ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 1, January 2015

A Comparative Study on Decision Tree and Random Forest Using R Tool

Prajwala T R

Department of Information Science and Engineering, CMRIT, Bangalore

Abstract: Data mining is a process of extracting valuable information from large set databases. Classification a supervised technique is assigning data samples to target classes. This paper discusses two classification algorithms namely decision trees and Random forest.. Decision trees are powerful and popular tools for classification and prediction. Decision trees represent rules, which can be understood by humans and used in knowledge system such as database. Random forest includes construction of decision trees of the given training data and matching the test data with these. Rattle an open source R-GUI is used for analysis of weather data for prediction of rainfall using 256 data samples. Based on results obtained a comparative analysis is done.

Keywords: Classification, Decision Trees, Random Forest, supervised learning, confusion matrix, Entropy, Information Gain.

1. INTRODUCTION

Data mining or knowledge discovery is computer assisted process for analysis of data from perspectives. Data mining tools predict the behavior of data and help take knowledge driven decisions. In supervised learning, classes are predetermined. The classes are seen as a finite set of data. A certain segment of data will be labeled with these classification [1]. The task is to search for patterns and construct mathematical models. The training set consists of unlabeled data. The task of classification, which is one of supervised data mining technique, is to predict accurately the class to which the data samples belong to. For example consider weather data samples. Based on training dataset predict whether it is going to rain next day or not and accordingly classify into data labels named "yes" and "no". Random forests are an ensemble learning method used for classification [13]. The methodology includes construction of decision trees of the given training data and matching the test data with these. Random forests are used to rank the importance of variables in a classification problem.

2. DECISION TREES

Decision trees are powerful and popular tools for classification and prediction. Decision trees represent rules, which can be understood by humans and used in knowledge system such as database. Decision tree is a classifier in the form of a tree structure which consists of [5]
Decision node: specifies a test on a single attribute.
Leaf node: indicates the value of the target attribute.
Edge: split of one attribute
Path: a disjunction of test to make the final decision.
Decision trees classify instances or examples by starting at the root of the tree and moving through it until a leaf node. For the decision tree algorithm, ID3 (Iterative Dichotomiser 3) was selected as it creates simple and efficient tree with the smallest depth[3].

Methodology

The ID3 decision makes use of two concepts when creating a tree from top-down:
1. Entropy
2. Information Gain

Entropy is the measurement of uncertainty where the higher the entropy, then the higher the uncertainty[4].

$E(S) = -(p+)*\log2(p+) - (p-)*\log2(p-)$

"S" represents the set and "p+" are the number of attributes in the set "S" with positive values and "p -" are the number of attributes with negative values. Information Gain uses the entropy in order to determine what attribute is best used to create a split with.

$Gain(S, A) = Entropy(S) - S ((|Sv| / |S|) * Entropy(Sv))$

In the formula, 'S' is the set and 'A' is the attribute. 'SV ' is the subset of 'S' where attribute 'A' has value 'v'. '|S|' is the number of elements in set 'S' and '|Sv|' is the number of elements in subset 'Sv'.

The ID3 algorithm is as follows[6]
1) Establish Classification Attribute
2) Compute Classification Entropy.
3) For each attribute, calculate Information Gain using classification attribute.
4) Select Attribute with the highest gain to be the next Node in the tree (starting from the Root node).
5) Remove Node Attribute, creating reduced table R^S.
6) Repeat steps 3-5 until all attributes have been used, or the same classification value remains for all rows in the reduced table.

Calculate the entropy of every attribute using the data set say S[9]. Split the set S into subsets using the attribute for which entropy is minimum or, equivalently, information gain is maximum. Make a decision tree node containing that attribute. Recur on subsets using remaining attributes.

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 1, January 2015

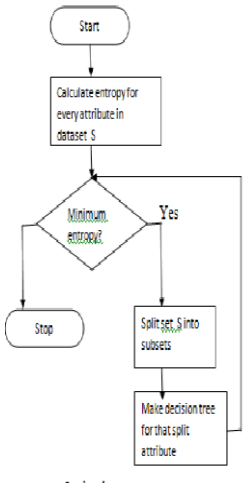

Figure 2.1: Diagrammatic representation of Decision Tree algorithm

Advantages of Decision trees[10]

1. Easy to interpret the decision rules
2. Nonparametric so it is easy to incorporate a range of numeric or categorical data layers and there is no need to select unimodal training data.
3. Robust with regard to outliers in training data.

