A Data Mining Approach for secure Cloud using Enhanced Random Forest

Shikha Pathania[1], Rajdeep Kaur[2]

[1] *Resarch Scholar ,Department of Computer Science and Engineering, Lovely Professional University*
Kharar, Jalandhar, India
[2] *Assistant Professor, Department of Computer Science and Engineering,Lovely Professional University*
Kharar, Jalandhar, India

***Abstract*— Data mining is the process of extracting and analyzing the large datasets to find out various hidden relationship patterns and much other useful information. Random forest is an ensemble method which is widely used is application having large datasets because of its interesting features like handling imbalanced data, identifying variable importance and detecting error rate. For building random forest randomness is established in two ways: Firstly by creating samples from original datasets randomly and Secondly at the time of creation of each tree, randomly selecting subsets of attributes at each node for best splitting decisions. But by using randomness in Random forests we are likely to have uninformative attributes which will lead to poor accuracy results and bad performance of the algorithm. In this paper we are providing an improved Feature selection Random Forest that improves the performance of the algorithm in terms of accuracy. In this first we are selecting the good features by applying the consistency on attributes after that we are combining this consistency based feature with the Random forest. Also most of the organizations today are moving towards the cloud computing services, so we are performing the mining operation on the cloud based data. To protect the data from the unauthorized user we are securing the cloud data using AES algorithm through this no unauthorized user can access the data.**

***Keywords*— Data mining, cloud computing, Feature selection.**

I. Introduction

Data mining is the process of analysing large data sets to find out various hidden and useful relationships and patterns. There are various machine learning algorithms in the data mining. Machine learning algorithms are those techniques that are applied to datasets to analyse and then acquiring new knowledge. Random forest is one of the machine learning algorithms that is used for prediction and classification purpose. Various studies have been done on the RF algorithm that shows that it gives best result for regression and classification problems. This algorithm is known to be one of the best algorithms as it provides best accuracy among the entire classification algorithm. Various researches have also been done to show that it has one of the best accuracy among the other algorithms. So in this paper the aim is to increase the accuracy of RF and also use the concept of cloud computing to secure the data from unauthorised users. [20],[21]

Recently organisations are using Cloud services for their business. Cloud computing is the process of providing on demand services over the internet.

But cloud is insufficient to analyse the data that is why we are using enhanced RF algorithm. By combining data mining and cloud computing organisation can have maximum profit. In RF there is the concept of randomness in this. But when we are selecting the attributes randomly there are chances that we can miss the important features and include uninformative attribute. This situation will lead to the formation of bad trees which in turn will affect the performance and accuracy of the algorithm.

This paper is organised in the following way: section II gives the brief introduction of the methods that has been used in this work. Section III describe the framework of the proposed approach, Section IV tells about the data source, Section V shows the results and discussion and section VI concludes the work.

II. Methods And Approach Used In The Research Work

In this RF and cloud computing methods are used along with the Feature selection approach that is used for improving the RF algorithm.

A. Random Forest

Random forest is a supervised machine learning algorithm originally proposed by Leo Breiman in 1999. It is an ensemble method in which multiple trees are used as base classifiers and the classification with the majority of votes by each tree is chosen. This algorithm develops set of trees by random selection of data or by random selection of variables. Since the trees in this are generated randomly and you have many trees it is called random forest algorithm[2].

Following are three steps that are use for building RF :

Step1:Create Tn bootstrap samples randomly with replacement from the original dataset T. Let us assume that the number of attributes in this dataset is M.

Step2: From each bootstrap sample generate a tree such that at each node rather than choosing the best split from all the features, select random subset of attributes. The number of attributes to be chosen randomly is m, where m =√M.

Step3: Classify the data by aggregating the classification of each tree. Means the class with the majority of the votes will be the classification result.

