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. 4, Issue 6, June 2016

An Efficient Approach for Privacy Preserving Data Mining Based on Random Decision Tree

Rimi Kumari, Soumitra Das

M.E Student, Dept. of Computer Engineering, Dr.D.Y.Patil School of Engineering, Lohgaon , Pune, India

HOD, Dept. of Computer Engineering, Dr.D.Y.Patil School of Engineering, Lohgaon, Pune, India

ABSTRACT: Big data play very important role in recent year. Centralized approach is inappropriate for most of the distributed and global data mining applications. In distributed environment due to long response time and lack of satisfactorily use of distributed resource centralized data do not perform well. To solve this problem distributed and parallel algorithm were introduced. In distributed environment data are from different source so there is a chance of disclosure of private data and also user get less accurate data .To avoid this there are many approaches were applied like Cryptography , randomization, perturbation and anonymization etc. In this paper a new approach for generating random decision tree is introduces based on entropy and gain calculation .To preserve privacy of data cryptographic techniques are applied on the leaves node of the decision tree.

KEYWORDS: PPDM, Randomization, Cryptography, ID3, Decision tree

I. INTRODUCTION

Data mining is the procedure of detecting a pattern from large datasets. Today organisations are very much dependent on distributed data means data are from different source. Due to this there is a possibility of sensitive information leakage of data providers. To protect this disclosure of data many techniques were applied like cryptography, randomization, perturbation and anonymization.

In this paper, we propose a combination of two techniques randomization and cryptography for more efficient and secure data mining task. With my best knowledge the proposed techniques are more efficient and secure than the existing system.The proposed system based on random decision tree structure . Same RDT code can be used for multiple data mining task such as Classification, regression, ranking and multiple classifications. RDT gives a better solution to the Distributed data mining in terms of privacy preserving because of these reasons:Random structure of the tree gives more security because to obtain priori information one should discover the entire classification model and instances. Since simple cryptographic technique is slow and not much efficient with respect to RDT because the structure of RDT and its characteristics in which only the leaves of the tree are encrypted / decrypted , the branch of the tree are hidden for the outsider. Same code of RDT can be used for four types of data mining tasks. RDT has another advantage is that it can made differentially private without losing accuracy of data.

In this paper, proposed work is efficient random decision tree while preserving privacy of data. Our contribution is to provide privacy, accuracy of data and also to reduce the time complexity than existing system. Here a algorithm is develop to securely construct RDT for horizontally and vertically partitioned data. Decision tree is a flowchart tree like structure which is used for decision analysis .In decision tree every internal node shows the test attribute, the outcomes of this test is represented by the branch of the tree and the leaves shows the class label or final result.Decison tree also shows the alternative outcome for better comparision .Random decision tree can be defined as the attribute selection can be done by randomly .In this paper a mathematical calculation has been done for selecting the attribute i.e Entropy and Gain .Entropy is used to calculate the impurity or uncertainty and Gain is used to calculate purity. At each node of the decision tree this calculation is performed and the largest information gain attribute is selected for next test attribute .This is a greedy approach.The tree split process is a recursive in nature from root node to down and it stops when the purity level reached or the leaves having the same class label whether there is no need to split the tree further.

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. 4, Issue 5, May 2016

II. RELATED WORK

In distributed architecture there are many work have been done to provide security to the data .In the survey process many papers are studied and collected the information about the methods applied on data to be secured.G. Jaideep Vaidya, Basit Shafiq,Wei Fan,Danish Mehmood, And David Lorenzi [1] the authors provide a solution to preserve privacy in data mining by random decision tree approach for horizontal and vertical partitioned data .A homomorphic cryptographic technique is used to provide security to the leaf node . Hemlata B. Deorukhakar1 , Prof. Pradnya Kasture [2] the author provides a new approach in which ID3 and boosting algorithm used within RDT .It focused on to provide high accuracy of data. Sathya Rangasamy,P.Suvithavani [3] this paper introduces a approach to provide privacy without loosing accuracy of data. It distributes the original datasets into group of unreal datasets by adding some noise in the datasets so that imposter can only know about data when they have total unreal datasets.Priyank Jain ,Neelam Pathak, Pratibha Tapashetti ,A.S. Umesh [4] this paper explain the ensemble classifier which is a combination of clustering and classification techniques .It focused on to provide data privacy and data utility.Jintu Ann John, Neethu Maria John [5] this paper recommend a solution to the accuracy of data.It takes random partitioned datasets instead of taking horizontal and vertical patitioned datasets.It provides more data security and load balancing.Pui K. Fong and Jens H. Weber-Jahnke [6] this paper revealed the combination of three approach perturbation,randomization and secure multiparty computation for privacy preserving data mining.

