DD2434 ML II
Random Forests

Group 25

Oktay Bahceci Hui Pang Sugandh Sinha Diego Yus

Introduction - Aim of the Article

- Decision Trees
  - Predictive model
  - Predict value of target variable based on input variables
    - Split at each node, best attribute
- Bootstrap Aggregating - Bagging
  - Split dataset and generate new replicas
- Ensemble Methods

Aim of the Article

- Decision Trees can easily overfit training data
- Article introduces concept of *Forests*
  - Ensemble method of combining Trees by Bagging.
  - *Random* selection of features/attributes
  - Given a new datapoint *X*, classify X using all generated trees
    - Classification: Majority vote
    - Regression: Average of all trees
- Ensemble methods works for Decision Trees!
- Provides better accuracy than a single tree

Random Forests does not overfit, because ...

- Generalization error has an upper bound:

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

    - s: Strength of an individual classifier (tree)
    - ρ: Correlation between individual classifiers (trees)

Method : Datasets

- Used 6 datasets (binary or multi-classes, numerical or categorical, large attributes)
  - Categorical datasets - Breast Cancer, House Votes
  - Non - Categorical datasets - Diabetes, Sonar, Vehicle, Glass
  - Binary Classification - Breast Cancer, House Votes, Diabetes, Sonar
  - Multiple - class Classification - Vehicle, Glass
- Handling missing values
  - For non-categorical data points by taking median
  - For categorical data points by taking most frequent non-missing value

Method : Implementation

- Used CART to grow trees:
  - Grow to maximum size (no pruning)
  - Binary splitting at each node
  - Splitting rules based on feature values.

- Values of parameters such as number of trees and number of runs were same as in paper.

Results - Random input selection

Dataset	Selection	Single Input	One tree
Glass	31.5	34.6	40.0
Breast Cancer	3.4	2.7	7.1
Diabetes	27.7	32.1	33.3
Sonar	22.0	22.3	30.4
Vehicle	43.9	44.4	45.0
Votes	4.1	5.4	7.1

Data set	Adaboost	Selection	Forest-RI single input	One tree
Glass	22.0	20.6	21.2	36.9
Breast cancer	3.2	2.9	2.7	6.3
Diabetes	26.6	24.2	24.3	33.1
Sonar	15.6	15.9	18.0	31.7
Vowel	4.1	3.4	3.3	30.4
Ionosphere	6.4	7.1	7.5	12.7
Vehicle	23.2	25.8	26.4	33.1
German credit	23.5	24.4	26.2	33.3
Image	1.6	2.1	2.7	6.4
Ecoli	14.8	12.8	13.0	24.5
Votes	4.8	4.1	4.6	7.4

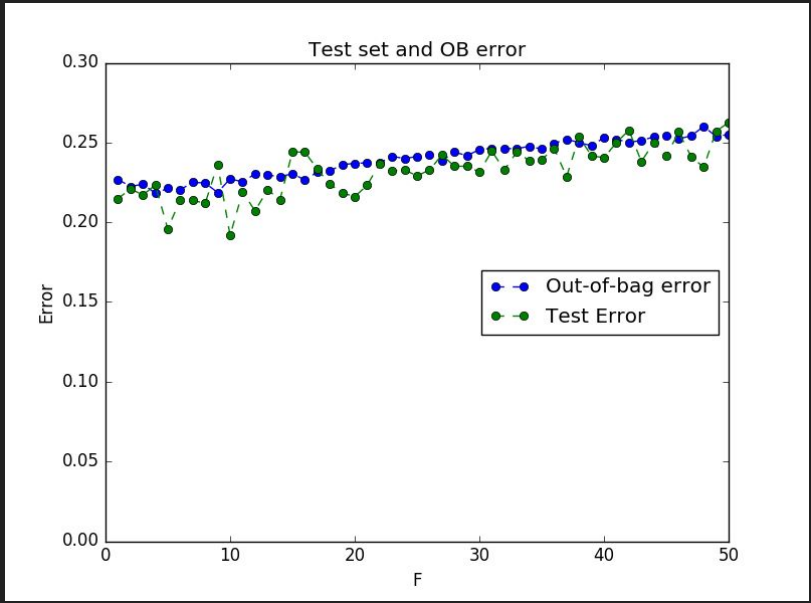
Reasons: “Best split” among values of attribute. Not stipulated in paper.

Our approach: Threshold = Middle point of domain

- Categorical: similar performance.

