Random Forests

Leo Breiman
Presented by Jizhou Xu

Summary

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. Each tree in the forest cast a unit vote for the most popular class at input.

Structure of Paper

Theoretical Background for Random Forests

Forests Using the Random Selection Features at Each Node to Determine the Split

Empirical Results for Two Different Forms of Random Features

Study on Why Selecting One or Two Features Gives Near Optimum Results

In Its Later Stages Adaboost Is Emulating a Random Forest

Combining Trees Grown Using Random Features Can Produce Improved Accuracy

Random Forests for Regression & Empirical Results

Opportunity

Its accuracy is as good as Adaboost and sometimes better.

Relatively robust to outliers and noise.

Faster than bagging or boosting.

Useful internal estimates of error, strength, correlation and variable importance.

Simple and easily parallelized.

Data with Many Weak Inputs.

Challenge

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

A large number of trees may make the algorithm slow for real-time prediction.

The Algorithm

Let the number of training cases be N, and the number of variables in the classifier be M.

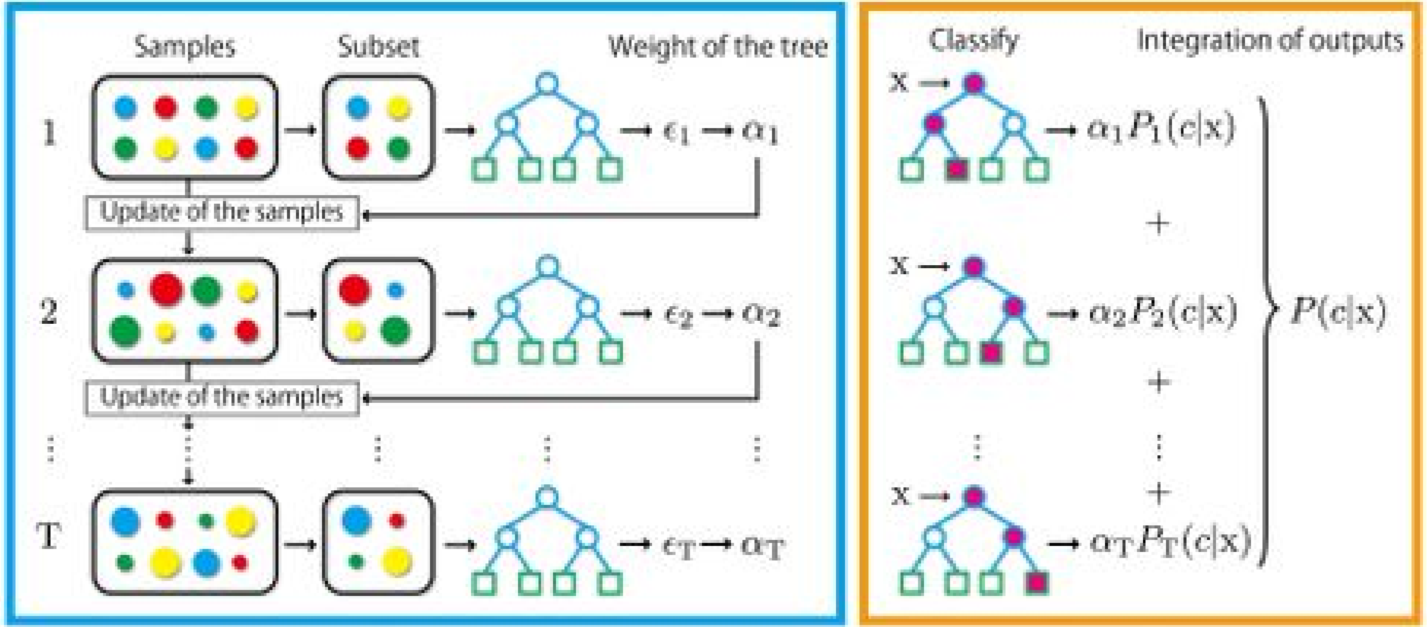
The number m of input variables are used to determine the decision at a node of the tree (m should be much less than M).

Choosing a training set for this tree by choosing N times with replacement from all N available training cases. Use the rest of the cases to estimate the error of the tree, by predicting their classes

For each node of the tree, randomly choose m variables on which to base the decision at that node. Calculate the best split based on these m variables in the training set.

Each tree is fully grown..

