FRANKFURT UNIVERSIY OF APPLIED SCIENCES, MARCH 2019, AI/ML SOFTWARE ENGINEERING SUPERVISED BY DAMIR DOBRIC

Decision Forest Regression

Yash Vyas

Abstract
Decision Forest is an ensemble of unpruned classification or regression trees created by using bootstrap samples of the training data and random feature selection in tree induction. Prediction is made by aggregating (majority vote or averaging) the predictions of the ensemble.

Keywords — Decision Forest, Regression, Bagging, Decision tree, Information Gain, Entropy,

I. INTRODUCTION

A random decision forest is an ensemble of randomly trained decision trees. The key aspect of the forest model is the fact that its component trees are all randomly different from one another. This leads to decorrelation between the individual tree predictions and, in turn, results in improved generalization and robustness. The forest model is characterized by the same components as the decision trees. The family of weak learners (test functions), energy model, the leaf predictors and the type of randomness influence the prediction/estimation In a forest with T trees we use the variable t ∈ {1,…,T} to index each component tree. All trees are trained independently. During testing, each test point **v** is simultaneously pushed through all trees (starting at the root) until it reaches the corresponding leaves Combining all tree predictions into a single forest prediction may be done by a simple averaging operation or with the partition function ensuring probabilistic normalization

II. METHOD

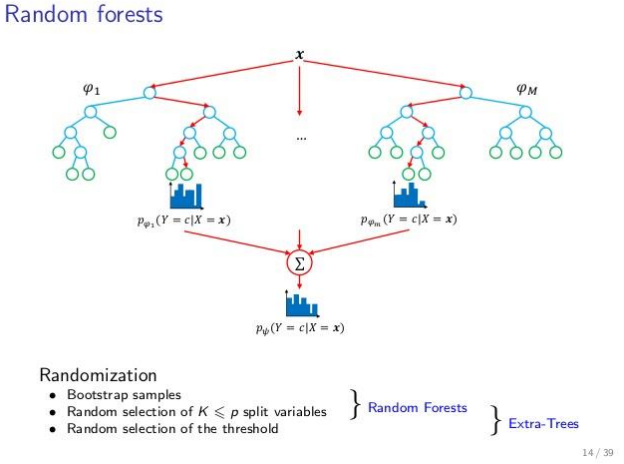
The decision forests, their construction and prediction abilities depend on the model parameters. The parameters that most influence the behaviour of a decision forest are:
• The maximum allowed tree depth D;
• The amount of randomness (controlled by ρ) and its type;
• The forest size (number of trees) T;
• The choice of weak learner model;
• The training objective function;
• The choice of features in practical applications.
Those choices directly affect the forest predictive accuracy, the accuracy of its confidence, its generalization and its computational efficiency.

Decision Forest Model for Regression

The regression task can be summarized as follows:

Given a labelled training set learn a general mapping which associates previously unseen independent test data with their correct continuous prediction.

Why regression forests? A regression forest is a collection of randomly trained regression trees. Just like in classification it can be shown that a forest generalizes better than a single over-trained tree. A regression tree splits a complex nonlinear regression problem into a set of smaller problems which can be more easily handled by simpler models. Next we specify the precise nature of each model component.



Regression trees are trained by minimizing a least-squares or least-absolute error function. Here, for consistency with our

General forest model we employ a continuous formulation of information gain.

$$I_j = \sum_{\mathbf{v}\in\mathcal{S}_j} \log(|\Lambda_{\mathbf{y}}(\mathbf{v})|) - \sum_{i\in\{\mathrm{L,R}\}} \left(\sum_{\mathbf{v}\in\mathcal{S}_j^i} \log(|\Lambda_{\mathbf{y}}(\mathbf{v})|) \right) \quad [1]$$

The ensemble model. Just like in classification, the forest output is the average of all tree outputs

$$p(\mathbf{y}|\mathbf{v}) = \frac{1}{T}\sum_{t}^{T} p_t(\mathbf{y}|\mathbf{v})$$

III. ADVANTAGES AND DISADVANTAGES OF DECISION FOREST REGRESSION

ADVANTAGES:

- It is one of the most accurate learning algorithms available. For many data sets, it produces a highly accurate classifier. [2]
- It runs efficiently on large databases.
- It can handle thousands of input variables without variable deletion.
- It gives estimates of what variables are important in the classification.
- It generates an internal unbiased estimate of the generalization error as the forest building progresses.
- It has an effective method for estimating missing data and maintaining accuracy.
- It has methods for balancing error in class population unbalanced data sets.
- Prototypes are computed that give information about the relationship between the variables and the classification.
- There are a number of cases that can be used in clustering, locating outliers, or (by scaling) give interesting views of the data.
- The capabilities of the above can be extended to unlabeled data, leading to unsupervised clustering, data views and outlier detection.
- It offers an experimental method for detecting variable interactions.

DISADVANTAGES:

- Random forests have been overlooked to over fit for some datasets with noisy classification / regression tasks. [3]

.

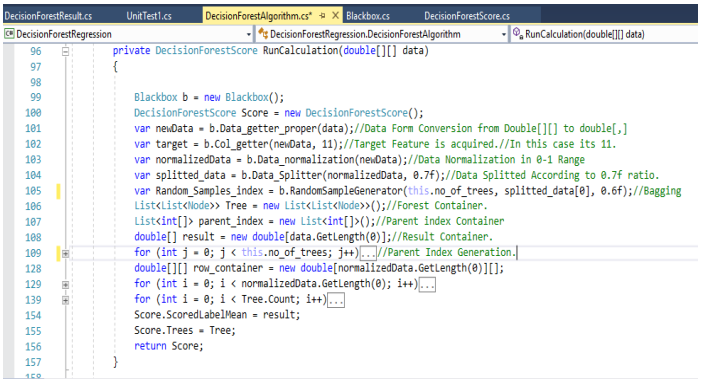
IV. PROJECT

A) Git Hub reference

https://github.com/UniversityOfAppliedSciencesFrankfurt/se-dystsys-2018-2019-softwareengineering/tree/YashVyas-YV/My_Works/Decision_Forest/DecisionForestRegression

B) Implementation

1. DecisionForest RunCalculation Method.



2. BlackBox Class Overview

VII. REFERENCES

[1] B. B. a. I.Reid, "Unsupervised learning of a scene specific coarse gaze estimator," in *International Conference on Computer Vision*, Barcelona, Spain, 2011.

[2] K. E, "On the Algorithmic Implementation of Stochastic Discrimination.," *IEEE TRANSACTIONS ON PAMI*, 2000.

[3] B. L, "Random Forest," 2001.