International Journal of Research Studies in Computer Science and Engineering (IJRSCSE)
Volume 1, Issue 1, January 2014, PP 76-82
www.arcjournals.org

A Framework for Privacy Preserving Data Mining

K. Srinivasa Reddy[1], N. Rajasekhar[2]

[1]Associate Professor, Department of Computer Science and Engineering, VNRVJIET, Hyderabad, India

[2]Assistant Professor, VNR VJIET, Hyderabad, India

Abstract: *Distributed data is universal in present day in- formation driven applications. As there are n number of sources for data today, the natural difficulty is to identify how to combine the data more effectively upon Organizational boundaries while using the data to maximum. As utilizing only the local data gives suboptimal use, different techniques must be developed for privacy preserving collaborating knowledge discovery. For the large scale data sets the existing system i.e., cryptography based function for privacy preserving data mining is very slow and effective to face present day big data challenge. Earlier work on random decision trees shows that it is possible to develop more accurate and effective with less cost. As studied we can conclude that RDTs can naturally accept a parallel and fully distributed architecture and accordingly generate protocols to implement privacy preserving RDTs that ensure basic and efficient distributed privacy preserving knowledge discovery.*

Keywords: *Random decision trees; Encryption; partitioning; Accuracy; Security; Privacy Preserving*

1. INTRODUCTION

Construction and implementation of any data mining model basically is assumed that it can be accessed freely but this isn't practically observed. Security and privacy maintenance are the major concerns and they restrict sharing of any centralized data. Therefore, a prosperous method ended up to solve this problem of privacy preserving data mining. There are distributed solutions proposed to preserve privacy while accessing the data mining. Today we are here to propose a solution that uses cryptography and random decision tree construction methods to provide better efficiency and security to tree based tasks.

Cryptography approaches in real are often very slow when observed practically and can become computationally expensive as the inter communications between different parties increase and the size of the data set increases. Random decision tree is a commonly used approach where same code applies for declaration and classification. Hence using the similar technique we can solve the different problems in the same framework, but if disturbances are not controlled with care the necessity of information gathered can be dissolved, or conversely, there might be a chance in the leakage if information itself is not perturbed properly.

Computer science and statistics are a large portion of studies with different fields where Data mining is one of the interdisciplinary sub-field with a aim is to mine the information using different intelligent methods from a dataset and convert or transfer them into an inclusive structure for the future use. Random decision tree is a process where it randomly selects the rows or observations and variables to construct multiple decision trees and then selects the best decisions i.e. which gives the high accuracy among the forest. A decision tree is built for the entire data set using the variables of interest and based on the entropy calculations. Here a Random Decision tree is builds n number of decision trees and then averages them for accurate and approximate high levels of management.

Existing framework is ID3-based conventions. Our fundamental commitment is to understand that Random decision trees or RDT's can furnish great security with high productivity. We center around characterization for simplicity of conversation, the essential issue in conveyed order is to prepare a classifier from the dispersed information and afterward arrange each new case and thus protecting the privacy worry. Protection and security concerns limit the sharing or centralization of information. Security protecting information mining has developed as a compelling.

The solution we propose here is that, both randomization and cryptographic techniques are used for providing better quality and more efficient and security for learning tasks of several decision trees. In fact, the proposed solution provides more order of magnitude improvement in efficiency and security

with cost efficient over the existing solutions. This is an essentially required solution for privacy-preserving data mining for the big data challenge or data science challenge. The performance of proposed protocols is improved immensely compared to other techniques and also to implement and analyze these protocols the communication cost and computation is effective. It is more secured compared to other processes.

