Machine Learning, 55, 251–270, 2004

Bagging Equalizes Influence

YVES GRANDVALET Yves.Grandvalet@hds.utc.fr
Heudiasyc, UMR CNRS 6599, Université de Technologie de Compiègne, France

Editor: Robert Schapire

Abstract. Bagging constructs an estimator by averaging predictors trained on bootstrap samples. Bagged estimates almost consistently improve on the original predictor. It is thus important to understand the reasons for this success, and also for the occasional failures. It is widely believed that bagging is effective thanks to the variance reduction stemming from averaging predictors. However, seven years from its introduction, bagging is still not fully understood. This paper provides experimental evidence supporting the hypothesis that bagging stabilizes prediction by equalizing the influence of training examples. This effect is detailed in two different frameworks: estimation on the real line and regression. Bagging's improvements/deteriorations are explained by the goodness/badness of highly influential examples, in situations where the usual variance reduction argument is at best questionable. Finally, reasons for the equalization effect are advanced. They support that other resampling strategies such as half-sampling should provide qualitatively identical effects while being computationally less demanding than bootstrap sampling.

Keywords: bagging, influence, leverage, bias/variance

1. Introduction

Bagging, introduced by Breiman in 1994 (1996a), is a procedure for building an estimator by a resample and combine technique. From an original estimator, a bagged estimator is produced by averaging several replicates trained on bootstrap samples.

A bootstrap sample (Efron & Tibshirani, 1993) is created by drawing with replacement n examples from the training set $\mathcal{T}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$. It has thus the same size as the original sample but contains replicates of some examples, while others are not represented. In bagging, bootstrap sampling is repeated many times (typically 25 or 50 times). Training is then performed on each bootstrap sample, by minimization of some empirical functional. The bagged estimate is finally obtained by averaging the resulting estimators.

In many studies, bagging decision trees, stumps, naive Bayes classifiers or neural networks almost systematically compares favorably with the original predictor, on artificial as on real data (Bauer & Kohavi, 1999; Breiman, 1996a, 1996b; Dietterich, 2000; Drucker, 1997; Maclin & Opitz, 1997; Quinlan, 1996; Schapire et al., 1998). Other ensemble methods such as boosting and arcing are often more effective in reducing prediction error, but, in situations with substantial noise, bagging performs better (Dietterich, 2000; Maclin & Opitz, 1997). Hence, it is important to understand how bagging works.

Available explanations for bagging's success are briefly reviewed in Section 2, where the present explanation is also summarized. We then provide experimental results supporting the main point of this paper: bagging equalizes the influence of points on the predictor. In these experiments, we characterize the influence of an example in the computation of the estimate, and illustrate how bagging distributes influence. Section 3 describes the effects of bagging regarding the weight given to each observation in point estimation. The transposition to regression is given in Section 4. After these experiments supporting our answer to the "how does bagging work?" question, we give the reason for "why does it work in this way?" in Section 5, and we end with a summary and a discussion in Section 6.

2. How does bagging work?

2.1. Available explanations

Breiman (1996a) states that the vital element for gaining accuracy thanks to bagging is the instability of the prediction method. A method is unstable if small perturbations of the learning set can cause significant changes in the predictor. Bagging is presented as a variance reduction procedure *mimicking* averaging over several training sets. The approximation taking place should be kept in mind, since this explanation would be trivial and definitive if averaging was performed on different training sets, but it acts on bootstrap replicates of a single training set. To quote Wolpert and Macready (1996), "the bias-plus-variance argument for bagging only suggests that bagging is worth using". For example, the work of Buja and Stuetzle (2000) on U-statistics (a family of estimates generalizing the concept of average) provides examples for which bagging is proved to increase squared bias and variance. Thus, although experimental results often show the expected variance reduction (Bauer & Kohavi, 1999; Breiman, 1996b; Schapire et al., 1998), several other stances have been explored to explain the success of bagging.

Friedman and Hall (2000) provide a theoretical analysis based on an asymptotic truncated Taylor series of the estimate. They conclude that, in the limit of infinite samples, bagging reduces the variance of non-linear components in the decomposition, while leaving the linear part unaffected. Bühlmann and Yu (2000) present another theoretical study dedicated to non-differentiable and even discontinuous predictors, where the results of Friedman and Hall do not apply. They focus on neighborhood of discontinuities of decision/regression surfaces where bagging is shown to have a smoothing effect. From these two studies and the one from Buja and Stuetzle (2000), we retain that bagging asymptotically performs some smoothing on the estimate. The latter clearly occurs also for finite samples, but it might not be the major effect. In particular, these asymptotic analyses did not address bagging's treatment of outliers, which seems to be particularly effective, as suggested by its good performance in the presence of noise (Dietterich, 2000; Maclin & Opitz, 1997).

Schapire et al. (1998) provide non-asymptotic bounds for voting algorithms, including bagging, relating the generalization performance of aggregated classifiers to the margin distribution of examples. Unlike boosting, bagging does not explicitly maximize margins, but the experiments provided by Schapire et al. suggest that bagging has a beneficial effect on the latter. The obtained bounds are acknowledged to be loose, and Breiman (1999) even claims they are qualitatively misleading, since for base classifiers of fixed VC-dimension, maximizing the margin of the aggregated classifier does not yield lowest error rates.

