Study of Random Tree and Random Forest Data Mining Algorithms for Microarray Data Analysis

[1]Ajay Kumar Mishra, [2]Bikram Kesari Ratha

[1,2]PG Department of CSA, Utkal University, Bhubaneswar, India

Abstract : The Weka workbench is an designed set of state-of-the-art machine learning techniques and data pre-processing tools. The primary way of relating with these methods is by calling up them from the command line. However, suitable interactive graphical user interfaces are provided for data exploration, for setting up large-scale test on distributed computing platforms. Classification is an significant data mining technique with major applications. It classifies data of different kinds. Random forests are a mixture of tree predictors such that every tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. This paper have been carried out to make a performance evaluation of Random Forest and Random Tree classification algorithm. The paper sets out to make comparative evaluation of classifiers Random Forest and Random Tree in the context of Microarray dataset. For processing Weka API were used. The data set from the UCI Machine learning Repository is used in this experiment.

Keywords— Random Tree, Random Forest, Weka, Pre-Processing

I. INTRODUCTION

Microarray technology has attracted increasing research interest in the modern years. It is a promising tool to simultaneously observe and measure the expression levels of thousands of genes of an organism in a sole experiment[1,3,5,7]. Basically a microarray is a glass slide that includes thousands of spots. Each spot may contain a few million copies of identical DNA molecules that uniquely correspond to a gene. The amount of data in the world and in our lives seems ever-increasing and there’s no end to it. We are overwhelmed with data. Today Computers build it too easy to save things. Inexpensive disks and online storage make it too easy to postpone decisions about what to do with all this stuff. In data mining, the data is accumulated electronically and the search is automated or at least augmented by computer. Even this is not particularly new. Economists, statisticians, and communication engineers have long worked with the idea that patterns in data can be sought automatically, identified, validated, and used for prediction. What is new is the staggering increase in opportunities for finding patterns in data. Data mining is a topic that involves learning in a practical, non theoretical sense.

The rest of the paper is structured as follows. In the next section, Algorithm selected for comparison is described. Section 3 provides experimental design methodology and section 4 summarizes the results. Then the paper concludes in section 5.

II. ALGORITHM SELECTED FOR COMPARISON

2.1Random Forests

Random forests is a idea of the general technique of random decision forests that are an ensemble learning technique for classification, regression and other tasks, that control by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests accurate for decision trees' habit of overfitting to their training set.

The first algorithm for random decision forests was created by Tin Kam Ho using the random subspace method which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg. An extension of the algorithm was developed by Leo Breiman[7] and Adele Cutler,[8] and "Random Forests" is their trademark[9].The extension combines Breiman's "bagging" idea and random selection of features, introduced first by Ho[1] and later independently by Amit and Geman[10] in order to construct a collection of decision trees with controlled variance.

2.2 Random Tree

Random Tree is a supervised Classifier; it is an ensemble learning algorithm that generates lots of individual learners. It employs a bagging idea to construct a random set of data for constructing a decision tree. In standard tree every node is split using the best split among all variables. In a random forest, every node is split using the best among the subset of predicators randomly chosen at that node. Random trees have been introduced by Leo Breiman and Adele Cutler.The algorithm can deal with both classification and regression problems. Random trees is a group (ensemble) of tree predictors that is called forest. The classification mechanisms as follows: the random trees classifier gets the input feature vector, classifies it with every tree in the forest, and outputs the class label that received the majority of “votes”. In case of a regression, the classifier reply is the average of the responses over all the trees in the forest. Random Trees are essentially the combination of two existing algorithms in Machine Learning: single model trees are merged with Random

Forest ideas. Model trees are decision trees where every single leaf holds a linear model which is optimised for the local subspace explained by this leaf. Random Forests have shown to improve the performance of single decision trees considerably: tree diversity is created by two ways of randomization[4,6,11]. First the training data is sampled with replacement for each single tree like in Bagging. Secondly, when growing a tree, instead of always computing the best possible split for each node only a random subset of all attributes is considered at every node, and the best split for that subset is computed. Such trees have been for classification Random model trees for the first time combine model trees and random forests. Random trees uses this produce for split selection and thus induce reasonably balanced trees where one global setting for the ridge value works across all leaves, thus simplifying the optimization procedure. [2] [8] [9] [10]

III. EXPERIMENTAL DESIGN METHODOLOGY

In the Expt, we use Weka data mining tool to conduct the experiment. We compared the classification performance of the Decision tree i.e Random Forest and Random Tree models employing attribute selection filter. The Breast Cancer DataSet from UCI repository is used in this expt.

We use 10-fold cross validation as the test mode to record classification accuracy. This approach is suitable to avoid biased results and provide robustness to the classification. Also, the parameters of a classification algorithm are chosen to their default values.

