Study of Random Forest

Vipul Gupta[1], Anil Dhankhar[2] , Ravi Kumar Jain[3]
[1]Dept of MCA
[2] Assoc. Professor,
[3]Asstt. Professor
[1, 2, 3] RIET Jaipur

Abstract- *Modern biology has experienced an increased use of machine learning techniques for large scale and complex biological data analysis. In the area of Bioinformatics, the Random Forest (RF) technique, which includes an ensemble of decision trees and incorporates feature selection and interactions naturally in the learning process, is a popular choice. It is nonparametric, interpretable, efficient, and has high prediction accuracy for many types of data. Recent work in computational biology has seen an increased use of RF, owing to its unique advantages in dealing with small sample size, high-dimensional feature space, and complex data structures.*

I. UNDERSTANDING RANDOM FOREST

A random forest is a supervised machine learning algorithm that is constructed from decision tree algorithm. This algorithm is applied in various industries such as banking and e-commerce to predict behaviour and outcome.

What Is a Random Forest?

A random forest is a machine learning technique that is used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms.

Features of a Random Forest Algorithm

- It is more accurate than the decision tree algorithm.
- It provides an effective way of handling missing data.
- It can produce a reasonable prediction without hyper-parameter tuning.
- It solves the issue of over fitting in decision trees.
- In every random forest tree, a subset of features is selected randomly at the node's splitting point

Advantages of Random Forests

- Random forests present estimates for variable importance i.e., neural nets.
- They also offer a superior method for working with missing data.
- Missing values are substituted by the variable appearing the most in a particular node.
- Among all the available classification methods random forest provide the highest accuracy
- The random forest technique can also handle big data with numerous data running into thousands
- It can perform both regression and classification tasks.

Disadvantages of Random Forest

- When using a random forest, more resources are required for computation.
- It consumes more time compared to a decision tree algorithm.

Why use Random Forest Algorithm

There are lots of benefits of using random forest Algorithm, but one of the main advantages is that it reduces the risk of overfitting and the required training time. Additionally, it offers a high level of accuracy. Random Forest algorithm runs efficiently in large databases and produces highly accurate predictions by estimating missing data.

Important terms to know

- **Entropy:** It is a measure of randomness or unpredictability in the data set.

- **Information Gain:** Ameasure of the decrease in the entropy after the data set is split is the information gain.

- **Leaf Node:**A leaf node is a node that carries the classification or the decision.

- **Decision Node:** Anode that has two or more branches.

- **Root Node:**The root node is the topmost decision node, which is where you have all of your data.

How Random Forest Algorithm Works Understanding Decision Trees

Decision trees are the building blocks of a random forest algorithm. A decision tree is a decision support technique that forms a tree like structure. A decision tree consists of three components:

- Decision Nodes
- Leaf Nodes
- Root Nodes

A decision tree algorithm divides a training dataset into branches, which further separates into other branches. This process keeps on traversing until a leaf node is attained. The leaf node cannot be separated further.

The nodes in the decision tree represent properties that are used for forecasting the outcome. Decision tree provides a connection to the leaves. Entropy and information gain are the building blocks of a decision trees. Entropy is a metric for calculating uncertainty. Information gain is a measure of how uncertainty in the target variable is reduced, given a set of independent variables.

Applying Decision trees in Random Forest

The main difference between the decision tree algorithm and the random forest algorithm is that in random forest establishing root nodes and separating nodes is done randomly. The random forest applies the bagging method to generate the required prediction. Begging involves using different samples of data (training data) rather than just one sample. A training dataset comprises observations and properties that are used for making predictions. The decision trees construct different outcomes, depending on the training data fed to the random forest algorithm. These outputs will be ranked, and the highest will be selected as the final output.

Classification in Random Forest: -

Classification in random forest employs an ensemble methodology to attain the outcome. The training data is fed to train various decision trees. This dataset consists of observations and features that will be selected randomly during the separation of nodes.

A rain forest system relies on various decision trees. Every decision tree consists of decision nodes, leaf nodes, and a root node. The leaf node of each tree is the final output produced by that specific decision tree. The selection of the final output follows the majority-voting system.

Regression in Random Forest: -Regression is the other task performed by a random forest algorithm. A random forest regression follows the concept of simple regression. Values of dependent (features) and independent variables are passed in the random forest model. We can run random forest regression in various programs such as SAS, R, and python. In random forest regression, each tree produces a specific prediction. The mean prediction of the individual tree s is the output of the regression. This is contrary to random forest classification, whose output is determined by the mode of the decision trees class.

Applications of Random Forest

Random forest is applied in various sectors to predict the results. Some such sectors are as follows:-

Banking: -

Random forest issued in banking to predict the creditworthiness of a loan applicant. This helps the lending institutions make a good decision on whether to give the customer the loan or not. Banks also use the random forest to detect the fraudsters.

Health Care: -

Health sectors use random forest systems to diagnose patients. Patients are diagnosed by assessing their previous medical history. Past medical records are judged to establish the correct dosage for the patients.

Stock Market: -

Financial analysts use it to identify potential markets for stocks. It also enables them to identify the behaviour of stocks.

E-commerce: -

Through rain forest algorithms, e-commerce vendors can predict the preference of customers based on past consumption behaviour.

When should one Not use random Forest: -

Random forest is not useful in some cases. Some of them are discussed below: -

Extrapolation: -Random Forest regression is not ideal in the extrapolation of data. Unlike linear regression, which uses existing observations to estimate values beyond the observation range? This explains why most applications of random forest relate to classification.

Sparse Data: -

Random forest does not produce good results when the data is very sparse. In this case, the subset of features and the bootstrapped sample will produce an invariant space. This will lead to unproductive splits which will adversely affect the outcome.

II. CONCLUSION

The random forest algorithm is a machine learning algorithm that is easy to use and flexible. It uses ensemble learning, which enables organizations to solve regression and classification problems.

This is an ideal algorithm for those who develop because it solves the problem of over fitting of datasets. It is a very resourceful tool for making accurate forecasts needed in strategic decision making in organizations.

REFERENCES

[1] Decision Tree & Random Forest by Chris Smith. Originally Published:2017
[2] How Many Trees in a Random Forest? By Thais Mayumi Oshiro Pedro Santoro Perez José Augusto Baranauskas
[3] Handbook of Random Forests: Theory and Applications for Remote Sensing by Ronny Hansch
[4] Random forest in remote sensing: A review of applications and future directions
[5] Improved Random Forest for Classification Angshuman Paul; Dipti Prasad Mukherjee; Prasun Das; Abhinandan Gangopadhyay; Appa Rao Chintha; Saurabh Kundu Published in: IEEE Transactions on Image Processing (Volume: 27, Issue: 8, Aug. 2018)