Domingos (1997) gives a Bayesian treatment of bagging. He performs several empirical tests pertaining to two hypotheses: either bagging works (1) because it approximates sampling from the posterior distribution or (2) because it shifts the prior to an appropriate model space. Domingos rejects the first hypothesis and accepts the second one, by verifying that bagged estimators are more complex than the original ones. However these conclusions depend strongly on the base predictor, and they rely on many assumptions which could not be verified on real data. This point of view is furthermore challenged by Rao and Tibshirani (1997), who qualify the bootstrap distribution as a "poor man's Bayes posterior" stemming from an approximation of a symmetric Dirichlet non-informative prior. Bagging is then interpreted as a Monte Carlo integration over the posterior distribution. This interpretation is used by Rao and Tibshirani to propose an averaging method resembling the Bayesian approach, but the link with a non-informative prior is of little help in understanding the reasons of bagging's success.

2.2. Present explanation

Unlike the preceding studies, the present approach does not consider the global effect of bagging on the predictor, but focuses on the potential influence of training examples in the estimation process. We argue that bagging systematically equalizes the potential influence of examples. This equalizing may lead to opposite modifications in the performance of the estimate, according to the goodness of leverage points, i.e. examples that may have a high influence on the predictor. In this respect, bagging's equalization can be interpreted as the primary effect possibly causing a secondary effect on global characteristics (e.g. variance or prediction error) of the estimator.

Leverage is illustrated in figure 1 for ordinary linear regression. A single outlier in the explicative variables can have a drastic effect in setting the coefficients of the regression line. According to the value of the explained variable, this effect can improve/deteriorate the estimator accuracy (good/bad leverage).[1]

In *most* situations, leverage is badly influential, and decaying leverage reduces the variance of the estimator. If an "unstable predictor" is defined as a predictor for which there are highly influential points, then, in *most* situations, our analysis is in accordance with

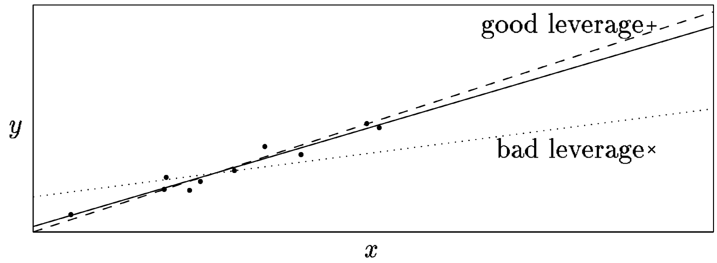

Figure 1. Least squares linear regression without the outlier in *x* (solid), with the good leverage point (dashed), and with the bad leverage point (dotted).

Breiman's: bagging stabilizes estimators and reduces variance. But the present explanation, which is not rooted in the averaging argument, also applies when bagging fails. This is illustrated in the two forthcoming Sections in point estimation and regression.

3. Point estimation

This first experiment presents a simple point estimation problem where variance reduction arguments do not hold. Instead, it is shown that bagging systematically balances the weights given to each example in the computation of the estimate. Variance reduction may occur as a consequence, but the converse may also be true.

We consider a mean estimation problem, where data are generated from a distribution "contaminated" by a widespread component centered on the same location. This type of mixture distribution is routinely used to illustrate the interest of robust statistics (Huber, 1981). In our experimental setting, $n = 20$ examples are generated by independent drawings from the mixture distribution

$$p(x) = \frac{1-P}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) + \frac{P}{\sqrt{2\pi}10} \exp\left(-\frac{x^2}{2(10)^2}\right). \tag{1}$$

The mean of the two normal components is zero, but the spread of the second one is ten times larger.

Four estimates of the mean are computed: the sample average, the sample median and their bagged estimates. The latter are obtained from 100 balanced bootstrap replications.[2] In fact, the bagged average is not computed since it is identical to the sample average; it is a linear statistics unaffected by bagging (Bühlmann & Yu, 2000; Buja & Stuetzle, 2000; Friedman & Hall, 2000).

The number of bootstrap replications is chosen arbitrarily and is unimportant for our present purpose. The experiment is repeated 1000 times on independent samples.

Table 1 reports the variance of all mean estimates for five different values of the contamination proportion P. As all estimators are unbiased in this experimental setting, their expected squared errors are equal to their variances.

The unbagged and bagged average being identical, they have the same variance.

To summarize, we observe that:

Table 1. Variance of the unbiased mean estimates.

