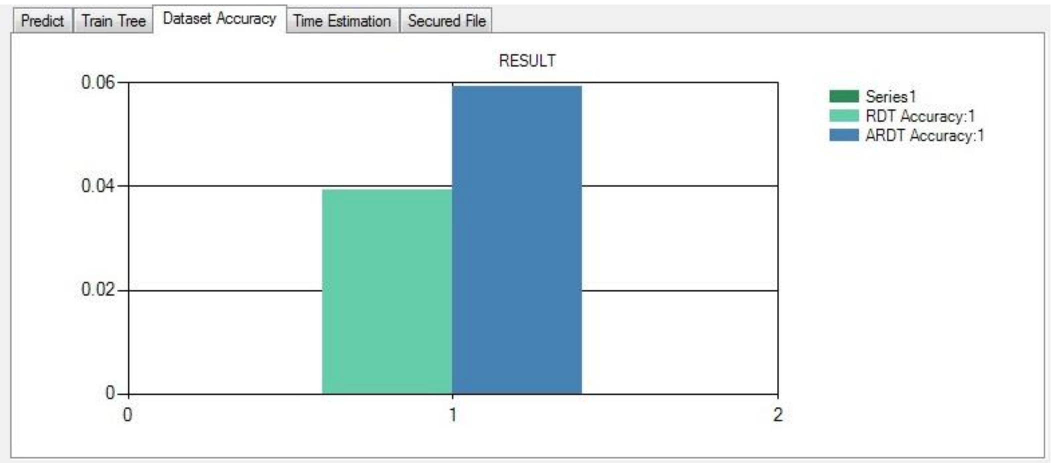

Figure 2: Actual Results (Accuracy)

In this, generate RDT with ID3 using training dataset and prediction values with performance measures based on it. Again train tree with test dataset using Boosting and Predication values based on generated ARDT then secured file of predication. Verification of accuracy and computational time of RDT and ARDT .It also reduces the computational time as compare to RDT framework shown in fig 3. ARDT privacy preserving enhanced the performance of RDT framework and efficiency. This determining accuracy of resultant classified data and computational time to evaluate performance of system, Conclusion and future work. Actual results to be shown in fig. 2 and 3.

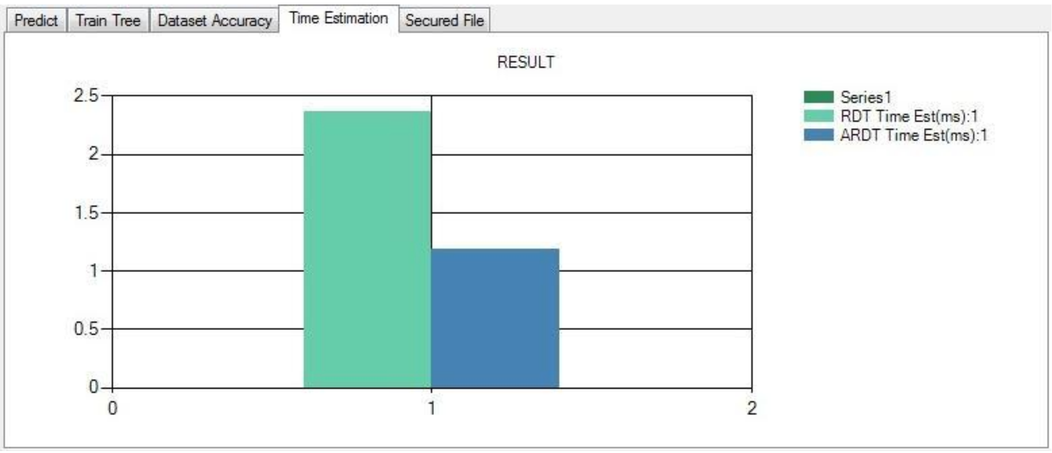

Figure 3: Actual Results (computational time)

V. Conclusion and Future Work

Privacy preserving ARDTs framework provides better accuracy with security and efficiency than RDT with privacy preserving framework. It can be used to generate accurate and sometimes better models with much smaller cost. It enhanced the performance of RDT with privacy preserving framework. It can efficiently handle distributed data than RDT with privacy preserving framework. It can reduce computational time than RDT with privacy preserving framework. Privacy preserving ARDT produces a highly accurate classifier and learning is fast. Privacy preserving ARDT is growing tree so it grows as instances are created.

In the future work, plan to generate general solutions system that can work for arbitrarily partitioned data and overlapping transaction.

VI. Acknowledgment

The authors heartfelt thankful towards the people whose help and support was extremely useful to accomplish this dissertation work on to Adaptive Random Decision Tree: A new approach for Data Mining with Privacy Preserving. It is extremely honour to say sincerest regards to P.G. Guide Prof. Pradnya Kasture for her precious inputs, whole co-operation, Direction, encouragement, suggestion and comment the during this dissertation work. Also grateful towards H.O.D., Director and the management. The Dissertation is based on research work in A Random Decision Tree Framework for Privacy-Preserving Data Mining by G. Jaideep Vaidya, Senior Member of IEEE, Basit Shafiq, Member of IEEE, Wei Fan, Member of IEEE, Danish Mehmood, and David Lorenzi. Author is extremely thankful to them for such a valuable work.

References

[1] 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

[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] 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 Intl Conf. Knowledge Discovery and Data Mining (KDD 06), pp. 136-146, 2006.

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

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

[6] K. Wang, Y. Xu, R. She, and P.S. Yu, “Classification Spanning Private Databases,” Proc. 21st National Conf. Artificial Intelligence, pp. 293-298, 2006.

[7] 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.

[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] 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] 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.

[11] J. Vaidya and C. Clifton, “Privacy Preserving Association Rule Mining in Vertically Partitioned Data,” Proc. Eighth ACM SIGKDD Intl Conf. Knowledge Discovery and Data Mining, pp. 639-644, July 2002.

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

[13] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, “On the Privacy Preserving Properties of Random Data Perturbation Techniques,” Proc. Third IEEE Intl Conf. Data Mining (ICDM 03), Nov. 2003.

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

[15] Rupali Bhardwaj , Sonia Vatta, “ Implementation of ID3 Algorithm,” Proc. International Journal of Advanced Research in Computer Science and Software Engineering, pp. 845-851,June 2013.

[16] Hemlata B. Deorukhakar, Vina M. Lomte, “A new approach for data mining with privacy preserving using Adaptive Random Decision Tree” Proc. CPGCON, March 2015.

BIOGRAPHY

Hemlata B. Deorukhakar received the B.Tech. Degree in Computer Science and Engineering from IGNOU, New Delhi in 2013. Currently appearing M.E. 2nd year Computer Engineering from Pune university in RMD SSOE, Pune, India.

Prof. Pradnya Kasture received the B.E. degree in Computer Engineering from Nagpur University, RPCE, Nagpur and M.E. degree in Information Technology from Pune University, SCOE, Pune Currently working as Assistant Professor of Computer Engineering Department in RMD SSOE Pune, India.