Disadvantages of Decision Trees [10]

1. Decision trees tend to overfit training data which can give poor results when applied to the full data set.
2. Not possible to predict beyond the minimum and maximum limits of the response variable in the training data.

Application of Decision Trees

1. It is used in filtering of spam emails.
2. Decision tree is used in field of medicine. Ex: To predict the type of people prone to specific type of Virus.

3. RANDOM FOREST

Random forests are an ensemble method used for classification. The methodology includes construction of decision trees of the given training data and matching the test data with these. Random forests are used to rank the importance of variables in a classification problem.

To measure the importance of a variable in a data setDn={(Xi,Yi)}i=1n we fit a random forest to the data. During the fitting process the error for each data point is calculated and averaged over the forest.

To measure the importance of the i-th feature after training, the values of the i-th feature are permuted among the training data and the error is again computed on this data set. The importance score for the i-th feature is computed by averaging the difference in error before and after the permutation for all the trees. Normalization of the score is done by the standard deviation of these differences[10].

Features which produce large values for this score are more important than features which produce small values. Random forests provide information about the importance of a variable and also the proximity of the data points with one another.

Methodology:

Algorithm for Construction of Random Forest is:

Step 1: Let the number of training cases be ‘n’ and let the number of variables included in the classifier be ‘m’.

Step 2: Let the number of input variables used to make decision at the node of a tree be ‘p’. We assume that p is always less than ‘m’.

Step 3: Choose a training set for the decision tree by choosing k times with replacement from all ‘n’ available training cases by taking a bootstrap sample. Bootstrapping computes for a given set of data the accuracy in terms of deviation from the mean data. It is usually used for hypothesis tests. Simple block bootstrap can be used when the data can be divided into non-overlapping blocks. But, moving block bootstrap is used when we divide the data into overlapping blocks where the portion ‘k’ of overlap between first and second block is always equal to the ‘k’ overlap between second and third overlap and so on. We use the remaining cases to estimate the error of the tree. Bootstrapping is also used for estimating the properties of the given training data[2].

Step 4: For each node of the tree, randomly choose variables on which to search for the best split. New data can be predicted by considering the majority votes in the tree. Predict data which is not in the bootstrap sample. And compute the aggregate.

Step 5: Calculate the best split based on these chosen variables in the training set. Base the decision at that node using the best split.

Step 6: Each tree is fully grown and not pruned. Pruning is used to cut of the leaf nodes so that the tree can grow further. Here the tree is completely retained.

Step 7: The best split is one with the least error i.e the least deviation from the observed data set.

Advantages:

1. It provides accurate predictions for many types of applications
2. It can measure the importance of each feature with respect to the training data set.
3. Pairwise proximity between samples can be measured by the training data set.

Disadvantages:

1.For data including categorical variables with different number of levels, random forests are biased in favor of those attributes with more levels.

2. If the data contain groups of correlated features of similar relevance for the output, then smaller groups are favored over larger groups

Factors on which the construction of decision tree depends are[10]:

1.The shape of the decision to use in each node.
2.The type of predictor to use in each leaf.
3.The splitting objective to optimize in each node.
4.The method for injecting randomness into the trees.