P	Average	Median	Bagged median
0	0.050 ± 0.002	0.069 ± 0.003	0.061 ± 0.003
0.05	0.321 ± 0.020	0.077 ± 0.003	0.068 ± 0.003
0.2	1.062 ± 0.050	0.109 ± 0.005	0.102 ± 0.005
0.7	3.565 ± 0.148	1.121 ± 0.084	1.544 ± 0.091
1	5.020 ± 0.216	6.933 ± 0.296	6.143 ± 0.261

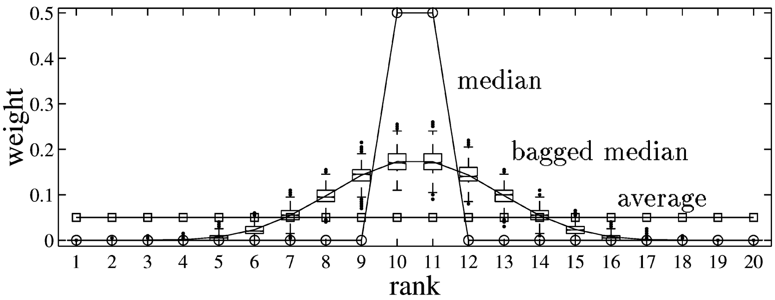

Figure 2. Boxplot of the weight given to the examples x_i versus the rank of x_i, for original and bagged mean estimates.

- bagging the average does not reduce variance;
- bagging the median sometimes reduces variance, sometimes not. Improvements are not related to the variability of the original estimate.

These results are hardly compatible with variance reduction arguments, and hence require another explanation. All previous attempts were looking at bagging's "macroscopic" effects on the estimate. Here, we focus on "atomic" effects, by looking at the weight given to each example of the training sample (i.e. the coefficient attached to one observation in the computation of the estimate).

Figure 2 summarizes the distribution, over the 1000 experiments, of the weight assigned to each example according to its rank in the original sample. On the one hand, all points contribute equally to the computation of the average, and bagging has no effect. On the other hand, the point contributions to the median are unequal. Let $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(20)}$ denote the ordered values of the x_i's, the median gives a weight of $1/2$ to $x_{(10)}$ and $x_{(11)}$, and 0 to the remaining part of the sample. Bagging balances the weights, which are spread on the neighboring order statistics of $x_{(10)}$ and $x_{(11)}$. This effect is a direct consequence of bootstrap sampling: for bootstrap sample b, $\mathcal{T}_n^b = \{x_i^b\}_{i=1}^{20}$, the median is defined as $(x_{(10)}^b + x_{(11)}^b)/2$, but $x_{(10)}^b$ and $x_{(11)}^b$ may not correspond to $x_{(10)}$ and $x_{(11)}$, which may not occur in $\mathcal{T}_n^b$, or whose ranks may be shifted due to the absence or multiple occurrence of other examples.

For all values of P, the same random draws were repeated to create bootstrap samples, thus, figure 2 refers to any line of Table 1, and displays a systematic effect which either improves or deteriorates the bagged median accuracy according to the parameter P.

For $P = 0$ and $P = 1$, the average is the efficient mean estimate: it has minimum variance among unbiased estimates. The median is extremely robust, but ignoring much of the training sample decreases its efficiency. The bagged median makes a more intensive use of the sample, which improves accuracy in low-mix settings, where only the extreme ranks are likely to be generated from the contamination component. Conversely, for $P = 0.7$, only the very middle ranks are likely to be generated from the narrow spread component. The original median being more robust to data contamination (as measured by the smallest proportion of contaminated data that can cause non-informative estimates (Rousseeuw, 1997)), it performs better than the bagged median.

To summarize, this experiment illustrates that bagging distributes the influence of examples in the computation of the estimator. The same bagging atomic effects (on the weighting of examples) can have opposite macroscopic effects (on the estimator error/variance) according to the data distribution. Bagging's observed effects are not in accordance with the usual variance reduction arguments, but they are easily explained from the influence modification viewpoint.

4. Regression

The first example showing the limits of variance reduction arguments was given by Breiman (1996a). The original aim was to illustrate the benefits of bagging for unstable estimation procedures such as subset selection. A surprising by-product is that bagging is harmful for ordinary least squares linear regression involving all variables. Breiman explains the failure of bagging by the stability of ordinary least squares: for stable procedures, averaging predictors trained on several independent datasets is better approximated by averaging over a single dataset drawn from the data distribution (original predictor), than by averaging over several bootstrap samples (bagged predictor). This statement acknowledges that the variance reduction argument reaches its limits when bagging fails. Here, we show that the modification of influence, together with the goodness of leverage points, explains bagging's success and failure.

4.1. Measuring the potential influence of examples

Most regression and discrimination techniques are based on the minimization of some criterion, where all examples are identically weighted. Bagging is however effective in these settings. The previous experiment showed that the equalizing effect acts on the weights given to examples in the computation of the estimator. These weights may dramatically differ. This phenomenon is well known in the framework of linear regression, where statistics have been devised to spot potentially highly influential examples, the so-called leverage points.

