Random Forests

Giulio Leonardi, Fabio Melasi

24th july 2024

What is a Random Forest?

- Definition: a classifier composed of a collection of k decision trees $h(x, \Theta_k)$ where Θ_k are independent identically distributed random vectors.
- Typically, also bagging is used: each tree is trained on a bootstrap sample of the training set.
  - Necessary for the out-of-bag estimation $\rightarrow$ presented soon!
- At inference time, for an input vector x, each tree votes for a class, and the majority class is the prediction of the random forest.

Characterizing Random Forests: convergence

- The margin measures the average of the votes for the correct class Y, minus the maximal average of the votes to every other class.
  $\uparrow$ *margin* $=\uparrow$ classification confidence

$$mr(X,Y) = av_k I(h_k(X) = Y) - \max_{j \neq Y}(av_k I(h_k(X) = j))$$

- Generalization Error:

$$PE^* = P_{X,Y}(mr(X,Y) < 0)$$

- As the number of trees increase, a.s. PE^* converges to the limit:

$$P_{X,Y}(P_\Theta(h(X,\Theta) = Y) - \max_{j \neq Y} P_\Theta(h(X,\Theta) = j) < 0)$$

- So, as the number of weak learners increases, the probability that the margin is negative stabilizes, **without causing overfitting**.

Characterizing Random Forests: PE^* upper bound

- Two important metrics for a forest:
  - **Strength** s: how accurate are the individual classifiers.

$$s = E_{X,Y} mr(X, Y)$$

  - **Correlation** $\bar{\rho}$: the dependence between the individual classifiers.

$$\bar{\rho} = \frac{E_{\Theta_i,\Theta_j}\left(\rho(\Theta_i, \Theta_j) \cdot \mathrm{sd}(\Theta_i) \cdot \mathrm{sd}(\Theta_j)\right)}{E_{\Theta_i,\Theta_j}\left(\mathrm{sd}(\Theta_i) \cdot \mathrm{sd}(\Theta_j)\right)}$$

- **Generalization error** upper bound:

$$PE^* \leq \frac{\bar{\rho}(1 - s^2)}{s^2}$$

- Strength and correlation impact the generalization error of the random forest, so the goal should be to **maximize strength and minimize correlation**.

Out-of-bag Error Estimation

- Using bagging, each tree is trained on a bootstrap sample of the training set.
  - Records **selected** at least once in a tree's sample are its **in-bag** records.
  - The training set records **not selected** in the tree's sample are its **out-of-bag** records.
- **OOB Classification**: classify each training set record by considering only the votes from the decision trees that have that record out-of-bag.
  - Therefore, each weak learner makes predictions only on the records it was not trained on.
- **OOB Error**: error rate of the OOB classification.

Experimental Setup

Name	Train Size	Test Size	Features	Classes
Breast Cancer	699	-	9	2
Glass	214	-	9	6
Sonar	208	-	60	2
Sat-Images	4435	2000	36	6
Letters	15000	5000	16	26

Table: Datasets selected for our experiments.

- For Small Datasets 100 iterations with random 10% of the data as test set.
- For Large datasets, the training set is used directly.
- Observed metrics:
  - test error of the model;
  - selected test error between two models with different hyperparameters, through the OOB error estimation.

First approach: Forest-RI

- Simplest type of random forest: during the training of each tree, at each node, F features are randomly selected to find the best split.
- Tested parameters:
  - $F = 1$;
  - $F = \lfloor log_2 M + 1 \rfloor$ with $M =$ number of inputs in the training set.

	Selection (paper)	Selection (own)	Single input (paper)	Single input (own)
Breast	2.9	3.2	2.7	3.1
Glass	20.6	21.2	21.2	21.4
Sonar	15.9	17.2	19.0	18.1
Sat-images	8.6	9.1	10.5	9.7
Letters	3.5	4.2	4.7	5.7

Table: Test set error (%) comparison between the original paper and our implementation.

Second approach: Forest-RC

- Random forests using linear combination of inputs.
- The splitting decisions at various nodes in the trees are based on features created from random linear combinations of multiple input variables.
  - L: number of variables to be combined (*to_combine* in the scripts).
  - F: number of newly generated linear combined variables.
  - At a given node, L variables are randomly selected and added together with coefficients that are uniform random numbers on [-1, 1]. F linear combinations are generated.