DOI 10.17148/IJARCCE.2015.4142

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 1, January 2015

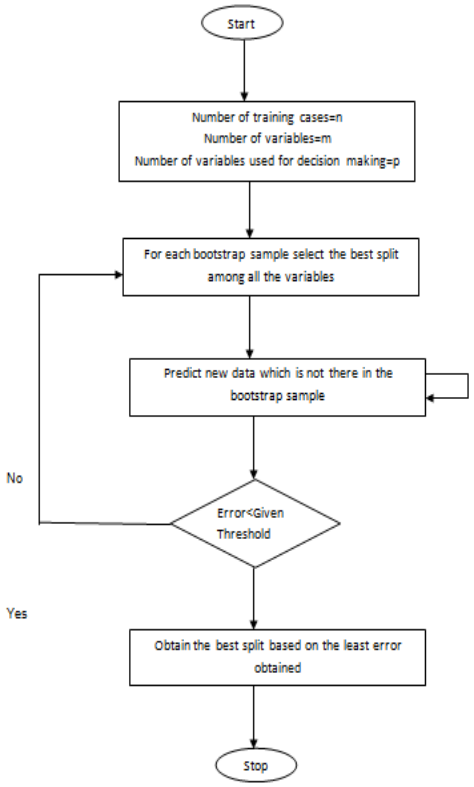

Figure 3.1: Diagrammatic representation of Random Forest algorithm

Applications:

1. Is used for image classification for pixel analysis.
2. Is used in the field of Bioinformatics for complex biological data analysis.
3. Is used for video segmentation (high dimensional data)

4. RESULTS AND DISCUSSION

R is a sophisticated statistical software package, which provides new approaches to data mining. Rattle, an R GUI, an open source tool for analysis of data mining algorithms. Rattle is built on the statistical language R. Rattle is simple to use, quickly to deploy, and allows to rapidly work through the data processing[12]. The paper discusses weather dataset for analysis of the above mentioned algorithms.

The algorithm is executed using Rattle to predict whether it is going to rain the next day of not. The Variables used in tree construction are: Evaporation, Humidity, Temperature, wind speed and direction, pressure, sunshine which is considered in following timings i.e 3PM and 9 AM[1].The number of instances used for analysis of weather data is 256. The confusion matrix shows that 205 instances predicted that it will not rain next day. The confusion matrix shows that 14 instances predicted that it will rain next day. 37 instances showed ambiguity in classification of data samples. Hence were considered to be incorrectly classified. Redistribution error rate (i.e., error rate computed on the training sample) for Random Forest is 0.0465. The time taken to construct the model is 1.68 seconds.

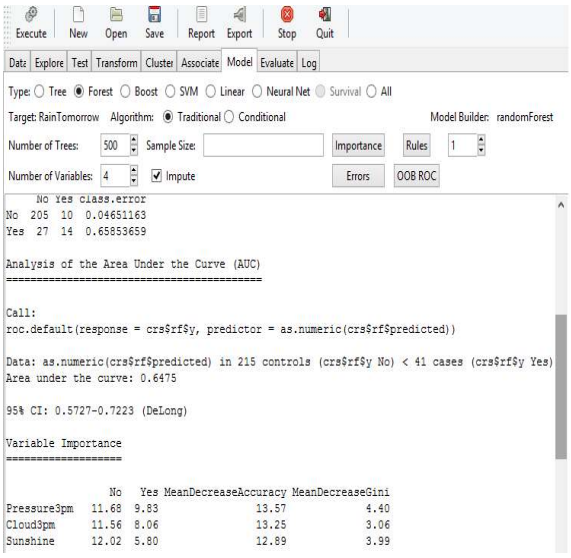

Figure 4.1: Confusion matrix of Random Forest algorithm

Figure 4.1 shows confusion matrix of random forest and the class error rate. The redistribution error rate for Decision Tree is 0.097659.It is calculated as follows: Product of relerror (0.6097) and root node error(0.16016). Redistribution error=0.6097*0.16016.The time taken to build the model is 0.16seconds.