Let $\mathcal{T}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be the training set. For a linear smoother (such as a linear regression, a kernel estimate or a smoothing spline), the value of the prediction on the training sample can be written as $\hat{\mathbf{f}} = \mathbf{S}\mathbf{y}$, where $\hat{\mathbf{f}}$ is the n-dimensional vector of fitted values at $\{\mathbf{x}_i\}_{i=1}^n$ $\hat{f}_i = \hat{f}(\mathbf{x}_i)$, $\mathbf{y}$ is the n-dimensional vector of response variables and $\mathbf{S}$ is the $n \times n$ smoothing or hat matrix (Hastie & Tibshirani, 1990). The ith row of $\mathbf{S}$ represents the sequence of weights (or equivalent kernel) given to y_j, $j = 1, \ldots, n$ to define $\hat{f}(\mathbf{x}_i)$. Each element of $\mathbf{S}$ is thus relevant for our analysis, but the diagonal elements S_{ii} provide a good summary, which is commonly used to flag leverage points, since $\partial \hat{f}(\mathbf{x}_i)/\partial y_i = S_{ii}$. As trace($\mathbf{S}$) is a possible definition of the degrees of freedom of the smooth (Hastie & Tibshirani, 1990), S_{ii} can also be interpreted as the degrees of freedom spent to fit $(\mathbf{x}_i, y_i)$. Leverage points are thus associated with large S_{ii}. There is no general agreement on what "large" means, but this is not crucial for the point made here.

The smoothing matrix of the bagged estimate $\mathbf{S}^{\text{bag}}$ is defined identically by $\hat{\mathbf{f}}_{\text{bag}} = \mathbf{S}^{\text{bag}}\mathbf{y}$, where $\hat{\mathbf{f}}_{\text{bag}}$ is the vector of fitted values at $\{\mathbf{x}_i\}_{i=1}^n$. It is easily computed for any linear smoother (see Appendix A).

4.2. *Experiment*

The experimental setup (see Breiman, 1996a for more details) consists in replicating 250 times:

1. Draw samples of size $n = 60$ from the model $y = \boldsymbol{\beta}^T\mathbf{x} + \varepsilon$, where ε is drawn from a normal distribution $\mathcal{N}(0, 1)$, $\mathbf{x} \in \mathbb{R}^d$, $d = 30$ is drawn from a normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^2)$, with $\Sigma_{ij}^2 = \rho^{|i-j|}$, and ρ is drawn from a uniform distribution on [0, 1].
2. Compute the estimates of $\boldsymbol{\beta}$ for subset sizes ranging from 1 to d (subsets are determined by forward subset selection).
3. Generate 50 bootstrap samples to compute bagged estimates.

In the example presented below, $\boldsymbol{\beta}$ has 27 non-zero coefficients:

$$\beta_k = c\alpha_k, \alpha_k = \sum_{\ell=1}^{3}\{(5 - |k - \mu_\ell|)_+\}^2,$$

with $\boldsymbol{\mu} = (5, 15, 25)$ and

$$c = \sqrt{3/\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^2\boldsymbol{\alpha}},$$

so that $R^2 \simeq 0.75$. The results are qualitatively equivalent for the two other setups described in (Breiman, 1996a).

Figure 3 displays the quadratic prediction error PE according to the subset size, for the original and bagged estimates. Note that the difference in prediction error is highly in favor of bagging for little subset sizes, and that it decreases as the subset size increases. There is a cross-over point past which bagging is detrimental.

The difference in prediction error for ordinary least squares (with all the variables entering the subset) could seem to be an artifact, since it contradicts the analysis of Friedman and Hall (2000), by which linear predictors are not affected by bagging. However, this analysis does not apply here, as the OLS estimate is only linear in $\mathbf{y} = (y_1, \ldots, y_i, \ldots, y_n)^T$, and not in $\mathbf{z}_i = (\mathbf{x}_i, y_i)$. In the regression framework, linear predictors in the sense of Friedman and Hall can be obtained if $\mathbf{x}_i$ are fixed, in which case bootstrap sampling consists in resampling residuals $y_i - \hat{f}(\mathbf{x}_i)$ instead of resampling the training pairs $(\mathbf{x}_i, y_i)$.

4.3. *Bagging's effect on bias and variance*

First, we stress that the failure of bagging in ordinary least squares regression can be proved to be due to an *increase* of variance. Let $\mathbf{X}$ denotes the $(n \times d)$ matrix of explicative variables

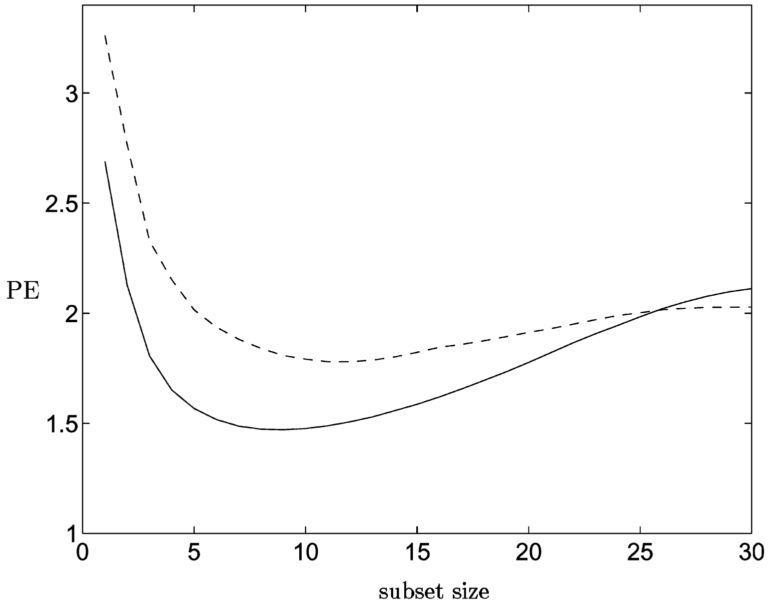

Figure 3. Prediction error (averaged over 250 experiments) for forward subset selection (dashed) and bagged forward subset selection (solid) vs. subset size.

