Why and How to exploit OOB Validation for Ensemble Size

Philip Kegelmeyer, Sandia National Labs, wpk@sandia.gov

CASIS, November 16, 2007
**Traditional:** Use 100% of training data to build a sage.

**Ensembles:** Use randomized 100% of training data to build an expert. Repeat to build many experts. Vote them.

The experts beat the sage[1]!
“Bagging” is the Formal Name for This Method

Sampling with replacement
How Big An Ensemble Do You Need?

*Don’t* use fixed size ensembles. They will short-change you and deceive you. Instead, stop when accuracy levels off.

But how to measure accuracy? *Don’t* just use the training data. Use a separate validation set? Sure, but they are rare and costly. Out-of-bag (OOB) validation is easy and cheap.
Every Classifier Lacks a Fraction of the Samples

Sampling with replacement

Vote
Every Sample Lacks a Fraction of the Classifiers!!

The classifiers that didn’t see the sample can be fairly used to test it.

Sample 2 can be tested by E3 and E4; Sample 4 by E1, E2, E3 and E4. Each sample can be tested by a substantial fraction of the classifiers. So the overall accuracy is accumulated, one sample at a time.
When To Stop? When Accuracy Flattens Out

![Graph showing the relationship between model size and accuracy]

- The x-axis represents the model size, ranging from 0 to 120.
- The y-axis represents accuracy, ranging from 0 to 1.
- The graph shows a curve that flattens out at around 80, indicating that the accuracy stops improving significantly after this point.

Kegelmeyer, Why and How to Exploit OOB Validation for Ensemble Size

Page 7 of 16
But: Accuracy Can Increase Erratically!
Can’t stop at first peak or plateau; accuracy curve must be smoothed.
So Smooth . . .

Smooth with a running average over a small window $w_{\text{small}}$.

\[ w_{\text{small}} = 5 \]
...and Check “Flatness” over Broad Window

Apply “set to maximum” filter over a broad window $w_{\text{large}}$, set ensemble size to first point that achieved max accuracy.

$w_{\text{large}} = 20$
Summary: Stopping Point Selection

Three step algorithm for selecting a stop point[2]:

1. Maintain a running average over \( w_{small} \) samples, to smooth.

2. Track maximum accuracy over windows of size \( w_{large} \) until it doesn’t increase.

3. Return size of ensemble that first achieved that accuracy.

Raw Accuracy Curve

Smoothed Accuracy

Maximum Filter Accuracy
Smooth ...
...and Check Flatness over Broad Window

\[ w_{\text{large}} = 20 \]
So: Smoothed Maximum Accuracy is Effective . . .

. . . but theoretically unsatisfying.

Next Steps:

• Generate a menagerie of real curves; build intuition.

• Estimate parameters from the curve itself?
  – Extract a non-parametric measure of variability from the raw ensemble data?
  – Explicitly model the “noise”, the variation in accuracy?

• Consult with a trained 1D signal processor.