random forest tuning - tree depth and number of trees

I have basic question about tuning a random forest classifier. Is there any relation between the number of trees and the tree depth? Is it necessary that the tree depth should be smaller than the number of trees?

random-forest

asked Jan 25 at 16:11

Vysh
52 7

	Are you sure you have the right forum? Should this maybe be in "The Great Outdoors" instead, or is this really a programming question? – B. Clay Shannon Jan 25 at 16:15
2	@B.ClayShannon Random forests is a machine learning method. His question totally belongs here. – Tim Biegeleisen Jan 25 at 16:16
1	I have never heard of a rule of thumb ratio between the number of trees and tree depth. Generally you want as many trees as will improve your model. The depth of the tree should be enough to split each node to your desired number of observations. – Tim Biegeleisen Jan 25 at 16:20
	@TimBiegeleisen here's my thumb rule :) – Soren Havelund Welling Jan 26 at 12:06

2 Answers

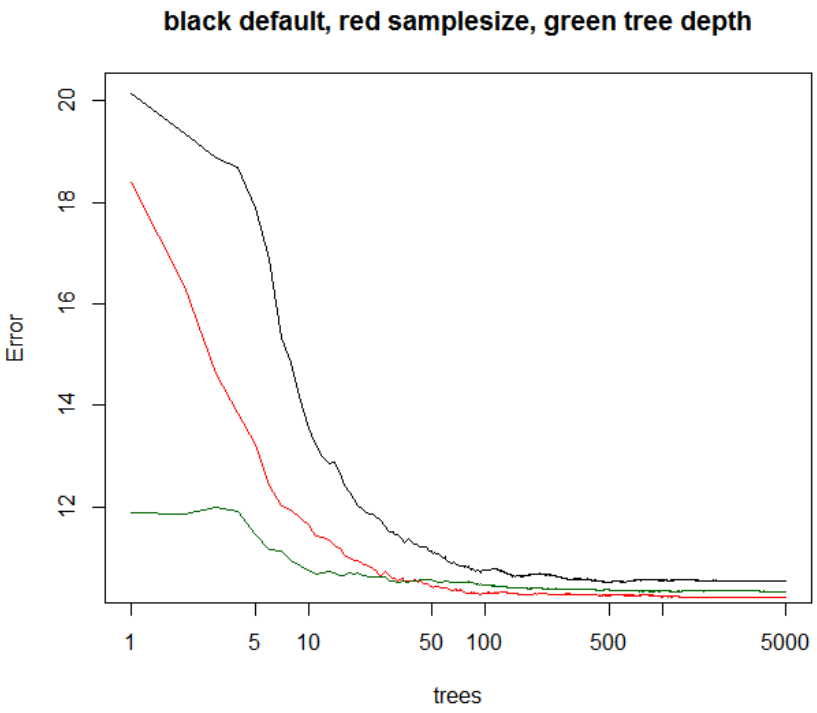
For most practical concerns, I agree with Tim.

Yet, other parameters do affect when the ensemble error converges as a function of added trees. I guess limiting the tree depth typically would make the ensemble converge a little earlier. I would rarely fiddle with tree depth, as though computing time is lowered, it does not give any other bonus. Lowering bootstrap sample size both gives lower run time and lower tree correlation, thus often a better model performance at comparable run-time. A not so mentioned trick: When RF model explained variance is lower than 40%(seemingly noisy data), one can lower samplesize to ~10-50% and increase trees to e.g. 5000(usually unnecessary many). The ensemble error will converge later as a function of trees. But, due to lower tree correlation, the model becomes more robust and will reach a lower OOB error level converge plateau.

You see below samplesize gives the best long run convergence, whereas maxnodes starts from a lower point but converges less. For this noisy data, limiting maxnodes still better than default RF. For low noise data, the decrease in variance by lowering maxnodes or sample size does not make the increase in bias due to lack-of-fit.

For many practical situations, you would simply give up, if you only could explain 10% of variance. Thus is default RF typically fine. If your a quant, who can bet on hundreds or thousands of positions, 5-10% explained variance is awesome.

the green curve is maxnodes which kinda tree depth but not exactly.




```
library(randomForest)

X = data.frame(replicate(6,(runif(1000)-.5)*3))
ySignal = with(X, X1^2 + sin(X2) + X3 + X4)
yNoise = rnorm(1000,sd=sd(ySignal)*2)
y = ySignal + yNoise
plot(y,ySignal,main=paste("cor="),cor(ySignal,y))

#std RF
rf1 = randomForest(X,y,ntree=5000)
print(rf1)
plot(rf1,log="x",main="black default, red samplesize, green tree depth")

#reduced sample size
rf2 = randomForest(X,y,sampsize=.1*length(y),ntree=5000)
print(rf2)
points(1:5000,rf2$mse,col="red",type="l")

#limiting tree depth (not exact )
rf3 = randomForest(X,y,maxnodes=24,ntree=5000)
print(rf2)
points(1:5000,rf3$mse,col="darkgreen",type="l")
```

edited Jan 26 at 12:00

answered Jan 26 at 10:44

Soren Havelund Welling
621 2 9

Thank you so much for the explanation. I could understand to some extent what you mean, however, since I am still getting used to this whole concept of developing random forest models, I have a few more questions based on your answer. What exactly is the tree correlation and how do you measure it? Is the OOB estimate and ensemble error the same things? Since these could be very basic, you could let me know if there is an article if I can read up to understand the terms better.Thanks a lot! – Vysh Jan 30 at 3:09

It is true that generally more trees will result in better accuracy. However, more trees also mean more computational cost and after a certain number of trees, the improvement is negligible. An article from Oshiro et al. (2012) pointed out that, based on their test with 29 data sets, after 128 of trees there is no significant improvement(which is inline with the graph from Soren).

Regarding the tree depth, standard random forest algorithm grow the full decision tree without pruning. A single decision tree do need pruning in order to overcome over-fitting issue. However, in random forest, this issue is eliminated by random selecting the variables and the OOB action.

Reference: Oshiro, T.M., Perez, P.S. and Baranauskas, J.A., 2012, July. How many trees in a random forest?. In MLDM (pp. 154-168).

answered Jan 28 at 23:24

Sharp Yan
35 7

Add Another Answer