in the training sample. The bagged estimate of the regression coefficient β is derived from the general form of smoothing matrices for bagged estimates provided in Appendix A:

$$\hat{\beta}^{\text{bag}} = \frac{1}{B}\sum_{b=1}^{B}(\mathbf{X}^T\mathbf{W}^b\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^b\mathbf{y}, \tag{2}$$

where B is the number of bootstrap samples and $\mathbf{W}^b$ is the diagonal matrix whose ith element counts the number of occurrence of point i in bootstrap sample b.

Let Y denote the $n \times 1$ random variable of explained variables, $Y = \mathbf{X}\beta + \varepsilon$. As $\mathbb{E}(\varepsilon) = 0$, we have

$$\begin{aligned}\mathbb{E}(\hat{\beta}^{\text{bag}}|\mathbf{X}) &= \frac{1}{B}\sum_{b=1}^{B}(\mathbf{X}^T\mathbf{W}^b\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^b\mathbb{E}(Y\,|\,\mathbf{X})\\ &= \frac{1}{B}\sum_{b=1}^{B}(\mathbf{X}^T\mathbf{W}^b\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^b\mathbf{X}\beta\\ &= \beta.\end{aligned} \tag{3}$$

As the ordinary least squares estimate and the bagged estimate are unbiased, an increase of variance is the only possible explanation for the increase in prediction error observed in the experiments. Furthermore, noting that Eq. (2) shows that $\hat{\beta}^{\text{bag}}$ is a linear transformation of $\mathbf{y}$, the increase of variance can also be explained by the Gauss-Markov theorem

(Saporta, 1990). In the current experimental setup, the latter states that the ordinary least squares estimate is, among all unbiased estimates which are linear in **y**, the one with minimum variance.

Theoretical analysis does not provide an analytic expression for the bias and variance of the subset selection model. Their plug-in estimates show however that prediction error is dominated by bias. Bagging hardly affects the latter (the squared bias is 1.4 with bagging compared to 1.5 without), but is effective at reducing variance (the variance is 0.2 with bagging compared to 0.7 without).

4.4. Potential influence equalization

The S_{ii} statistics are computed for ordinary least squares. Subset selection is *not* a linear predictor, since the observations y_i take part in the subset choice. Influence is thus measured by a generalized leverage statistic $\tilde{S}_{ii}$, which is based on the data perturbation approach described in Appendix B.

Figure 4 compares the histograms of S_{ii} and S_{ii}^{bag} computed for all 250 training samples. The histogram obtained for ordinary least squares (left) shows that the distribution is unimodal (there are no gross outliers), centered on 0.5 (30 free parameters are to be set by 60 points). Bagging yields also a unimodal distribution, centered on 0.5, which renders that complexity is not modified by bagging. The spread is however sensibly reduced (the standard deviation is divided by two): the influence of each point on the estimator is equalized.

The histograms are more complex for subset selection (right). First, note that a log-scale is used to highlight that the distribution for subset selection is bimodal (on a linear scale, the high spread of the minor mode makes it hardly visible). The main mode contains about 90% of points with small $\tilde{S}_{ii}$ values (mean slightly less than 1/60); it gathers points with little influence on the choice of the element entering the subset, but which intervene in the tuning of the single regression coefficient. The other mode contains highly influential

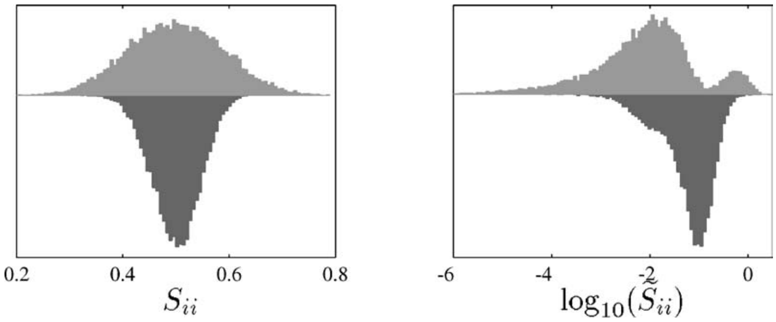

Figure 4. Left: histograms of S_{ii} (up, light grey) and S_{ii}^{bag} (down, dark grey) for ordinary least squares (all variables); right: histograms of $\log_{10}(\tilde{S}_{ii})$ (up, light grey) and $\log_{10}(\tilde{S}_{ii}^{\text{bag}})$ (down, dark grey) for subset selection (one variable).

points, with a mean $\tilde{S}_{ii}$ value of about 0.6. These points play a leading part in choosing the variable entering the subset. Only one variable is selected, but it brings an average of 4.6 degrees of freedom: this inflation in degrees of freedom is due to the subset choice, as discussed in detail in Tibshirani and Knight (1999). The distribution for the bagged subset selection estimate is unimodal, centered on 0.1, with an average 5.9 degrees of freedom. Thus, bagging increases slightly the complexity of the predictor (an average of nine different variables enter the subset, but the regression coefficients are not computed by least squares). The spread of $\tilde{S}_{ii}^{\mathrm{bag}}$ is also halved: leverage statistics are equalized. We now verify that, in the present setting, influence equalizing happens to be beneficial for subset selection and detrimental for ordinary least squares.

