MJRET

Open Access

Multidisciplinary Journal of Research in Engineering and Technology

Journal homepage: www.mjret.in

ISSN:2348 – 6953

PRIVACY PRESERVING DATA MINING BASED ON RANDOM DECISION TREE FRAMEWORK

Rimi Kumari, Mr. Soumitra Das

Department of Computer Engineering
Dr.D.Y.Patil School of Engineering
Pune, Maharashtra, India

Abstract:*Data processing with information privacy and information utility has been emerged to manage distributed information expeditiously. In this paper, we tackle the problem of privacy in sensitive data based on RSA encryption algorithm within RDT. We introduce a generalized privacy preserving variant of the ID3 algorithm for vertically partitioned data distributed over multiple parties.The RSA technique provide two keys private and public which gives strong privacy as well as easy sharing of data in multiparty computation.It also maintain high utility of data and low computation cost because of boosting algorithm applied on the classified data. We have used this as an application in libraries which has less space storage.*

Keywords: *Random decision Tree, RSA, vertical partitioning, RDT****.***

1. INTRODUCTION

Recently data mining has attracted more attention due to popularity of big data. Data mining involves analysis of big data. It is the computational process of discovering patterns in big datasets .In data mining process data shared between many users while sharing data provider wants to secure his sensitive data .To achieve this there is a technique called privacy preserving in data mining developed. Previously data stored in a centralized fashion causes inefficiency and security related issue in big data. Now the data is distributed between two or more locations (sites), and these sites cooperate to achieve the global data mining results by maintaining the privacy of individual sensitive data .There are many approaches applied to preserve privacy of data mining like perturbation, anonymization, cryptographic etc. But these techniques gave less privacy, and also not increases the efficiency and infeasible to analyze the big data. In implemented RDT with privacy preserving, uses both approach randomization and cryptographic techniques [1] but it is still slower than proposed privacy preserving RDT.

In proposed work, we are implementing a classification rule for data mining with privacy preserving. Classification can be defined as storing the data in a class with the similar features of the other data. This can be done by referring the original data or by following the model of data. Classification can be done in two steps, the first step is supervised learning (where a classification model is constructed) and the second step a classification step (where the model is used to predict class labels for given data) or it is used to classify the accuracy of data.

In our proposed work we are using a classification technique called Random Decision Tree which is used for many data mining tasks i.e. classification, regression, multiple classification ,ranking [1], [2], [3], [4].

In this paper, we provide better accuracy as well as reduce the computation time compare to RDT [1] by using RSA cryptography algorithm and boosting algorithm while maintaining privacy in data mining.

2. RELATED WORK

To protect sensitive data in data mining process there are many methods applied and experimented. But still this is a problem to give accurate data and protect individual private data.

Sheikh, B. Kumar, D. K. Mishra proposed system [5] dk-Secure Sum Protocol for multiparty computation to preserve individual private data. This protocol distributes the data segment into different parties before computation to provide zero leakage of information but it increases the computation time and also the cost factor.

H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar [6] proposed system masking the individual private data by adding some noise in the original data. It is an additive random – matrix data filtering technique adds only white noise which is not provide stringent privacy to data.

Benny Pinkas[7] Proposed system secure the multiparty computation based on cryptography ID3 algorithm. It provides a generic construction which is a combination of many circuit which in not so efficient and infeasible in nature.

J. Vaidya and C. Clifton Proposed system [8] provide security on vertical partitioning data using scalar product on individual data based on association rule. It is applicable only for two party computation and limited on Boolean association rule.

M. Kantarcioglu and C. Clifton[9] proposed system provide security on horizontally partitioned data using cryptographic approach by adding little overhead to mining task.