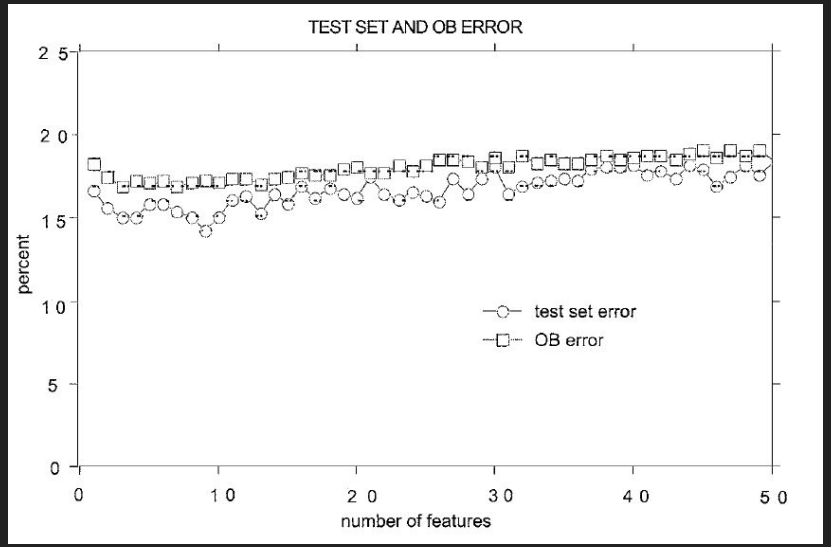
- Numerical: error increase of 5-10%

Results - OOB error and test error



Same shape/behavior

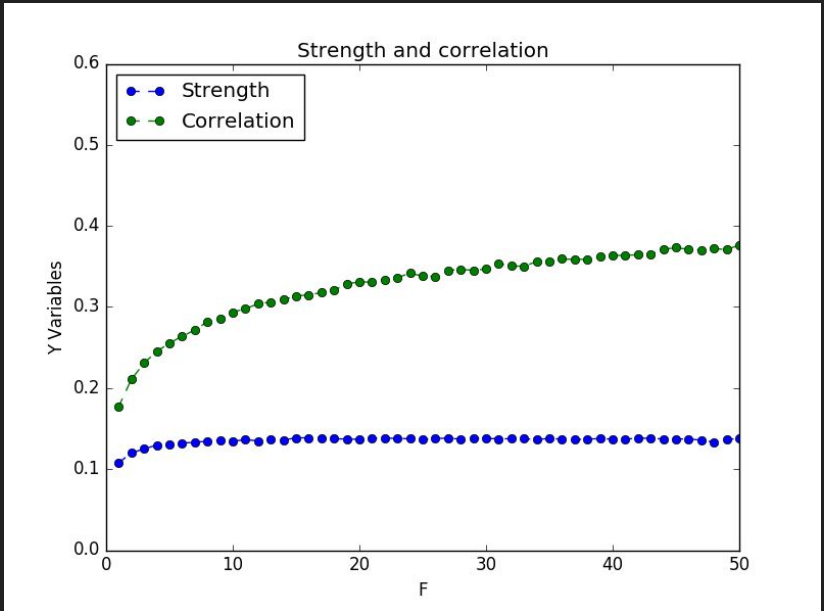
a small drop from $F = 1$ out to F about 4–8, and then a general, gradual increase.



Different absolute value for both OOB and test error.

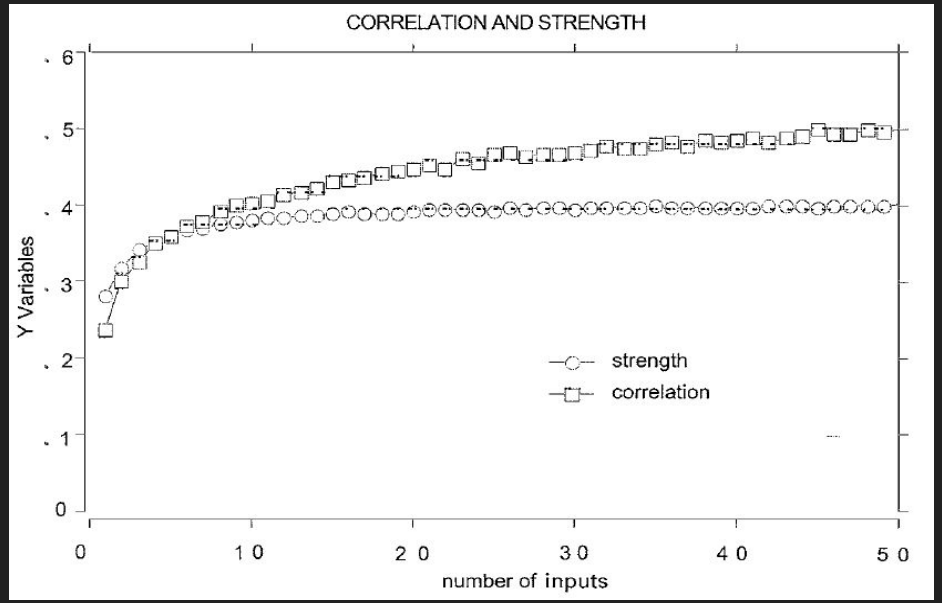
OOB overestimates error. More estable.

Results - Strength and correlation



Same shape/behavior

Past about $F = 4$ the strength remains constant; the correlation continues to increase.



Different absolute value for strength: Caused by worse classification: Lower margin function.

Correlation: slightly lower value.

What is good about RF (and we have tested)？

No worries for overfitting -- Convergence of generalization error guaranteed according to the Law of Large Numbers;

Out-of-bag estimates **avoids the need for cross-validation;**

Handle both **categorical** and **continuous** data sets;

Favorable results compared with other state-of-the-art methods(e.g. Adaboost);

Easily parallelizable, potential to deal with large real-life systems.

What is NOT good about Random Forests?

Drawbacks:

- “Black box” -- the mechanism of how variables interact with each other inside random forest is still not clear;

- As the number of tree grows and without parallelization, the computing can be time-consuming.

Possible Improvements

Well-known efforts:

- ExtraTrees, that introduces more randomness by selecting random points for splitting, and can be computationally faster;
- Rotation Forest, manipulate the description space by building subsets with Principle Component Analysis;

Improvements:

- We may introduce more dynamic ways to optimize the selection of tree numbers and group sizes, by monitoring the strength and correlation and therefore the upper bound of generalization error.

Thank you!

Results - Strength and Correlation - Zoomed in

Background slide

Decrease until F = 4-8. Then continuous increase.