Forest-RC: Implementation and results

- The same procedure as for Forest-RI was used: for small datasets, 100 iterations with random 10% of the data as a test set. For large datasets, only one iteration was performed, and the model was tested on the test set.
- Selected parameters: $L = 3$ and $F = 2, 8$.
  - Also in this case, two forests with different F values has been trained. The final error was selected depending on the out-of-bag error.

	Selection (paper)	Selection (own)	Two features (paper)	Two features (own)
Breast	3.1	2.9	2.9	2.9
Glass	24.4	25.9	23.5	25.5
Sonar	13.6	17.7	13.8	17.5
Sat-images	9.1	9.9	10.2	10.7
Letters	3.4	3.9	4.1	4.6

Table: Test set error (%) comparison between the original paper and our implementation.

Empirical results on strength and correlation

- Out-of-bag methods can be used to estimate not only error but also **strength** and **correlation**.
- This enables us to empirically study the impact of these two metrics on generalization error across various datasets.
  - *Sonar*, *Breast cancer* and *Sat-images* datasets $\rightarrow$ results in a while!
- For each dataset, i iterations were performed. In each iteration, a *maxF* number of forests were grown, each with a different F value ranging from 1 to *maxF*.
  - The *error*, *strength*, and *correlation* metrics were estimated from each forest and averaged across all iterations. This approach allows us to study the impact of F on these metrics.

How strength and correlation were estimated?

OOB estimate for strength

- As we saw before, the strength is given by the expectation of margins

$$mr(x, y) = P_\Theta(h(x, \Theta) = y) - \max_{j \neq y} P_\Theta(h(x, \Theta) = j)$$

- We estimate the probabilities with the out-of-bag proportion of votes cast at the input **x** for class **j**, calculated as

$$Q(x, j) = \frac{\sum_k I(h(x, \Theta_k) = j; (y, x) \notin T_{k,B})}{\sum_k I((y, x) \notin T_{k,B})}$$

- Then, the strength is empirically obtained substituting $Q(x, y)$ and $Q(x, j)$ for $P_\Theta(h(x, \Theta) = y)$ and $P_\Theta(h(x, \Theta) = j)$ and taking the average over the training set.

$$s = E[Q(x, y) - \max_{j \neq y} Q(x, j)]$$

OOB estimate for correlation

- To estimate the correlation, we need the variance of the margins:

$$var(mr) = E[mr^2] - s^2$$

- and the standard deviation of Θ

$$sd(\Theta) = [p_1 + p_2 + (p_1 - p_2)^2]^{1/2}$$

where p_1 is the proportion of OOB instances correctly classified while p_2 the proportion of OOB instances wrongly classified*.

$$p_1 = E_{X,Y}(h(X, \Theta) = Y) \quad and \quad p_2 = E_{X,Y}(h(X, \Theta) = \hat{j}(X, Y))$$