B. Cloud computing

Cloud computing refers to the process of providing on demand services over the internet. In this the data is kept on the internet instead of your hard drives. In cloud computing the hardware and software are provided to each individuals and business organizations. These hardware and software are kept on the remote locations and are managed by the third parties. Since the data is managed by the third parties concept of privacy is major concerns in this. This model allows you to access the resources from anywhere if there is a network connection. Some examples of cloud computing are emails, social networking sites and online file storage.

There are three service models in the cloud:

- IAAS: (Infrastructure as a service) provider provides only network and hardware services.
- PAAS:(Platform as a service) provider provides application development environment.
- SAAS:(Software as a service) provider provides the application running on the cloud.

C. Feature Selection approach

Feature selection is the process of evaluating each feature and then finding the best feature subset that is relevant to the class. There are three types of methods in feature selection: Filter, wrapper and embedded method. In filter method the features are selected on the basis of the various statistics such as consistency. In wrapper methods the importance of the features are calculated by evaluating the models. And lastly in embedded methods the features selection is incorporated in the model building process[25].

III. Framework Of The Proposed Method

In this section we describe the step by step working of our proposed approach. Fig: 1 shows the overall design of our approach.

The approach is explained through following steps:

Step1: Data is partitioned into three parts and are then encrypted using the AES algorithm in order to provide security on the data.

Step2: These encrypted files are then store in the cloud environment. Single cloud is used in this environment but the cloud is partitioned into three parts each one storing a single encrypted file.

Step3: After this the data is decrypted to apply Random forest algorithm.

Step 4: Now the improved algorithm is applied on the same data. In the Enhanced Forest algorithm Feature selection is integrated with the algorithm.

Step5: Both RF and Enhanced RF algorithm is compared on the basis of some parameters like FP-measure, precision, recall, TP rate, FP rate.

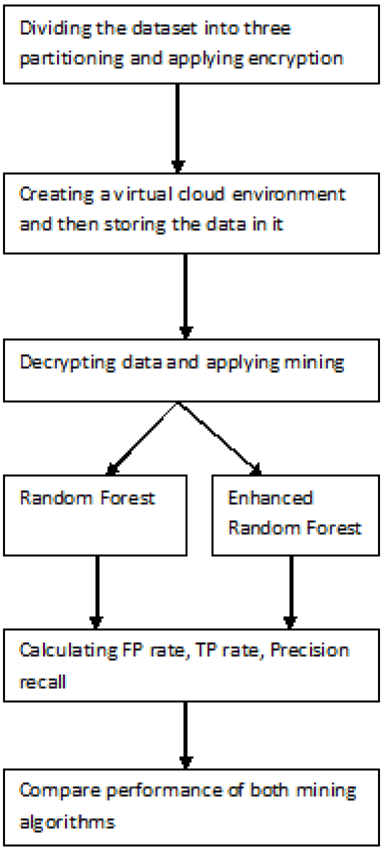

Dividing the dataset into three partitioning and applying encryption
Creating a virtual cloud environment and then storing the data in it
Decrypting data and applying mining
Random Forest
Enhanced Random Forest
Calculating FP rate, TP rate, Precision recall
Compare performance of both mining algorithms

Fig 1: Basic design of proposed method

A. Consistency based features selection

In our work we are using consistency based feature selection. This means that we find the optimal subset on the basis of the consistency of the subset. The feature subset with best consistency is selected. Then we apply the RF on that feature subset. By creating the trees with these features it will give us best accuracy result because we are performing the operation on those features which are related to the class label and uninformative features will not be used.

The consistency of feature subset is measured by calculating the inconsistency rate. Two patterns of feature subset is said to be inconsistence if all of their values are same except the class labels. The inconsistency count is the number of all matching instances minus the largest instances with different class labels. The inconsistency rate is the sum of all the inconsistency count divided by the total number of instances.

IV. Data Source

The data set used for our work is collected from the UCI repository. We are using ionosphere dataset in which radars returning from ionosphere are classified in good or bad classes.

