Unit4: Task2 - Random Forests

Pattern Recognition SS 2016

Fausto Heraldo Sifuentes Caccire (0607000)

May 9, 2016

1 Theory

1.1 Random Forest

The Random Forest classifier is an algorithm introduced by Leo Breiman [5] wich consists of an ensemble of decision trees each grown with an independent equally distributed subset of random samples out of a superset (see Bagging for details). Classification is achieved by using the same testing value as input for every single tree. After each tree has made a decision, the class with the most votes is chosen. As the trees work independently, parallelism can be used to reduce the computation time.

1.2 Bagging

Breiman described this approach for the growth (training) of every tree in his Random Forest. Bagging is an abbrevation for *bootstrap aggregating*. It consists of choosing n random samples out of the training set T and saving them in a vector (or subset) T_i. This can be done s times. At the end of this selection stage, a superset RF can be created containing s different and independent T_i, each one representing a random decision tree. At every node of each tree, the split is done by selecting m random features out of the M features of every training input. With this m features, the best one can be computed or selected to create new tree branches. The vector T_i is also called *bootstrap* and it may not contain all the cases (a training value and its corresponding class) avaiable in the training set. As described by Breiman [5], his bagging approach works without replacement (every bootstrap contains unique samples), in other approaches found on the internet, a *bootstrap* may contain the same sample several times and its size is equal to the training set's size. [1] [2].

1.3 Out-Of-Bag Error

As stated above, a bootstrap may not contain all possible cases. For each case a subset of trees without this case can be found. This left out cases are called *out-of-bag* data. With this data, an estimate of the generalization error can be calculated. [3] [4]

2 Practice

For this part of the task, the Landset satellite dataset and the MATLAB function TreeBagger with Breiman's split approach were used. The training set consists of 5148 and the test set of 1287 36th-dimensional feature vectors. A test consists of a unique combination of the forest, the traning set and the split sizes. Because of the randomness, running a test several times would return different classification results, so every test (except for the last two) was run 10 times and the average result was taken.

# Trees	# Features	# RFeats	avg. TP	avg.Precission
50	100	5	644.6	0.5055167055
50	100	30	630.8	0.4901320901
50	515	5	787.6	0.6119658120
50	1030	5	798.2	0.6202020202
50	1544	5	976.8	0.7589743590
50	2059	5	1060.0	0.8236208236
50	2574	5	1150.2	0.8937062937
50	2574	30	1144.4	0.8891996892
100	2574	5	1153.0	0.8958818959
100	2574	30	1150.8	0.8941724942
50	5148	5	1176.2	0.9139083139
50	5148	30	1164.2	0.9045843028
1000	5148	30	1171.4	0.9098679099
4000	5148	30	1170.0	0.9090909090

This table shows, that only the size of the training set biases the result of the classification the most. Having a forest size of 50 trees and only 5 random features for split, only half of the training set ist needed to achieve satisfactory results. As for why 50 trees are sufficient, Figures 2 and 3 show a visual explanation.

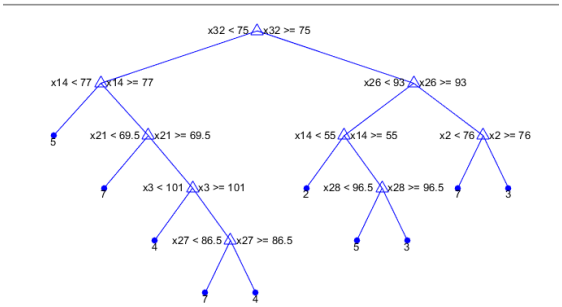

Figure 1: Example of a tree created with TreeBagger

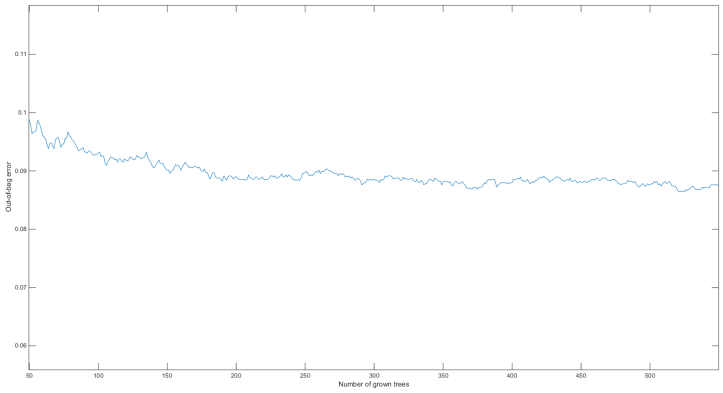

Figure 2: OOB Error with 4000 trees and the whole set. It can be seen, that having more than 50 trees reduces the error to 9 - 10%.