3. PROPOSED WORK

In privacy preserving data mining Random decision tree algorithm create multiple decision tree randomly. How the random decision tree is constructed? Firstly start with creating attribute lists from the training datasets. Now generate a tree by choosing an attributes randomly. The tree stops growing when reached the height limit. Before applying update

statistics on each leaf node a pattern dataset is created using ID3 algorithm. With the help of this pattern which is used for predicting the class label and test dataset is classified again using boosting algorithm and then put RSA encryption technique on each classified data .During generation of tree each time select the new attribute which is not yet accessed from root to current node. To update the statistics in multiple random trees the training dataset is accessed only once.

Here we apply Divide and conquer strategy.

1. We will select best attribute for splitting.
2. For each attribute create new child nodes.
3. For each child nodes
   a. If node is leaf node then stop
   b. Else keep splitting.
   c. End if.

3.1 Architecture of proposed system

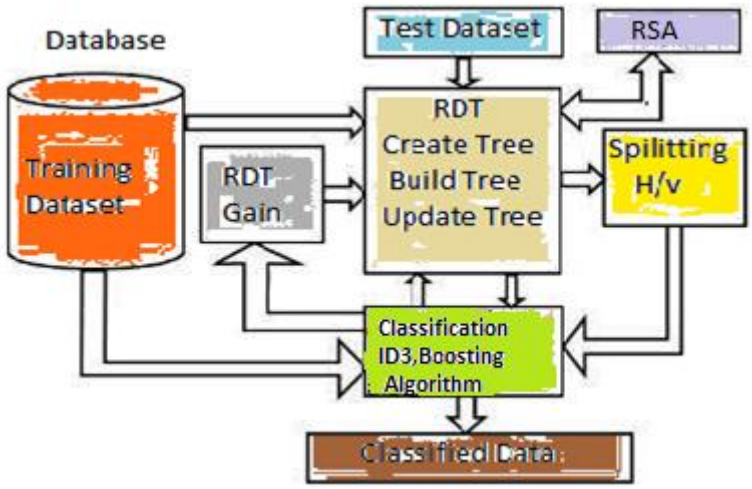

Fig.1 Architecture of Random decision tree

In this architechture input is taken from test dataset as data tuples which has unknown class label after applying ID3 classification algorithm within RDT .Training dataset is used to remove noise from dataset.After this update the training dataset and then apply RSA encryption algorithm to provide privacy in the classified sensitive data .For taking decision on creating random classifier we use threshold function gain and entropy.To speed up the classification process we use boosting algorithm before creating a pattern dataset for again classify test data set.

In the existing architecture [11] they used a random key approach for protecting the private data without giving any specified approach. we use RSA technique to provide privacy in classified sensitive data. In RSA cryptography approach it uses two keys:Public key to encrypt messages known to all party and private key to decrypt messages. This mechanism gives better security to the data than existing technique.

3.2 Data partitioning

We consider a scenario for predicting library book use. For this we consider set of attributes as name, last use, publication date, language, country, and alphabetic prefix of the Library. We will only consider here attributes

Last use	Publication date	Language
Feb 2015	1994	Hindi
Dec 2014	2010	English
Sept 2012	2015	Hindi
Oct 2014	1992	English
Jan 2015	1994	Hindi
March 2013	2010	English