4.5. *From influence equalization to prediction error*

For ordinary least squares, the *expected* difference in prediction error between the ordinary and the bagged estimate is a quadratic function in smoothing matrices $\mathbf{S}$ and $\mathbf{S}^{\mathrm{bag}}$. In the present setup, it is however difficult to exhibit a clear-cut relationship between S_{ii} equalization and the actual changes in prediction error. Bagging's action on potential influence may have some outcome on the effective influence, which may in turn have positive or negative consequences on prediction error according to the goodness/badness of leverage. Furthermore, goodness/badness does not describe an intrinsic quality of a single point with respect to a predictor. It is defined relatively to a learning set and is subject to interactions: two leverage points may be badly influential separately and beneficial jointly.[3] The relationship between influence equalization and prediction error is shown here in a setup where the relationships in the chain linking potential influence to goodness/badness are controlled.

Figure 5 displays maps of differences in prediction error $\mathrm{PE} - \mathrm{PE}^{\mathrm{bag}}$ according to potential influence S_{ii} and badness of examples. Positive values thus correspond to bagging's improvements and negative ones to failures. To control goodness and avoid interactions, we first set $y_i = \boldsymbol{\beta}^T \mathbf{x}_i$, $i = 1, \ldots, n$ which is the ultimate goodness for all examples. The badness for example i is then controlled by adding a perturbation to y_i while $y_j = \boldsymbol{\beta}^T \mathbf{x}_j$, $j \neq i$. The magnitude of this perturbation is reported on the badness axis. The maps are produced

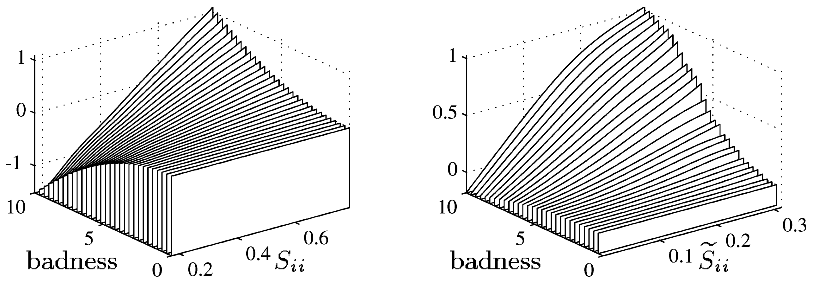

Figure 5. Maps of difference in prediction error $\mathrm{PE} - \mathrm{PE}^{\mathrm{bag}}$ vs. badness and leverage statistics, left: for ordinary least squares (all variables); right: for subset selection (one variable).

by curves detailing the effects of influence equalization according to the influence statistic of the contaminated point at the given badness level. These curves are obtained by a local line smoother on the 60 × 250 points (on the 250 experiments one example is perturbated at a time). They are faithful summaries: they explain most of total variance ($R^2 \simeq 70\%$).

The OLS predictors being linear, S_{ii} and S_{ii}^{bag} are not affected by badness. Hence, for each badness value, the marginalized influence histogram of S_{ii} and S_{ii}^{bag} is identical to the one displayed in figure 4. Furthermore, the same trend is obtained for all badness levels, which have only a multiplicative impact on the difference in prediction error. For one level, the difference in prediction error varies almost linearly with the S_{ii} statistics. When perturbations are applied to the most influential point, bagging is highly beneficial, when they are applied to the less influential ones, it is highly detrimental. When all examples are corrupted identically, the overall result is slightly negative.

For subset selection, $\tilde{S}_{ii}$ and $\tilde{S}_{ii}^{\text{bag}}$ are affected by badness, but influence equalization is similar for all badness values (not shown here). Here, we observe slight differences in trend according to badness, with a saturation effect for highly perturbated influential points. Bagging is highly beneficial when the most influential points are perturbed, and slightly detrimental in the opposite situation. Overall, when all examples are corrupted identically, the outcome is beneficial.

To summarize this section, we illustrated that bagging has an equalizing effect on leverage statistics which improves or deteriorates estimation in the presence of respectively bad or good leverage points. For ordinary least squares, bagging increased variance. This negative effect is explained by the experimental setting in which the original predictor is optimal within a set including bagged ordinary least squares. In setups including bad leverage points, ordinary least squares benefits from bagging. For subset selection, the fact that more variables are used in the bagged estimate is not essential to bagging improvements, as bias is not affected in the process. The important point is that one point cannot be influential in all bootstrap samples; hence, more points intervene in the subset choice and variance is reduced.

5. A paradox explained

In bagging, all examples being represented the same number of times in bootstrap samples as a whole, it may seem paradoxical that their influence is modified. The distribution of influence in bootstrap samples provides an explanation for this phenomenon. It furthermore indicates that bootstrap sampling, which is a key element according to the "reduce variance by averaging" argument, is not essential to the algorithm effectiveness.