III. PROPOSED ALGORITHM

A. *RDT Generation*

In this phase training datasets are taken here Car datasets for horizontal partition data and Mushroom datasets for vertical partition are taken from UCI machine learning repository . Random decision tree is generated by first selecting attributes from training datasets . Now calculate :

The Set theory:

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

Where

S=Input data set = $\{s_1,s_2,s_3,\ldots.s_i\}$

A= No. of attributes = $\{a_1,a_2,a_3,\ldots a_n\}$

C = No. of data classes = $\{c_1,c_2,c_3,\ldots.c_i\}$

M = Total no. of trees = $\{m_1,m_2,m_3,\ldots m_i\}$

P= Total no. of parties = $\{p_1,p_2,p_3\ldots..p_k\}$

To calculating gain information;

If a dataset S contains examples of C classes , the Entropy(S) which gives how uncertain a attribute is:

$$\textbf{Entropy(S)} = -\sum_{i=1}^{c} P_i \log_2(P_i) \ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

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

$$\textbf{Gain(S,A)} = \textbf{Entropy(S)} - \sum_{v \in V(A)} |S_v|/|S| \ \textbf{Entropy(S)} \ldots\ldots\ldots \quad (2)$$

Where ,v represents any possible values of attribute A ; $|S_v|$ is the subset of S for which attribute A has value v; S is the number of elements in S_v ; $|S|$ is the number of elements in S. Gain gives clarity to choose the attribute from training dataset for generating tree .

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. 4, Issue 5, May 2016

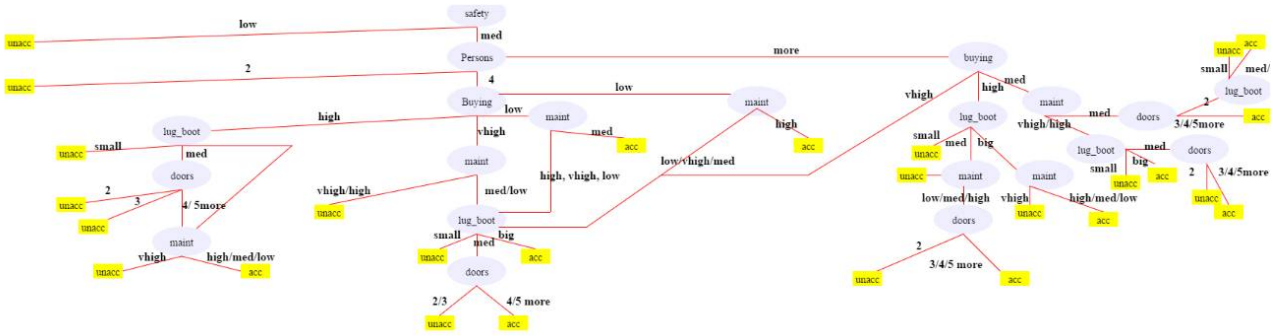

Fig 1: Snapshot of RDT generation

B. *RDT Encryption*

In this module The class label of random decision tree is get encrypted by using Two-Fish algorithm.This algorithm works on 128 bits block cipher and different size of key having 16 rounds of fiestel networks .

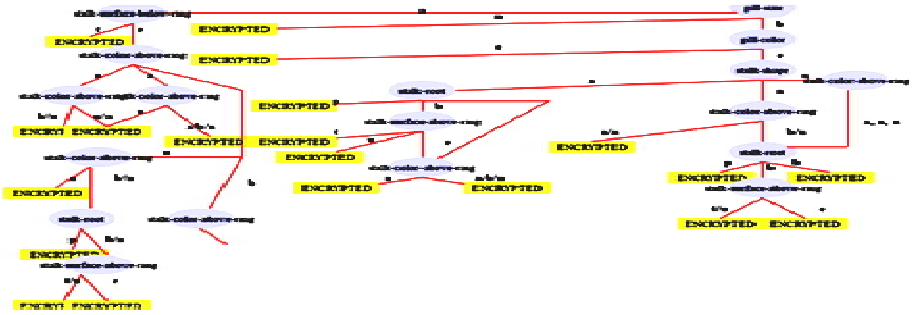

Fig 2: Snapshot of encrypted leaf nodes

C. *New Instance Classification*

In this module when a new instance classification request comes all the sites communicate securely by passing a token between them and return class distribution vector.