Table 1:Distributed library dataset

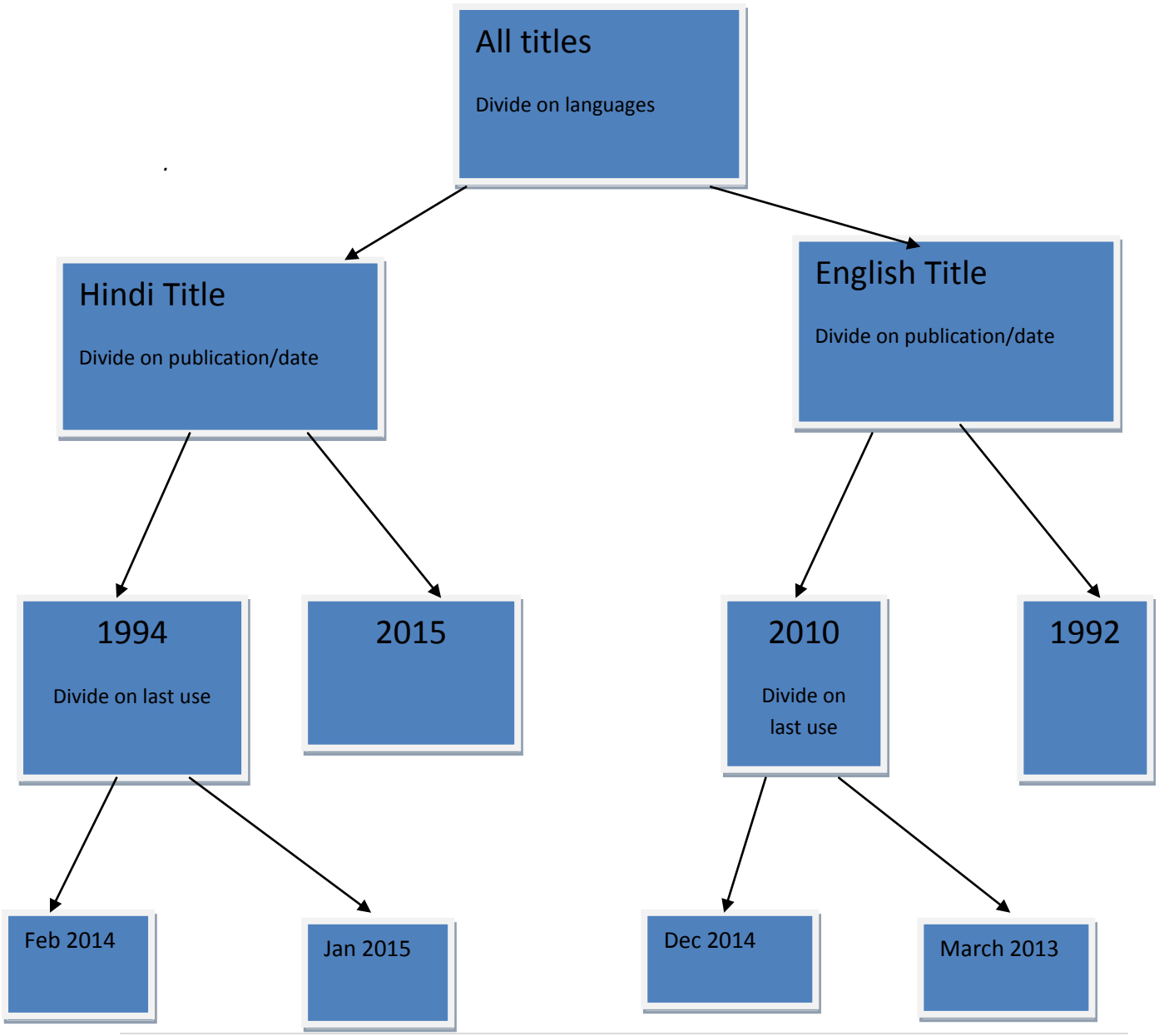

Fig:2 Random decision tree with vertically partitioning data

We consider dataset given in Table 1. When vertically partition all parties collects data for same set of entities. But each party collects data for different set of attributes. If the parties are willing to share information then only it is possible to create trees randomly.We explain the instance classification procedure using the above table. The root node is classified on basis of languages. Since more information needs to be classified, it is further divided on year. As on level 2, some attributes are classified. But some need to be classified further; hence we classify using last use attribute.

We keep dividing nodes until we reach upto leaf node. After reaching leaf node, we need to stop. In other words, we keep splitting until further partition is unavailable.

In our case we collected different information about same set of entities. One individual user can only know the class attribute. We must keep in mind that the class attributes should be known to all the parties. It is more general case, and is therefore considered. After reaching that stage we need to encrypt the decision tree obtained using appropriate encryption algorithm. For our case we will consider RSA algorithm. After encryption, the data tree can be shared with the user having suitable key. It will thus maintain accuracy in individual sensitive data. Also the node generated statistics can be seen as private data without loss in accuracy.