5.1. *Distribution of influence in bootstrap samples*

The distribution of influence in bootstrap samples is illustrated here in smoothing spline regression. This experiment enables to display two types of leverage, without interaction. Furthermore, the effects can be visualized and leverage statistics are rapidly computed (see Section 4.1).

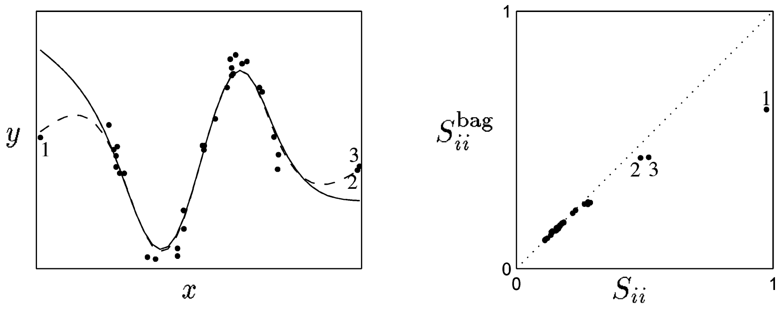

Figure 6. Left: cubic regression spline (dashed line) and bagged estimate (solid line); right: diagonal elements of the bagged smoothing matrix $\mathbf{S}^{\text{bag}}$ versus diagonal elements of the smoothing matrix $\mathbf{S}$.

We chose a large number of bootstrap replications (1000) to depict the distributions of bootstrap statistics. Other details of the experimental setup are omitted as they are not important for interpretation. Figure 6 displays the effects of bagging on a cubic regression spline.

The right hand side of figure 6 displays the diagonal elements of the bagged smoothing matrix. For the three leverage points (at the boundary) marked 1–3, S_{ii}^{bag} is smaller than S_{ii}, while the changes are minor for the other points. As a result, the bagged fit is hardly changed for internal values of x and clearly looser for the boundary points.

For a smoothing spline, S_{ii}^{bag} is the average of S_{ii}^{b} (see Appendix A), thus, the distribution of leverage statistics on bootstrap samples details how bagging affects influence. Figure 7 displays the cumulative distributions of S_{ii}^{b} for the three leverage points. The step at zero gathers the bootstrap samples where the considered point is absent, resulting in a null

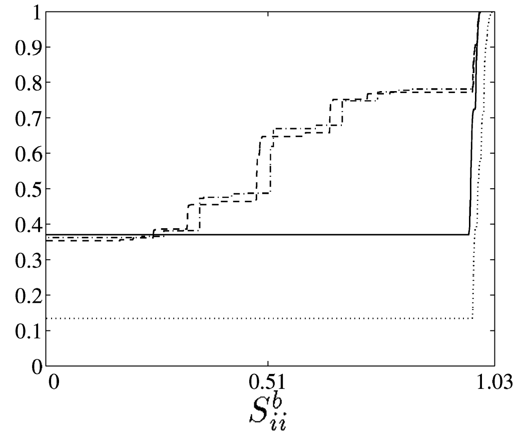

Figure 7. Cumulative distribution function of bootstrap leverage statistics S_{ii}^{b} for point 1 (solid), 2 (dashed), 3 (dashdot) and for the sum $S_{22}^{b} + S_{33}^{b}$ (dotted).

influence. Its height is close to the probability (with respect to random bootstrap sampling) of this event ($(1-1/n)^n \simeq 0.37$).

For the isolated outlier 1, the function shows a second step about the leverage statistic of the unbagged smooth $S_{11} = 0.97$. This second mode of the distribution gathers all bootstrap samples where point 1 is present: the number of replications of the example plays a minor role in the computation of S^b_{11}.

For the grouped outliers 2–3, the distribution has numerous modes, as shown by all the small steps of the cumulative distribution. They correspond to constant ratio of $w^b_2/(w^b_2 + w^b_3)$, where w^b_i denotes the number of replications of example i in bootstrap sample b. However, regarding prediction, the influences of points 2 and 3 should not be differentiated, since they have nearly identical x-values, and thus correspond to a single influential cluster whose influence is quantified by the sum $S^b_{22} + S^b_{33}$. The distribution of $S^b_{22} + S^b_{33}$ is bimodal. The prior probability of the first mode is sensibly lower than for single points, since it corresponds to the probability of absence of the whole cluster in a bootstrap sample ($(1 - 2/n)^n \simeq 0.14$).

Points 2–3 illustrate a *masking* effect: bagging down-weights more efficiently a single leverage point than a cluster thereof. The distribution of S^b_{ii} explains that this masking stems from the proportion of bootstrap samples not containing a cluster: the probability for m given points not to be in one bootstrap sample is $1-(1-m/n)^n \simeq 1-\exp(-m)$, hence $S^{\text{bag}}_{11} \simeq 0.63 \times S_{11}$, and $(S^{\text{bag}}_{22} + S^{\text{bag}}_{33}) \simeq 0.86(S_{22} + S_{33})$. The other influence statistics are hardly affected by bagging, since a typical inner point belongs to a large cluster for which $1-(1-m/n)^n \simeq 1$.

