Comparative Study of J48, AD Tree, REP Tree and BF Tree Data Mining Algorithms through Colon Tumour Dataset

Abhaya Kumar Samal1 Subhendu Kumar Pani2 Jitendra Pramanik3
1Trident Academy of Technology, Bhubaneswar 2Orissa Engineering College, Bhubaneswar 3BPUT, Odisha, Bhubaneswar

Abstract—The Weka workbench is a 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 a significant data mining technique with major applications. It classifies data of different kinds. Decision Trees 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 has been carried out to make a performance evaluation of J48, ADTree, REPTree and BFTree classification algorithm. For processing Weka API were used. The data set from the UCI Machine learning Repository is used in this experiment.

Key words: Decision Tree, J48, AD Tree, REP Tree and BF Tree, 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, 2, 3, 6]. 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

In this section, we present features of various data mining algorithms for foregoing comparative study:

A. J48

J48 implements Quinlan’s C4.5 algorithm for generating C4.5 pruned or unpruned decision tree. An extension of Quinlan’s earlier ID3 algorithm is C4.5. The decision trees J48 can be used for classification. Using the information entropy, J48 builds decision trees from a labelled training data. It uses the fact that each data attribute can be used to make a decision by the data splitting into smaller subsets. J48 which is examines the normalized information gain or difference in entropy that results for splitting the data from choosing an attribute. Highest normalized information gain of attributes is used to make the decision and then the algorithm recurs on the smaller subsets of element. If all instances in a subset belong to the same class then after the splitting procedure stops. Leaf node in the decision tree telling to choose that class which are used. Sometimes happens that none of the features will give any information gain.

Random forests is an 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 [5] and Adele Cutler [10], and "Random Forests" is their trademark [8]. The extension combines Breiman's "bagging" idea and random selection of features, introduced first by Ho [1] and later independently by Amit and Geman [9] in order to construct a collection of decision trees with controlled variance.

B. Rep Tree

RepTree uses the regression tree logic and creates multiple trees in different iterations. After that it selects best one from all generated trees. That will be considered as the representative. In pruning the tree, the measure used is the mean square error on the predictions made by the 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 predictors 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 are 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, 5, 9]. 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, 5, 7, 8].

C. AD Tree and BF Tree

An alternating decision tree (ADTree) is a machine learning method for classification. It generalizes decision trees and has connections to boosting. An ADTree consists of an alternation of decision nodes, which specify a predicate condition, and prediction nodes, which contain a single number. An instance is classified by an ADTree by following all paths for which all decision nodes are true; and summing any prediction nodes that are traversed and Class for building a best-first decision tree classifier is known as BFTree.

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 J48, ADTree, REPTree and BFTree Models. The Colon Tumor DataSet from UCI repository is used in this expr having instances 62 and attribute 2001. 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.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

We performed several runs in Weka tool and gathered the data for the inference. Table-1 summarizes the classifier’s performance analysis.

<table>
<thead>
<tr>
<th>Classifier’s Name</th>
<th>Accuracy</th>
<th>Time Taken to Build the Model (Seconds)</th>
<th>ROC Area</th>
<th>F-Measure</th>
<th>Root Mean Squared Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48</td>
<td>82.25</td>
<td>0.42</td>
<td>0.749</td>
<td>0.819</td>
<td>0.414</td>
</tr>
<tr>
<td>AD Tree</td>
<td>74.19</td>
<td>0.82</td>
<td>0.825</td>
<td>0.731</td>
<td>0.444</td>
</tr>
<tr>
<td>BFTree</td>
<td>77.41</td>
<td>2.25</td>
<td>0.781</td>
<td>0.754</td>
<td>0.421</td>
</tr>
<tr>
<td>REPTree</td>
<td>69.35</td>
<td>0.19</td>
<td>0.649</td>
<td>0.651</td>
<td>0.464</td>
</tr>
</tbody>
</table>

Table 1: Classifier’s Performance Analysis

It is observed in the tabulated data that the performance of J48, ADTree, REPTree and BF Tree classifiers. This is depicted in Figure-1.

Fig. 1: Performance of J48, AD Tree, REP Tree and BF Tree

V. CONCLUSION

In this analytical study, we consider 4 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 straightforward conclusion and find J48 based Decision Tree classification selection can produce better classification accuracy than that of other decision tree algorithms.

ACKNOWLEDGMENT

The authors would like to thank anonymous reviewers for their valuable comments and suggestions.

REFERENCES