2. LITERATURE SURVEY

As per data mining model, this is the time to figure out how the Reliable Data Transfer can be built and its classification to be performed, to have data for all the attributes; if data is horizontally partitioned to gather for variety of entities by all parties. Because all these parties share the logical description accordingly. When these are put together they form the group of random trees. Presumably individuals alone are able to design the structure of the tree as their own way separately. According to global data set, all parties should compute the parameters in a co-operative and secure manner i.e., values of every terminal node. Here, there are two possibilities: 1) Participant is known about the structure of tree. 2) Participant is unknown about the structure of tree. We could see three possibilities in Case 1 as categorized below: The performance of proposed protocols is improved immensely compared to other techniques and also to implement and analyze these protocols the communication cost and computation is effective. It is more secured com- pared to other processes. (1a) all parties can have a chance to know about the universal class distribution vector for each leaf node. (1b) Individual alone can know which is having their own tree structure. (1c) None of the Parties have idea on this. Of course the Case 2 is apparently more complicated than Case 1, since it doesn't make any sense in this context which it causes a problem since the other parties can no longer compute the local leaf values. Every party should be collaborated with the tree owner in some way to find out the leaf node values accordingly.

To start with, as the pattern is known to everybody, and the structure of tree is arbitrary, everybody has their imagination for that specific structure. In fact, it would be very smarter to get rid of irregular structures that might be unsuitable to certain gatherings because of security concerns. Besides, regardless of whether the structure is obscure, each grouping of another case can uncover some information about the tree. For instance, if various new examples are ordered to a similar leaf hub, it is anything but difficult to make sense of (or if nothing else limited down) what the structure of the branch is?. At long last, despite the fact that this might be conceivable from the safe calculation point of view, it is probably going to refute the productivity preferred position of RDTs, consequently disposing of the principle explanation behind picking the arbitrary tree approach in any case.

Here, all the individual users know the structure of all the trees. If an individual thinks a tree is not secure, they can reject the tree and check for other options with the help of several methods available to us. This process is done through a voting process where all the participants accept a tree. Even if a single participant rejects a tree other participants will know that it is rejected by someone else. In the adverse situations of two participants only one will guess the other party feels the tree as a breach. But the path of the breach in the tree is not known to others. If other individuals can also detect the path of the breach this leads to a larger data leakage and thus our data will be at risk. Once all the individuals accept a tree the execution of terminal values proceeds. For this all the users compute the leaf node values. All the leaf values are added up to attain the global tree values. Obtaining the global tree values is done utilizing the process secure sum protocol. Now all the users will have the tree pattern and can be used locally. Therefore it does not require classifying of any new instance. Weka created at the College of 'Waikato in New

Zealand', is an assortment of AI calculations for information extracting errands actualized in Java. Aside from giving calculations, it is a basic execution system, alongside help classes and documentation. It is extensible and advantageous for prototyping presents. In any case, the wekaframe work is a brought together framework intended to be utilized at a solitary site. Protection Saving Information Mining with Irregular choice tree system.

3. METHODOLOGY

Now-a-days multiple parties use same data to identify class name of their data and if we expose all data to all parties then privacy will be at risk. For example, multiple parties such as bank, insurance company or credit card companies will use same records but for different purposes. Bank will use it to

find past transaction. Credit card will use data attributes related to pass payment. Insurance Company will use to identify correct policy for that person.

These different companies will make access of our personal information which is private but with different at- tributes. So if this data is viewed by every company then the privacy will be at great risk. So, to overcome from privacy issue data mining algorithm called Random Decision Tree is built, by this algorithm data is randomly selected and homomorphism encryption is applied to provide privacy to user's data. All companies know only about the class name and dataset will be partition based on the company requirement. The partitioned dataset helps in construction of random decision tree. Classification model which is also random decision tree algorithm is given a dataset to build random decision tree. Main idea is to compare the accuracy rate of random decision tree when compared to other processes.

A. Modules

ARFF records have two particular segments. The primary segment is the Header data, which is followed the Information data. The Header of the ARFF document contains the name of the connection, a rundown of the properties (the sections in the information), and their sorts. The information that is gathered from the field contains numerous undesirable things that prompts wrong examination. For instance, the information may contain invalid fields, it might contain sections that are unessential to the present investigation, etc. Therefore, the information must be partitioned and preprocessed to meet the necessities of the sort of examination you are looking for. We used horizontal partitioning algorithm in this.