The following steps have been applied to generate experimental data in order to draw inference:

1. Find classification performance of the classifiers in the dataset.

2. Find classification performance using Attribute selection filter.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

we performed several runs in Weka tool and gathered the data for the inference. Table-2 summarizes the classification accuracy in percentage of all the classifiers across the datasets with original features while Table-3 provides the classification performance after a genetic algorithm based feature selection. Table-4 shows the classification performance of the classifiers when feature selection is performed using PSO based search.

Table 1:Classification Accuracy

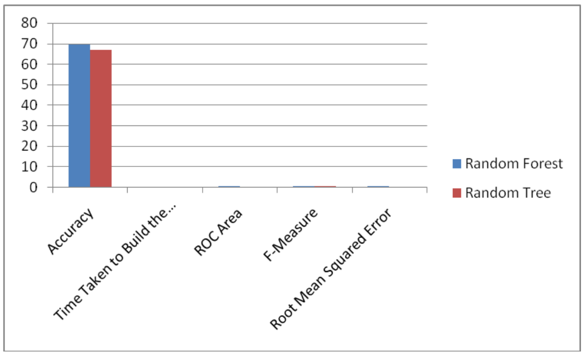
Classifier's Name	Accuracy	Time Taken to Build the Model (Seconds)	ROC Area	F-Measure	Root Mean Squared Error
Random Forest	69.58	0.22	0.637	0.67	0.45
Random Tree	66.78	0.05	0.588	0.664	0.56

Table 2:Classification Accuracy with Attribute selection Filter

Classifier's Name With Attribute selection Filter	Accuracy	Time Taken to Build the Model(Seconds)	ROC Area	F-Measure	Root Mean Squared Error
Random Forest	70.27	0.27	0.62	0.67	0.46
Random Tree	70.62	0.01	0.588	0.66	0.49

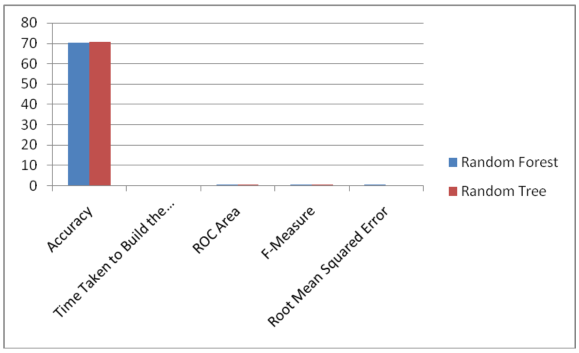
It is observed in the tabulated data that the performance of the Random Forest and Random Tree classifiers. This is depicted in Figure-2.

Figure2. Performance of Random Forest and Random Tree



It is observed in the tabulated data that the performance of the Random Forest and Random Tree classifiers with attribute selection filter. This is depicted in Figure-3.

Figure3. Performance of Random Forest and Random Tree with AttributeSElection



V. CONCLUSION

In this analytical study, we consider 2 popular decision tree classifiers with publicly available microarray datasets and one attribute selection filter for classification. Following a methodical approach, we gathered experimental data using Weka, a popular data mining tool. Based on data, it is found that the classification performance of the classifiers across the dataset is not very uniform to compare and draw straight forward conclusion on attribute selection filter techniques. Both the techniques compete with each other to provide classification accuracy across the dataset So, we lean towards statistical inference and find that attribute selection filter based Random Tree classification selection can produce slightly better classification accuracy than that of Random Forest.

REFERENCES

[1] Ho, Tin Kam (1995). Random Decision Tree (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.

[2] N. Landwehr, M. Hall, and E. Frank, ―Logistic model trees,‖ Mach. Learn., vol. 59, no. 12, pp. 161-205, 2005.

[3] An Implementation of ID3: Decision Tree Learning Algorithm Wei Peng, Juhua Chen and Haiping Zhou Project of Comp 9417: Machine Learning University of New South Wales, School of Computer Science & Engineering, Sydney, NSW 2032, and Australia.

[4] Wikipedia contributors, ―C4.5_algorithm,‖ Wikipedia, The Free Encyclopedia. Wikimedia Foundation, 28- Jan-2015.

[5] N. Landwehr, M. Hall, and E. Frank, ―Logistic model trees,‖ Mach. Learn., vol. 59, no. 1–2, pp. 161–205, 2005.

[6] L. Breiman, ―Random forests,‖ Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[7] Wikipedia contributors, ―Random tree,‖ Wikipedia, The Free Encyclopedia. Wikimedia Foundation, 13-Jul- 2014

[8] Breiman Leo (2001). "Random Forests". Machine Learning 45 (1): 5–32.

[9] Liaw, Andy (16 October 2012). “Documentation for R package random forest”. Retrieved 15 March 2013.

[10] U.S. trademark registration number 3185828, registered 2006/12/19.

[11] Amit, Yali; Geman,Donald (1997). "Shape quantization and recognition with randomized trees" (PDF). Neural Computation 9 (7): 1545–1588.