IV. PSEUDO CODE

In proposed algorithm there are two algorithm are used one is random decision tree using ID3 algorithm for tree generation and Two – Fish algorithm a cryptographic technique for encryption/decryption of data.

Algorithm Random Decision Tree(RDT)

Input: D, the training set, and
X, the set of attributes.
Output: A random decision tree R

R= **BuildTreeStructure**(X)
UpdateStatistics(R,D)
Prune subtrees with zero counts
return R

***Subroutine* BuildTreeStructure**(X)

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. 4, Issue 5, May 2016

if X= 0 **then**
return a leaf node
else
Randomly choose an attribute *F* as testing attribute
Create an internal node *r* with *F* as the attribute
Assume *F* has *m* valid values
for *i*= 1 to *m* **do**
c_i = **BuildTreeStructure**(*X*-{*F*})
Add c_i as a child of *r*
end for
end if
return *r*
Subroutine **UpdateStatistics(*r*,D)**
for each *x* in D **do**
AddInstance(*r*,*x*)
end for
Subroutine **AddInstance(*r*,*x*)**
If *r* is not a leaf node **then**
Let *F* be the attribute in *r*
Let c represent the child of *r* that corresponds to the value of *F* in *x*
AddInstance(*c*,*x*)
else
/* *r is a leaf node* */
Let t be the label of x
Let α[t] = # of t-labeled rows that reach r
α[t] <- α[t]+1
end if

Two-Fish algorithm

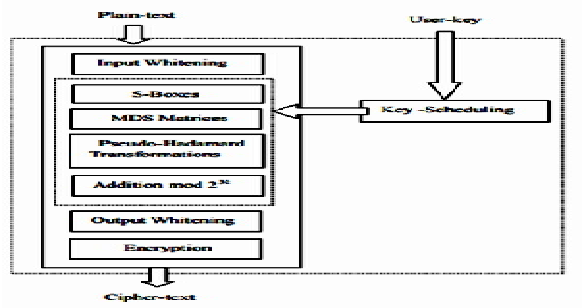

Fig 1: Steps of two-fish algorithm

V. RESULTS ANALYSIS

From the table 1 it shows the proposed system reduces the time complexity as compared to existing RDT framework.In existing system RDT is generated by choosing attributes randomly and tree stops growing when tree level reached half of no of attributes while in proposed system RDT is generated using ID3 algorithm and tree stops growing when purity level reached.Due to this time complexity of proposed system is less.

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. 4, Issue 5, May 2016

Input data	Existing system Execution time (1000s)	Proposed system Execution time (1000s)
Data 1	1133	538
Data 2	1176	527
Data 3	1138	428
Data 4	1185	417
Data 5	1291	431
Data 6	1405	467

Table1: Comparison of Execution time in milliseconds

From fig 3 the graph shows the execution time of existing and proposed system in vertical partition data.In vertical partition data different parties have different attributes.To gain some information from vertical partition data securely construct RDT without violating data privacy.

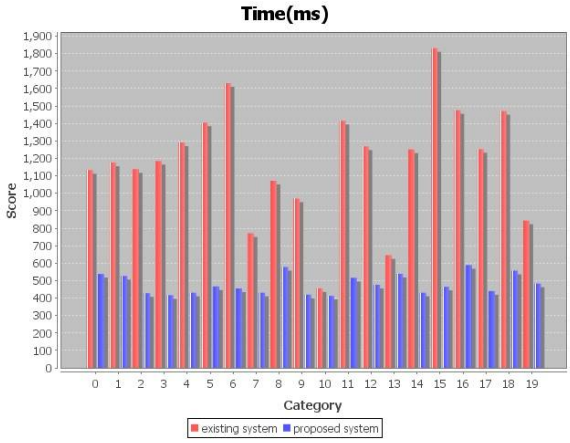

Fig 3: Graph for Comparison of Execution time in milliseconds

VI. Conclusion and Future Work

Today privacy of data for different organizations are very important because to expand their business they have to share information without loosing their privacy for sensitive data.Here randomization and cryptographic techniques are applied to the sensitive data .ID3 algorithm is used for decision tree generation and Two – fish algorithm is used for securing the private data.This algorithm provides high security to the data while maintaining purity or accuracy of the data.This paper deals with the horizontal and vertical partitioning of data.
In future this work is extended for arbitrarily partitioned data

References

[1] G. Jaideep Vaidya, Senior Member, IEEE, Basit Shafiq, Member, IEEE, Wei Fan, Member, IEEE, Danish Mehmood, And David Lorenzi "ARandom 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] Hemlata B. Deorukhakar1 , Prof. Pradnya Kasture" Adaptive Random Decision Tree: A New Approach for Data Mining with Privacy Preserving", Vol. 3, Issue 7,pp. 6378-6384, July 2015.