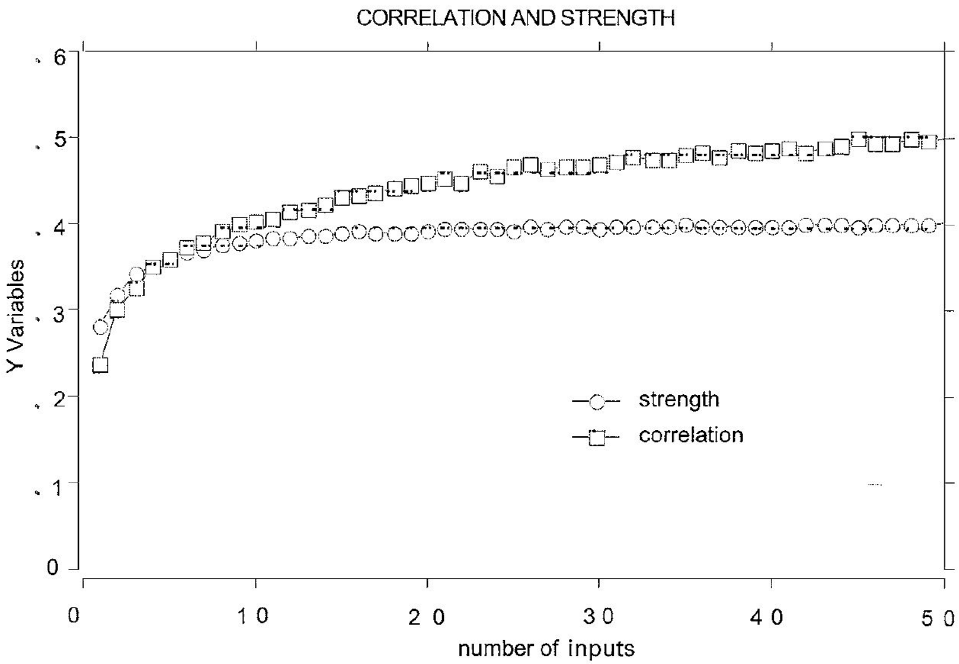
The Algorithm



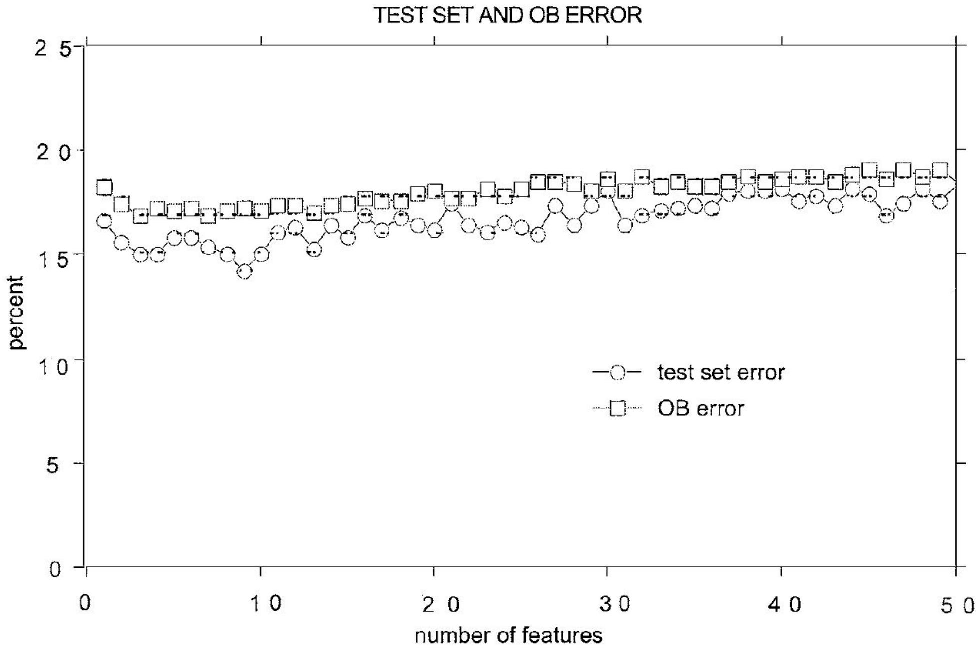
Learning Process

Classify Process

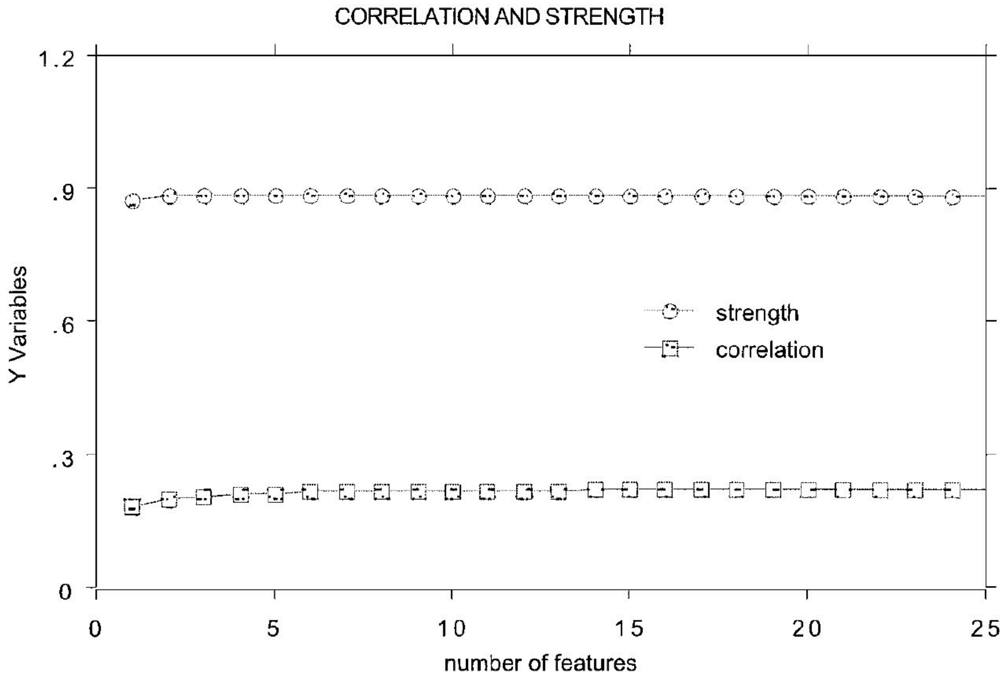
Resolution


CORRELATION AND STRENGTH
.6
.5
.4
.3
.2
.1
0
Y Variables
0
10
20
30
40
50
number of inputs
strength
correlation

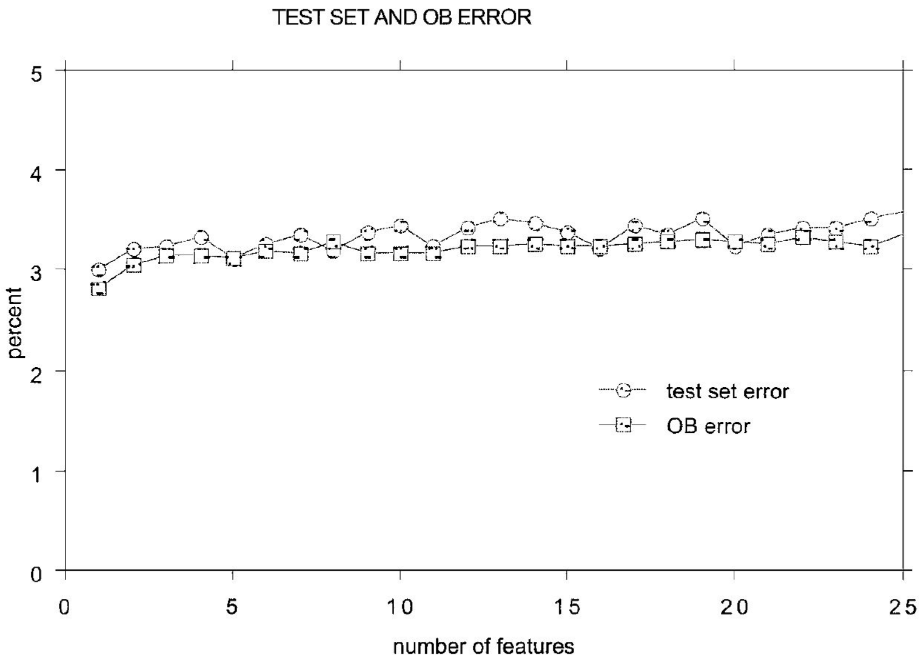
Resolution


TEST SET AND OB ERROR
percent
25
20
15
10
5
0
0
10
20
30
40
50
number of features
—○— test set error
··□·· OB error

Resolution


CORRELATION AND STRENGTH
1.2
.9
.6
.3
0
Y Variables
strength
correlation
0
5
10
15
20
25
number of features

Resolution


TEST SET AND OB ERROR
5
4
3
2
1
0
percent
0
5
10
15
20
25
number of features
test set error
OB error

“

Warren Buffett is one of the best learning machines on this earth. The turtles which outrun the hares are learning machines. If you stop learning in this world, the world rushes right by you.

Applications

Air Quality Prediction

Travel Time Prediction

UPenn and Mayo Clinic’s Seizure Detection

Future

Data Analysis

Screening

Information Processing

Grow Sales

THANKS!

Any questions?