Appendix S1: Bagging may decrease predictive mean-squared error

Let the training set be \( \{ (y_i, G_0^i), i = 1, 2, ..., N_{\text{Train}} \} \), where each case is drawn from some joint distribution \( F \), and let \( \hat{\phi}(G_0^i) = \hat{\phi}^{*} \) be the predictor at some fixed focal point \( G_0^i \). Define the "average" predictor to be an average over the distribution of an infinite number of training sets \( S \)

\[
\hat{\phi}_{\text{Ave}}(G_0^i) = E_S \left[ \hat{\phi}(G_0^i) \right],
\]

with the expectation taken over conceptual repetitions of training sets of size \( N_{\text{Train}} \). For some fixed focal point and fixed output \( y \) the mean squared error is

\[
E_S \left[ y - \hat{\phi}(G_0^i) \mid y, G_0^i \right]^2 = y^2 - 2yE_S \left[ \hat{\phi}(G_0^i) \right] + E_S \left[ \hat{\phi}^2(G_0^i) \right] .
\]

(1)

Since \( \text{Var}_S \left( \hat{\phi}(G_0^i) \right) = E_S \left[ \hat{\phi}^2(G_0^i) \right] - E_S \left[ \hat{\phi}(G_0^i) \right]^2 \geq 0 \), it must be that

\[
E_S \left[ \hat{\phi}^2(G_0^i) \right] \geq E_S \left[ \hat{\phi}(G_0^i) \right]^2 ,
\]

and using this in (1) produces

\[
E_S \left[ y - \hat{\phi}(G_0^i) \mid y, G_0^i \right]^2 \geq \left( y - E_S \left[ \hat{\phi}(G_0^i) \right] \right)^2 = \left( y - \hat{\phi}_{\text{Ave}}(G_0^i) \right)^2 .
\]

(2)

Taking expectations (deconditioning) over the joint distribution \( F \)

\[
E_F E_S \left[ y - \hat{\phi}(G_0^i) \mid y, G_0^i \right]^2 \geq E_F \left( y - \hat{\phi}_{\text{Ave}}(G_0^i) \right)^2 ,
\]

showing that the mean-squared error of \( \hat{\phi}_{\text{Ave}}(G_0^i) \) over all possible input focal points-output combinations is not larger than that of \( \hat{\phi}(G_0^i) \). How much lower the mean squared error is depends on the inequality \( E_S \left[ \hat{\phi}^2(G_0^i) \right] \geq E_S \left[ \hat{\phi}(G_0^i) \right]^2 \) and the larger \( \text{Var}_S \left( \hat{\phi}(G_0^i) \right) \) is, the more improvement averaging will produce. Now, \( \hat{\phi}_{\text{Ave}}(G_0^i) = E_S \left[ \hat{\phi}(G_0^i) \right] \) depends on the true distribution from which the training set \( S \) is drawn. However, the bagged estimator is constructed from a bootstrap distribution that assigns probability \( 1/N_{\text{Train}} \) to each case in \( S \). [31] stated that if the predictor is unstable, the improvement from averaging (aggregation) may be sizable, but if the predictor is stable the bootstrap approximation to the distribution may be inaccurate. Unfortunately, "stability" is a vague and relative concept. He concludes:

"Bagging goes a long way towards making a silk purse out of a sow’s ear, especially if the sow’s ear is twitchy. It is a relatively easy way to improve an existing method..."