A homomorphic cryptosystem resembles different types of open encryption in that it utilizes an open key to encode information and permits just the person with the coordinating private key to get to its decoded information (however there are likewise instances of symmetric key homomorphic encryptiontoo).

B. Umldiagrams

Let's see how the design phase of the Application. To understand the project, we use the UML diagrams. Let's say component diagram. Component Diagram also called as an UML component graph, depicts the association and wiring of the physical segments in a framework. Component diagrams in general follow abstraction property that is they hide the implementation details of a framework and check repeatedly if it every required aspect of a system is included and secured for improvement of system or not.

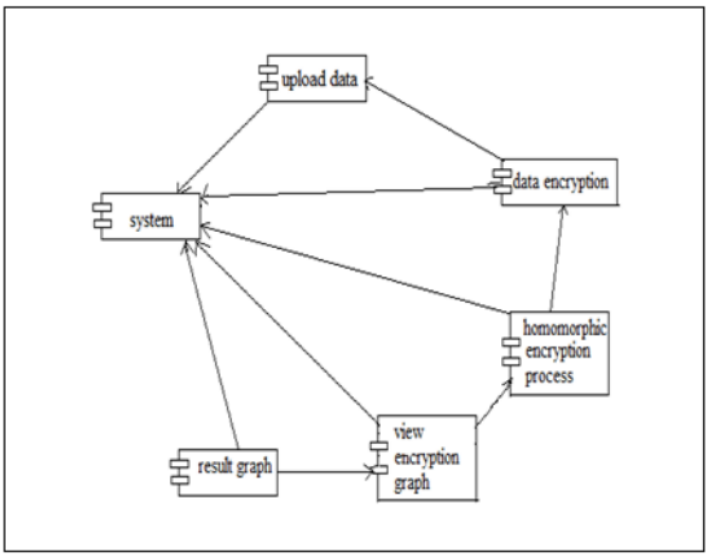


Fig. 3.1 Component Diagram for encrypting data The following steps are shown in the component diagram 1. First, the required dataset is uploaded and is under the control of system. 2. Then necessary action can be performed as per the choice either encrypting the data or constructing the decision trees as required. 3. We can view the data which is encrypted and it is confined only to the controller. 4. Result is viewed in the form of graph in comparison with different methods.

C. How does Performing Random Decision Tree Look like

Data mining is utilized to find information by utilizing existing or past information and new information class can be discovered by applying it to existing utilizing the characterization method. Presently a-days different gatherings utilize the same information to recognize the class name of their