[3] Sathya Rangasamy,P.Suvithavani,"Privacy preserving On Continous and Discrete Data sets – A Novel Approach",IJMER,Vol.3,Issue.2, pp-809-815, March-April,2013.

[4] 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.

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. 4, Issue 5, May 2016

[5] Jintu Ann John, Neethu Maria John,"Privacy Preserving Random Decision Trees over Randomly Partitioned Dataset",vol.3,Issue 8, pp. 7746-7750,2015.
[6] Pui K. Fong and Jens H. Weber-Jahnke,Senior Member, IEEE Computer Society," Privacy Preserving Decision Tree Learning Using Unrealized Data Sets", IEEE transactions onknowledge and data engineering, vol.24, No.2 ,February 2012 .
[7] D. Agrawal and C.C. Aggarwal, "On the Design and Quantification of Privacy Preserving Data Mining Algorithms," Proc. 20thACM SIGACT-SIGMOD-SIGART Symp. Principles of Database Systems, pp. 247-255, May 2001.
[8] J. Vaidya and C. Clifton, "Privacy-Preserving k-Means Clusteringover Vertically Partitioned Data," Proc. Ninth ACM SIGKDD Int'lConf. Knowledge Discovery and Data Mining, pp. 206-215, Aug. 2003
[9] Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining,"J. Cryptology, vol. 15, no. 3, pp. 177-206, 2002.
[10] K. Wang, Y. Xu, R. She, and P.S. Yu, "Classification Spanning Private Databases," Proc. 21st Nat'l Conf. Artificial Intelligence,pp. 293-298, 2006.
[11] W. Du and Z. Zhan, "Building Decision Tree Classifier on PrivateData,"Proc. IEEE Int'l Conf. Data Mining Workshop on Privacy, Security, and Data Mining, pp. 1-8, Dec. 2002.
[12] J. Vaidya, M. Kantarcioglu, and C. Clifton, "Privacy Preserving Naive Bayes Classification,"Int'l J. Very Large Data Bases, vol. 17,no. 4, pp. 879-898, July 2008.
[13] R. Wright and Z. Yang, "Privacy-Preserving Bayesian Network Structure Computation on Distributed Heterogeneous,"Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, Aug. 2004.
[14] M. Kantarcio^ glu 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.
[15] M. Kantarcioglu and C. Clifton, "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data,"Proc. ACM SIGMOD Workshop Research Issues on Data Mining and Knowledge Discovery (DMKD '02), pp. 24-31, June 2002.
[16] X. Lin, C. Clifton, and M. Zhu, "Privacy Preserving Clustering with Distributed EM Mixture Modeling,"J. Knowledge and Information Systems, vol. 8, no. 1, pp. 68-81, July 2005.
[17] J. Vaidya and C. Clifton, "Privacy-Preserving k-Means Clustering over Vertically Partitioned Data," Proc. Ninth ACM SIGKDD Int'lConf. Knowledge Discovery and Data Mining, pp. 206-215, Aug. 2003.
[18] G. Jagannathan and R.N. Wright, "Privacy-Preserving Distributed k-Means Clustering over Arbitrarily Partitioned Data," Proc.ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 593-599, Aug. 2005.
[19] J. Vaidya and C. Clifton, "Privacy Preserving Association Rule Mining in Vertically Partitioned Data,"Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 639-644, July 2002.

BIOGRAPHY

Rimi Kumari is a M.E Student in the Computer Science Engineering Department, Dr. D.Y. Patil School of Engineering, Lohgaon, Pune, India. Her research interests are Data Mining and Network Security.

Mr. Soumitra Das is HOD in Dr. D.Y.Patil School of Engineering, Lohgaon, Pune, India. Currently, he is PhD researcher at Sathyabama University, Chennai, India.His research interest includes Computer Networks,Wireless Sensor Networks, etc. He is a members of IEEE, CSI, LMISTE, IACSIT and IAENG. He is also an active reviewer of various conferences and jo m m