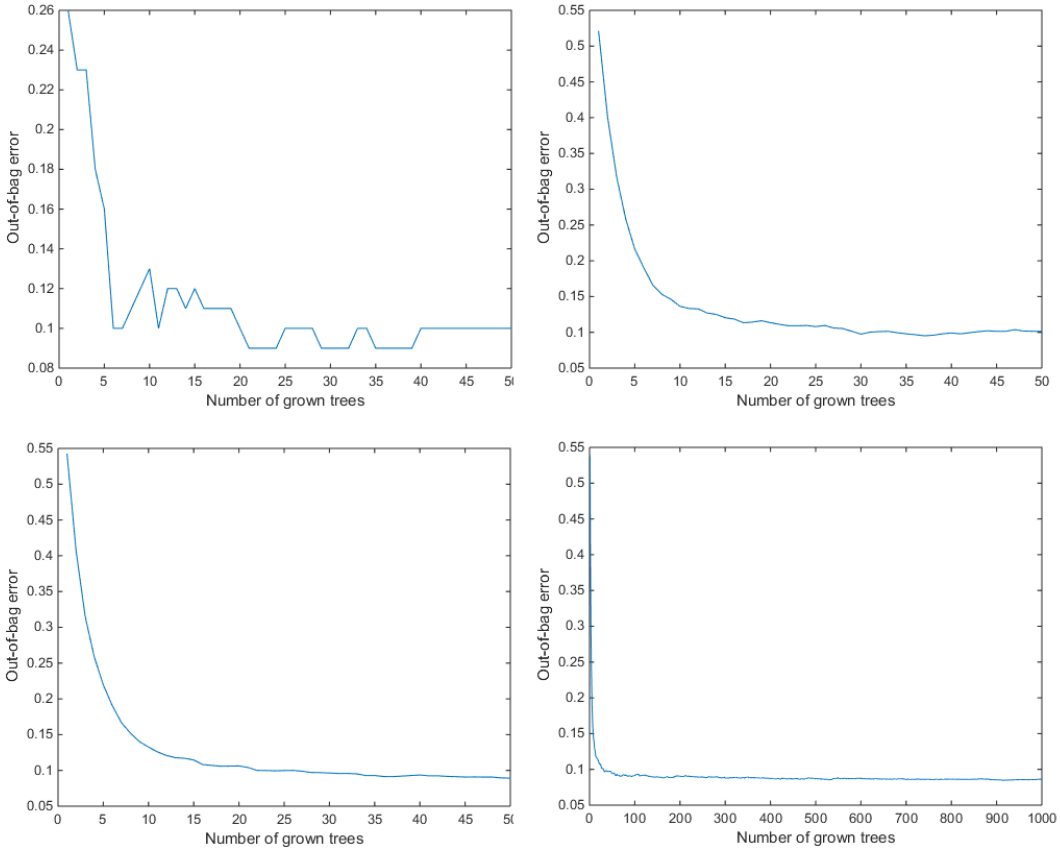

Figure 3: OOB Error for different values. Upper left: 50 trees & 100 features; upper right: 50 trees & 2574 features; lower left: 50 trees & all features; lower right: 1000 trees & all features

References

[1] https://en.wikipedia.org/wiki/Bootstrap_aggregating.

[2] https://www.statistik.tu-dortmund.de/fileadmin/user_upload/Lehrstuehle/Genetik/BS1213/Biostatistik_mit_R_Knorre.pdf.

[3] https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.

[4] https://www.r-project.org/conferences/useR-2009/slides/Cutler.pdf.

[5] Leo Breiman. Random forests. 45:5–32, 2001.