V. Results

After performing both RF and Enhanced RF algorithm on the cloud based data we find out that the accuracy of our Enhanced RF is better than the RF algorithm. We have compared the accuracy of both algorithms on the basis of some parameters. Following tables show the comparison of both algorithms.

TABLE I
PARAMETER MEASURE FOR BOTH ALGORITHMS

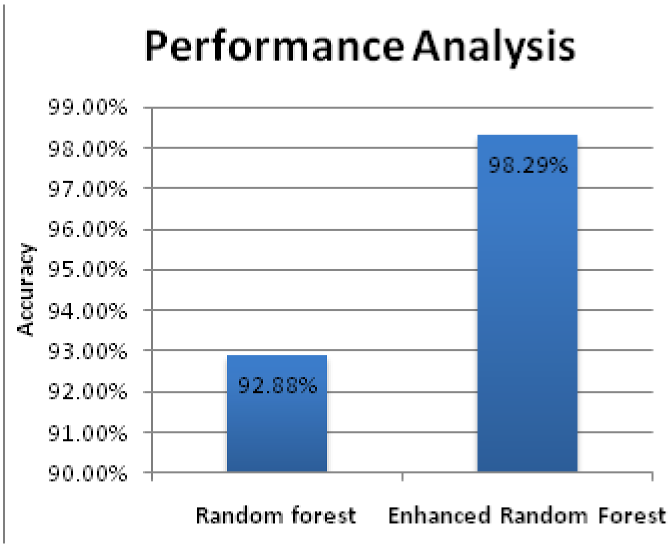
	Random Forest	Enhanced Random
Kappa Statistic	0.8439	0.9657
Mean absolute error	0.1211	0.0288
Root mean squared error	0.2377	0.1082
Relative absolute error	26.293	5.768
Root relative squared error	49.5448	21.6658

TABLE II
DETAILED ACCURACY MEASURES OF BOTH ALGORITHM

	Random Forest	Enhanced Random
TP Rate	0.929	0.983
FP rate	0.092	0.017
Precision	0.929	0.983
Recall	0.929	0.983
ROC area	0.973	0.999
F-measure	0.928	0.983

TABLE III
CORRECTLY AND INCORRECTLY CLASSIFIED INSTANCES

	Correctly classified	Incorrectly classified
Random Forest	326	25
Enhanced Random Forest	345	6



VI. Conclusion

In this paper we have done data mining on the cloud based data. We have used a classification algorithm named Random forest and then also improve that algorithm by improving the feature selection of the RF. To protect our data we partitioned the data and then apply AES encryption on the data so if even some portion of the data is leaked it can't be understand by the intruder. To improve the feature selection of RF algorithm first we apply consistency on the features which will give us the optimal subsets of relevant feature after that we integrate this feature selection with the RF algorithm. After comparing the RF results on the cloud based data we find out that the accuracy of our improved RF is better than the original RF.

The future work can be done in improving the security of cloud data. Also now every other organizations our moving to cloud so further various mining can be performed in the cloud based data.

References

[1] Amjad Hussain Bhat, Sabyasachi Patra, Dr. Debasish Jena, (2013) "*Machine learning approach for intrusion detection on cloud*", International Journal of Innovation in Engineering & Management, Volume 2

[2] Andy Liaw and Matthew Wiener,(2002) "*Classification and regression by random forest*"

[3] Baoxun Xu, Xiufeng Guo, Yunming Ye, Jiefeng Cheng (2012)"*An improved Random Forest classifier for Text categorization.*", Journals of Computers, Volume 7

[4] Bhagyashri U. Gaikwad, P.P.Halkarnikar ,(2013)"*Spam E-mail detection by Random Forest algorithm.*", Computer Science & Technology, Department of Technology, Shivaji University, Kolhapur, Maharashtra, India.

[5] Cuong Nguyen1, Yong Wang1, Ha Nam Nguyen,(2013)"*Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic.*" School of Business and Administration, Chongqing University, Chongqing, China ,College of Technology, Vietnam National University, Hanoi, Vietnam.