- Then, the correlation is obtained by

$$\bar{\rho} = \frac{var(mr)}{(E_\Theta sd(\Theta))^2}$$

Effect of number of inputs on Sonar data

- $iterations = 80,\ maxF = 50,\ to\_combine = 1$

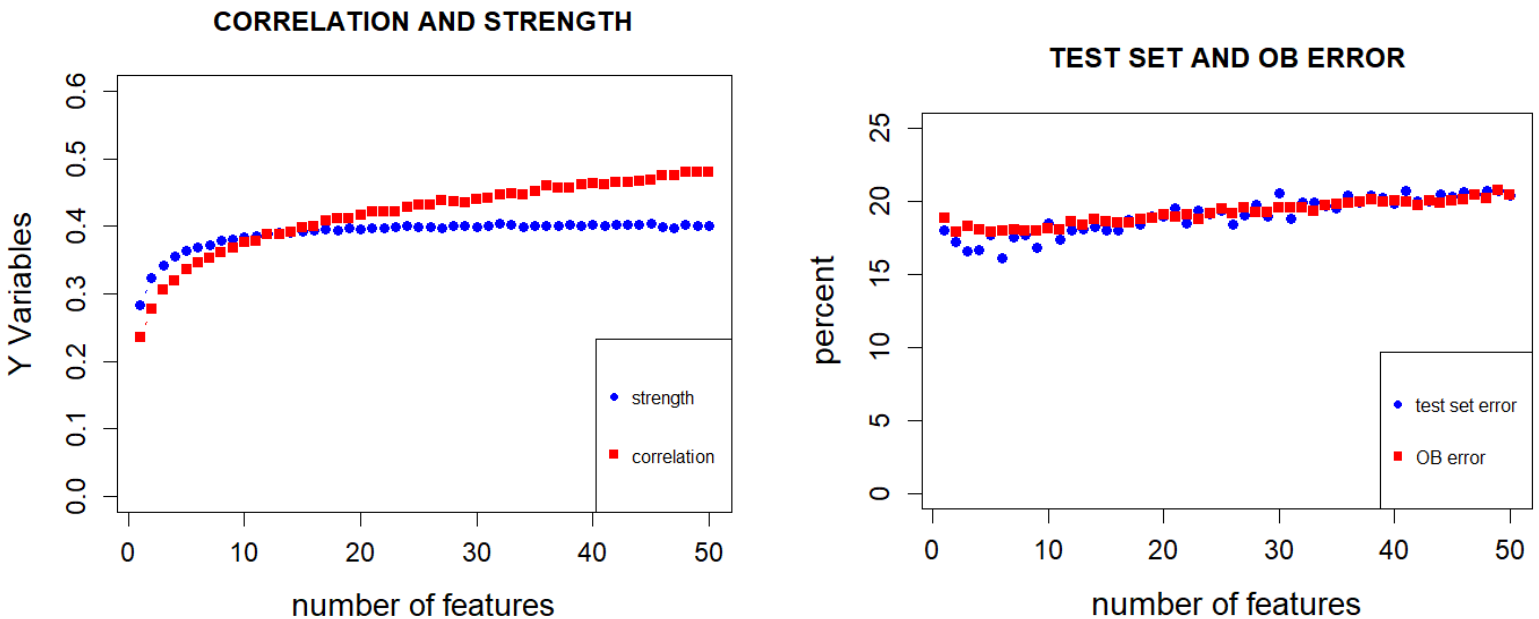

Figure: Reproduction of the Figure 1 of the original paper.