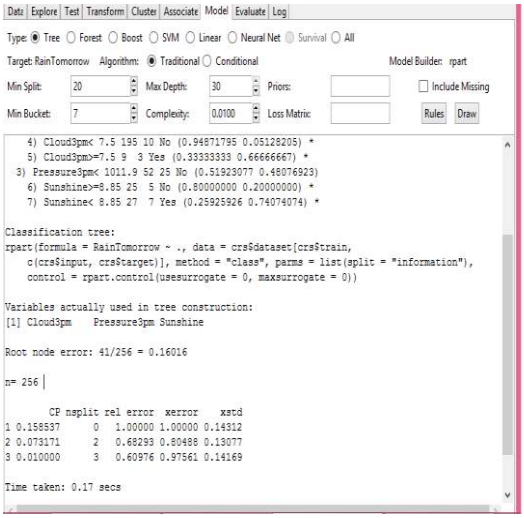

Figure 4.2 :Results of Decision Tree

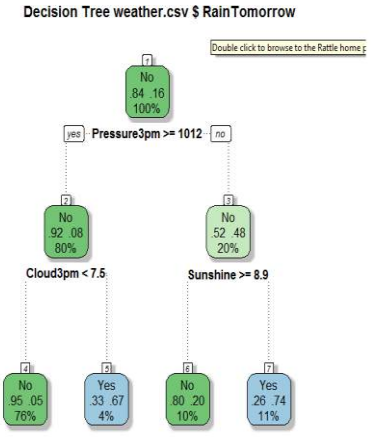

Figure 4.3 : Decision Tree for weather dataset.

DOI 10.17148/IJARCCE.2015.4142

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 1, January 2015

Figure 4.3 shows results of decision tree of weather dataset run on R platform. It depicts the rainfall for the next day based on pressure, cloud, sunshine.

Random Forest	Decision Tree
1.Time taken to construct the model more(7.68 seconds)	1.Time taken to construct model is less(0.17 seconds)
2.Redistribution error rate is less	2. Redistribution error rate is more

The table below shows the difference between random forest and Decision tree. The Redistribution error rate of random forest is less than decision tree since random forest is ensemble of trees.

5. CONCLUSION

Classification in data mining assigns data samples to target classes. For example identify loan applicants as medium, high, low in a bank data samples. The goal is to predict accurately the class to which the data samples belong to. A Random forest is ensemble of decision trees so it helps in predicting data accurately. . A random forest model is typically made up of tens or hundreds of decision trees. The above algorithm were compared and analysis was performed using the tool Rattle-R GUI [12], by considering 256 data samples of weather data set. The Redistribution error rate of random forest is less than decision tree since random forest. But time taken by random forest to execute the dataset is more compared to decision trees. Thus a technique of classification is finding importance in field of medicine.

REFERENCES

1. Data Mining, Southeast Asia Edition: Concepts and Techniques By Jiawei Han, Micheline Kamber.
2. Hong Bo Li ; Wei Wang ; Hong Wei Ding ; Jin Dong,"Trees Weighting Random Forest Method for Classifying High-Dimensional Noisy Data" IEEE 7th International Conference ,Publication Year: 2010 , Page(s): 160 – 163.
3. Ross, Peter (10/30/2000). Rule Induction: Ross Quinlan's ID3 Algorithm ,2010algo
4. Shannon, Claude E. Prediction and Entropy of Printed English.,2010-entropy
5. Yiwen Zhang ; Lili Ding ; Yun Wang ,Research and design of ID3 algorithm rules-based anti-spam email filtering, Software Engineering and Service Science (ICSESS), 2011 IEEE 2nd International Conference on , DOI: 10.1109/ICSESS.2011.5982380 , Page(s): 572 – 575,Publication Year: 2011
6. Wei Peng, Juhua Chen, and Haiping Zhou. An Implementation of ID3 -- Decision Tree Learning Algorithm. Retrieved March 10,
7. "How many trees in a random forest?", Proceedings of the 8th international conference on Machine Learning and Data Mining in PatternRecognition. Springer-Verlag Berlin, Pages 154-168 ,july 2012.
8. Martin,A. ; Aswathy,V. ; Balaji, S. Lakshmi, T.M. ,An analysis on Qualitative Bankruptcy Prediction using Fuzzy ID3 and AntColony Optimization Algorithm, Pattern Recognition, Informatics and Medical Engineering (PRIME), 2012 International Conferenceon DOI: 10.1109/ICPRIME.2012.6208382 ,Page(s):416 -421 Publication Year: 2012
9. Salperwyck,C. ; LemaireV.,Incremental decision tree based on order statistics Neural Networks (IJCNN), The 2013 International Joint Conference on NeuralNetworks, DOI: 10.1109/IJCNN.2013.6706907 , Page(s): 1 - 8,Publication Year: 2013
10. http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#prox
11. https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm
12. http://journal.r-project.org/archive/2009-2/RJournal_2009-2_Williams.pdf
13. http://mlg.eng.cam.ac.uk/zoubin/talks/lect3ssl.pdf