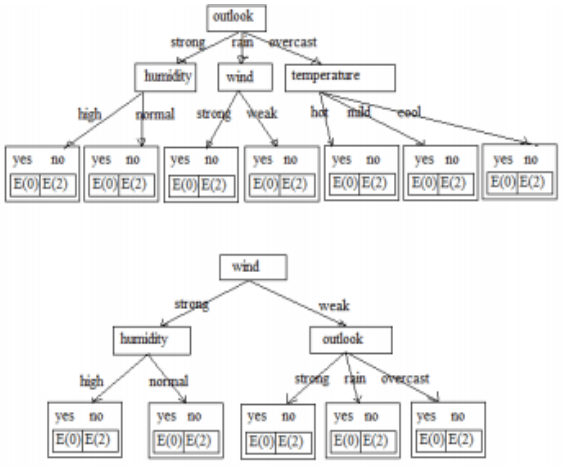


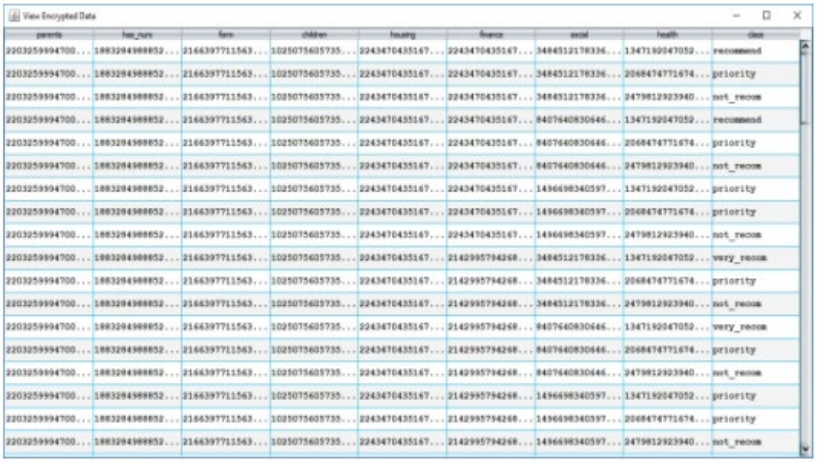
Fig.3.2 Horizontal partitioning on random decision tree Here we can observe the random decision of tree built using the horizontal partitioning of data. In the upper figure the data is understood to every individual and the details are clear which puts our data at risk. So, random branches are selected and encryption is performed where the exact details of the tree are hidden and not understood to the individual. That is how a random decision tree is constructed.

4. RESULTS AND DISCUSSION

We used Weka tool to describe about the entire process being carried out and easy understanding. Double click on 'run.bat' file to get below screen. To implement this along with WEKA tool we also used java API to develop our project.

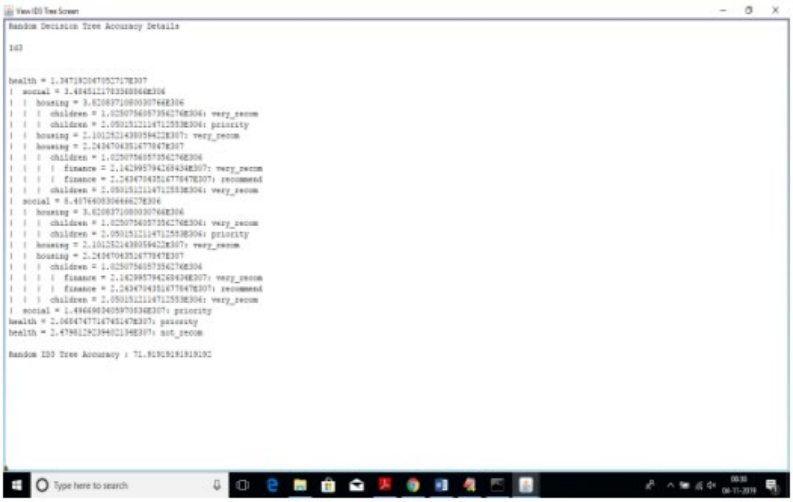
In above screen click on 'Upload Dataset' button and upload any dataset. As we used only two datasets in this system "mushroom" and "nursery" data will appear as below.

In above screen we can see entire dataset records in plain format, this can be easily understood to everyone. In order to provide protection to our data Homomorphic encryption is performed. Now click on 'View Encrypted Data' to view the data which is encrypted. Homomorphic encrypted data is seen as below.

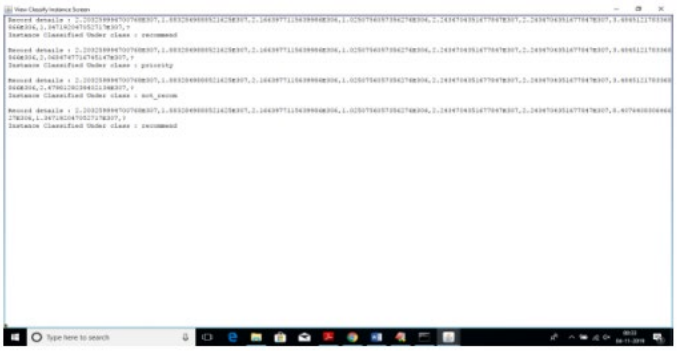


In above screen we can see all records are encrypted and only class name which are in last column are shown to parties. With this encrypted data nobody can understand anything. Now to build tree on this encrypted data click on 'Run Random Decision Tree' button to build tree.