Effect of number of inputs on Breast cancer data

- $iterations = 80,\ maxF = 25,\ to\_combine = 3$

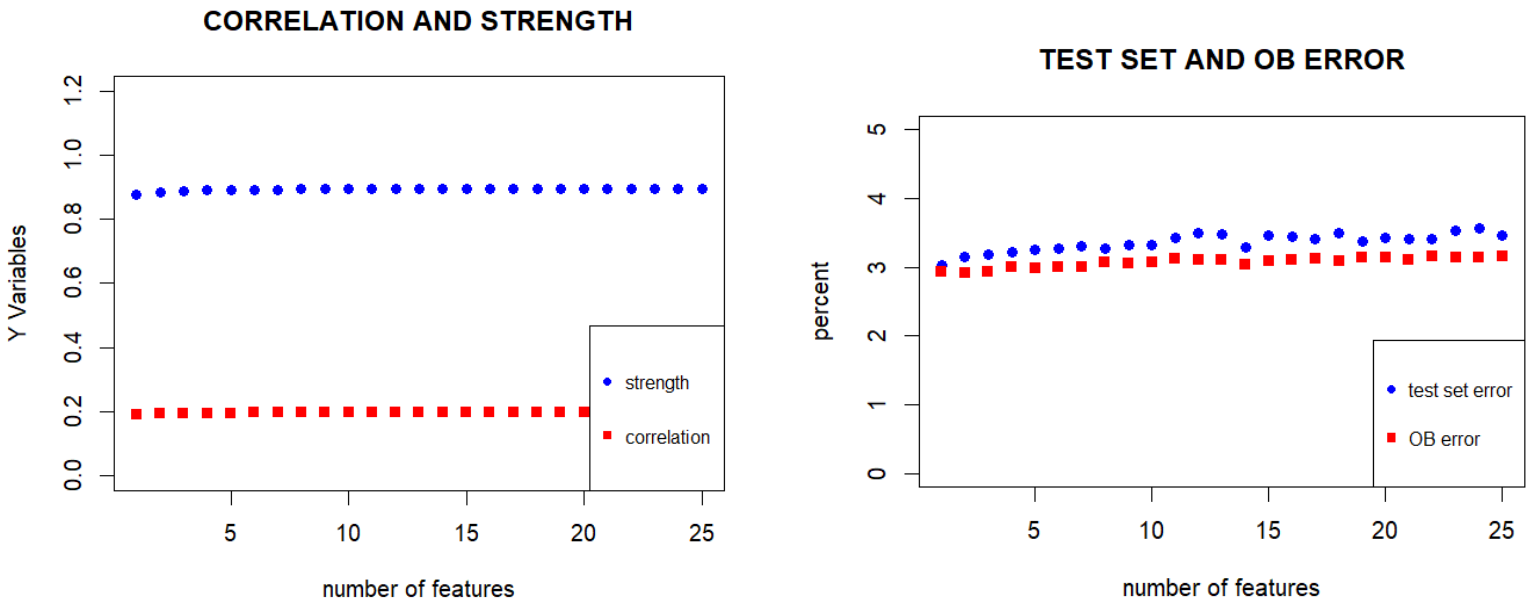

Figure: Reproduction of the Figure 2 of the original paper.

Effect of number of inputs on Sat-images data

- $iterations = 30,\ maxF = 25,\ to\_combine = 2$

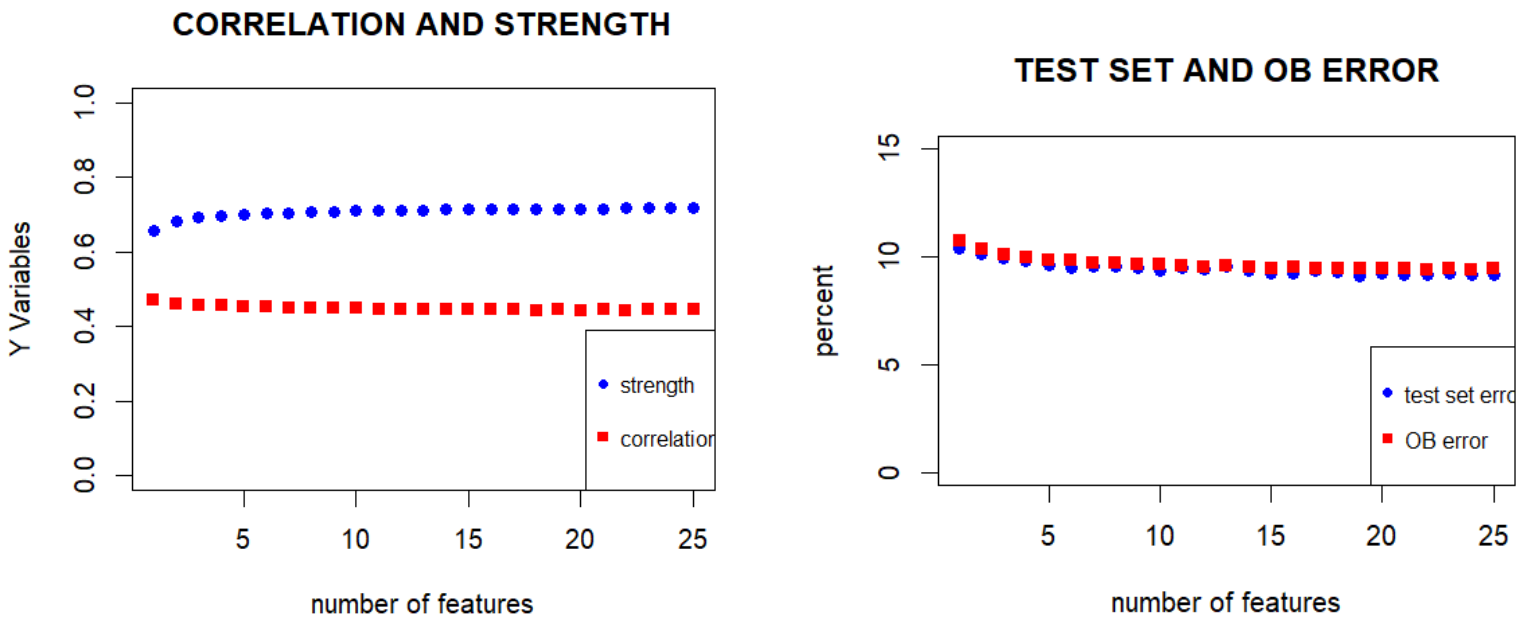

Figure: Reproduction of the Figure 3 of the original paper.