[6] Diego Marron, Albert Bifet , Gianmarco De Francisci Morales ,(2014)"*Random Forest of very fast decision trees on GPU for mining evolving big data streams.*"

[7] D.L. Gupta, A.K.Malviya , Satyendra Singh (2012) "*Performance analysis of classification tree learning algorithms*.", International Journal of computer applications, Volume 55

[8] Florian Baumann, Fangda Li, Arne Ehlers, Bodo Rosenhahn ,(2014) "*Thresholding a Random Forest classifier*."

[9] Jehad Ali, Rehanulla Khan, Nasir Ahmad, Imraan Masqood(2012) "*Random Forest and decision trees.*", Computer Systems Engineering, International Journal of Computer Science Issues, volume 9.

[10] Kashish Ara Shakil, Mansaf Alam,(2013) "*Data management in cloud based environment using K-Median clustering technique*", International journal of Computer applications

[11] Ke Sun, Wansheng Miao, Xin Zhang, Ruonan Rao, (2014) "*An improvement to feature selection of Random forest on Spark*", IEEE 17th International conference on Computer Science and Engineering.

[12] Kilho Shin, Danny Fernandes and Seiya miyazaki, (2011) "*Consistency measures for feature selection: A Formal Definition, relative sensitivity comparison and a fast algorithm*" International joint conference on Artificial intelligence

[13] Manoranjan Dash, Huan Liu,(2003) " *Consistency based search in Feature selection*" Elsevier Computer science

[14] Mohamed Bader-El-Den , Mohamed Gaber,(2012) "*GARF: Towards self –optimized Random Forest*.", School of Computing,

[15] S.Bharathidason, C.Jothi Venkataeswaran, (2014) "*Improving Classification accuracy based on Random Forest model with uncorrelated high performing trees.*", International Journal of Computer Applications Volume 101- No.13, September 2014

[16] Shengqiao Li, E James Harner, Donald A Adjeroh ,(2011)"*Random KNN feature selection- A fast and stable alternative to Random forest.*", BMC bioinformetics

[17] Sunita, Prachi, (2013) "*Efficient cloud mining using RBAC concept*", International Journal of Advanced Research in Computer Science and Software Engineering.

[18] Supraja.Y, T.V.Sai Krishna, P.VenkatasubbaReddy, Dr. M.A.D.Swamy, Dr.P.Srinivasulu,(2013) "*Random Forest machine learning algorithm to detect abnormal behavior in cloud-based mobile services.*" , International Journal of Computer Science and Information Technology and Security, Vol 3.

[19] Thanh-Tung Nguyen, Joshua Zhexue Huang, and Thuy Thi Nguyen,(2014) "*Unbiased Feature Selection in Learning Random Forests for High – Dimensional Data*", Hindwai Publishing corporation The scientific world journal

[20] Vrushali Y Kulkarni, Pradeep K Sinha, (2014) "*Effective learning and classification using Random Forest algorithm*", International Journal of Engineering and Innovative Technology(IJEIT) Volume 3, Issue11, May 2014.

[21] Vrushali Y Kulkarni, Pradeep K Sinha,(2013) "*Efficient learning of Random Forest Classifier using disjoint partitioning approach.*" Proceeding of the world Congress on Engineering, London , U.K

[22] Vrushali Y Kulkarni, Aashu Singh, Pradeep K Sinha,(2013) "An Approach towards Optimizing Random Forest using Dynamic Programming Algorithm," International Journal of computer applications ,Volume75

[23] Xiao Liu, Mingli Song, Dacheng Tao ,Zicheng Liu,Chun Chen and Jiajun Bu."*Semi-Supervised Node Splitting for Random Forest Construction*." IEEE

[24] YIN XiaoHong, DIAO Zhijian, (2014) "*Research and implementation of the data mining algorithm based on cloud platform*", 2014 IEEE workshop on Electronics, Computer and applications.

[25] Manoranjan das, Huan liu "*Consistency based search in feature selection"* , 2003