4. EXPECTED RESULT AND DISCUSSION

In RDT approach there are three dataset is used car ,nursery and mushroom dataset.The expected result will be like this.Classification,and computation time will be same but privacy of data will be more secure.

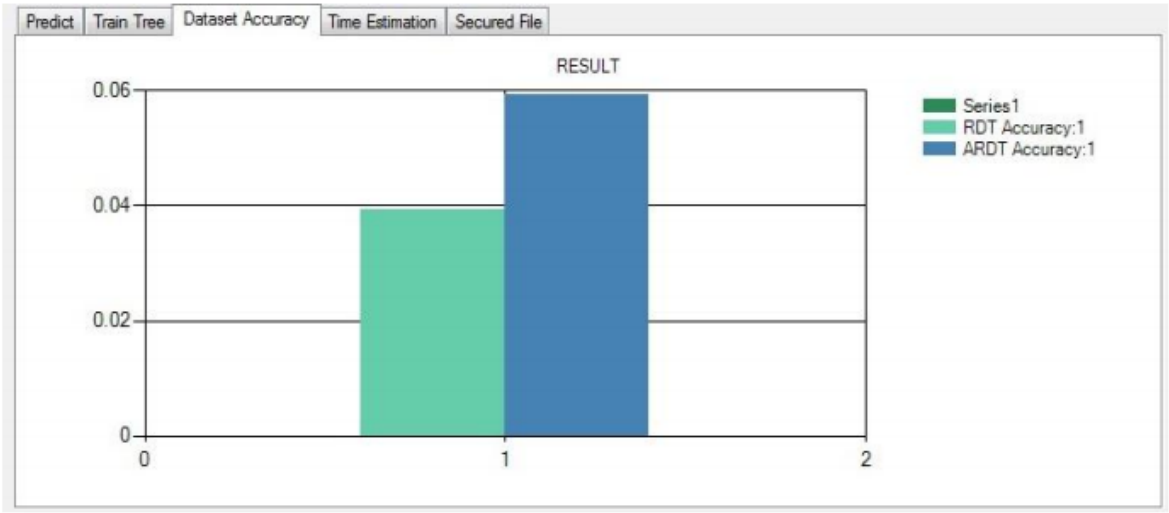

Fig.3: Actual result

5. CONCLUSION & FUTURE SCOPE

Privacy preserving in distributed data using random decision tree with RSA and boosting algorithm improves efficiency than simple RDT framework .It also maintain accuracy in data while preserving privacy in individual sensitive data.This mechanism also reduces computation time than previous implemented paper.

In future we have to be more focus on categorical data as well as arbitrary set of data.

ACKNOWLEGEMENT

We would like to thank our guide Mr. Soumitra Das, HOD of computer department and respected professors for giving guidance to understand the paper and motivate to develop new ideas.

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 MultiLabel Cost," Proc. SIAM Intl Conf. Data Mining (SDM 10), pp. 778-789, 2010.

[5]. R. Sheikh, B. Kumar, D. K. Mishra "A Distributed k-Secure Sum Protocol for Secure Multi-Party Computations",JOURNAL OF COMPUTING, VOLUME 2, ISSUE 3, MARCH 2010, ISSN 2151-9617

[6]. 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

[7]. Benny Pinkas "Cryptographic techniques for privacy-preserving data mining",HP Labs

[8]. J. Vaidya and C. Clifton, "Privacy Preserving Association Rule Mining in Vertically Partitioned Data,"Proc. Eighth ACM SIGKDDInt'l Conf. Knowledge Discovery and Data Mining, pp. 639-644, July 2002.

[9]. 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

[10]. Jaideep Vaidya, Senior Member, IEEE, Basit Shafiq,Member, IEEE, Wei Fan, Member, IEEE, DanishMehmood, 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

[11]. 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