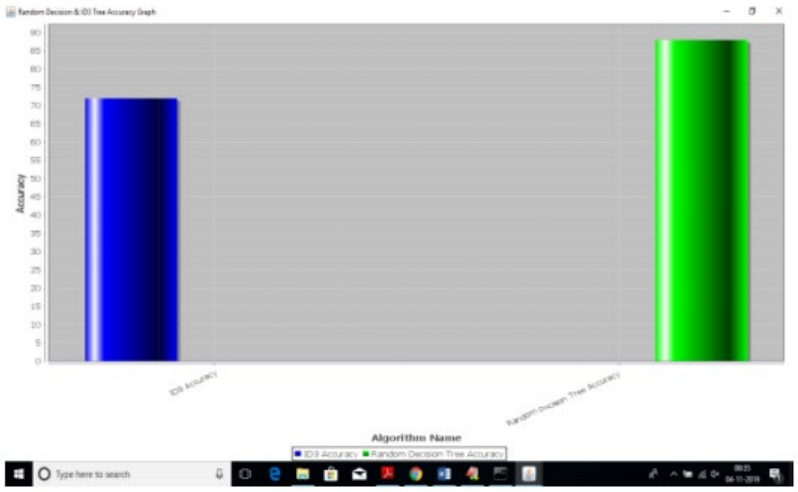
In above screen we can see tree generated by random decision and all nodes contains encrypted data and this tree got accuracy as 87perentage. In last line we can see accuracy. Now we will compare it with another algorithm. Now click on 'Build ID3 Tree' button to generate tree with ID3technique.



In above screen we can see ID3 tree also but its accuracy is 71Now click on 'Classify Instance' button to upload test file and get prediction or classification result. Here if u build decision tree with NURSERY dataset then upload nursery test dataset only. The dataset uploaded should be tested accordingly.



In above screen each records contains '?' at last column and in next line application has given or predict it class name. For example in above screen in first record is classified as 'recommend'. Now click on 'Random Decision ID3 Tree Accuracy Graph' button to get below accuracy graph of both algorithms.



In above graph x-axis represents algorithm name and y- axis represents accuracy of those algorithms. As shown above random decision tree has more accuracy than id3 tree algorithm.

5. CONCLUSION

We have shown that general and effective appropriated protection saving information revelation is really achievable. We have considered the security and protection suggestions when managing disseminated information that is divided either evenly or vertically over different destinations, and the trouble in performing information mining assignments on such information. As RDTs perform to create similar, exact and now and again better models with a lot littler cost, we have come up with a disseminated protection saving RDTs. This is our proposed system. Our methodology use the way that arbitrariness in structure can furnish solid protection with less calculation. The analyses results that the security safe- guarding adaptation of the RDT calculation scales directly with informational collection size, and requires altogether less time than elective cryptographic methodologies. Later on, we intend to create general arrangements that can work for self-assertively parceled information. You can in any case approve the estimations of encrypted fields utilizing approval rules or Summit. Both work whether or not the client has the "View Encoded Information" consent.

REFERENCES

[1] J. Vaidya, C. Clifton, and M. Zhu, Privacy-Preserving Data Mining.ser. Advances in Information Security first ed., vol. 19, Springer-Verlag, 2005.

[2] 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.

[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 Int'l 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 Int'l Conf. Data Mining (SDM '10), pp. 778-789, 2010.[5] A. Dhurandhar and A. Dobra, "Probabilistic Characterization of Random Decision Trees," J. Machine Research, vol. 9, pp. 2321-2348,2008.

[5] R. Agrawal and R. Srikant, "Privacy-Preserving Data Mining," Proc. ACM SIGMOD Conf. Management of Data, pp. 439-450, May 2000.

[6] Z. Huang, W. Du, and B. Chen, "Deriving Private Information from Randomized Data," Proc. ACM SIGMOD Int'l Conf. Management of Data, June2005.