The effectiveness of bagging is more tightly related to the probability of absence/presence of examples than to their multiple occurrences because replicates have usually little effect on predictors compared to absence/presence. This argument is trivially exact for interpolation, where having several copies of one point does not change the solution, while deleting one point has important consequences. A similar behavior can be expected for the complex predictors which are recommended in bagging (Taniguchi & Tresp, 1997). The question thus arises whether bootstrap sampling is really a key ingredient of the procedure.

5.2. *Is bootstrap sampling a key ingredient of bagging?*

Regarding the probability of absence/presence of clusters, sampling with replacement a proportion P of the original dataset is asymptotically equivalent to sampling without replacement a proportion $1-\exp(-P)$. This is illustrated in figure 8 where the probabilities of absence are reported for a sample size of 50.

We provide here an additional clue supporting that the main effect of bagging stems from the presence/absence of points in bootstrap samples. A predictor trained on a bootstrap sample can be viewed as the original predictor contaminated by case-weight perturbation. Case-weight is the multiplicative coefficient weighting each example in the criterion used to define the predictor; it may be not related to influence.

In bootstrap sample b, the case-weight for example i is its number of occurrence w^b_i. We test two alternative case-weights:

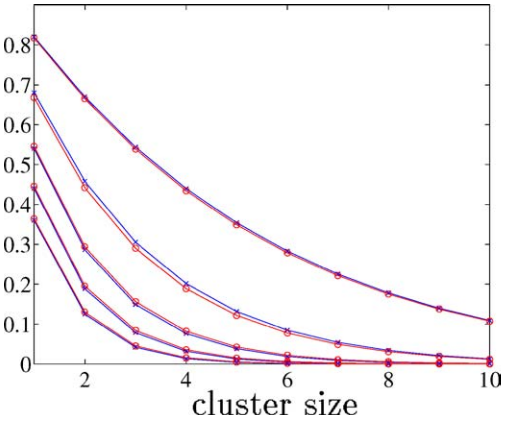

Figure 8. Probability of absence of a cluster vs. cluster size for subsampling with (○) and without (×) replacement for a training sample size of $n = 50$. From top to bottom, subsample size is $n' = 10, 20, 30, 40, 50$ with replacement, and $9, 16, 23, 28, 32 \simeq n(1 - \exp(-n'/n))$ without replacement.

1.

$$w_i^{b\prime} = w_i^b + \frac{\mathbf{I}_{[w_i^b=0]}}{\sum_{i=1}^n \mathbf{I}_{[w_i^b=0]}},$$

where $\mathbf{I}_{[w_i^b=0]}$ is one iff example i does not appear in bootstrap sample b, does not affect the number of replicates, except that absence is replaced by presence with a small case-weight (here, this average case-weight is below 0.1); for any ℓ^p norm, $\|\mathbf{w}^{b\prime} - \mathbf{w}^b\|_p \leq 1$.

2. $w_i^{b\prime\prime} = \mathbf{I}_{[w_i^b>0]}$ does not modify absence/presence, but transforms all multiple occurrences in a single one; for any ℓ^p norm, $\|\mathbf{w}^{b\prime\prime} - \mathbf{w}^b\|_p > 1$.

Figure 9 shows how the aggregated predictors are affected by these choices. The comparison with figure 6 clearly shows that the solution obtained with $\mathbf{w}^{b\prime\prime}$ is nearly identical to

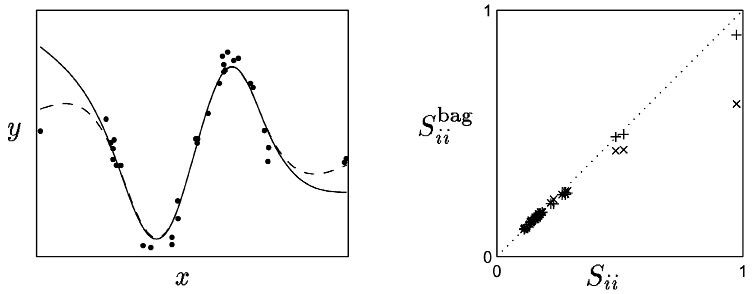

Figure 9. Left: pseudo-bagged cubic regression spline estimates for $\mathbf{w}^{b\prime}$ (dashed line) and $\mathbf{w}^{b\prime\prime}$ (solid line); right: diagonal elements of the respective pseudo-bagged smoothing matrices $\mathbf{S}^{\mathrm{bag}\prime}$ (+) and $\mathbf{S}^{\mathrm{bag}\prime\prime}$ (×) versus diagonal elements of the smoothing matrix $\mathbf{S}$.

the solution obtained with bagging ($\mathbf{w}^b$), while $\mathbf{w}^{b'}$ yields a solution which is closer to the original predictor than to its bagged version. As $\mathbf{w}^b$ is closer to $\mathbf{w}^{b'}$ than to $\mathbf{w}^{b''}$ for any ℓ^p norm, the importance of absence/presence relative to number of replicates is demonstrated.

The interpretation of bootstrap sampling as a case-weight perturbation technique suggests another interpretation of bagging. Suppose we want an estimator to be robust to some predefined perturbations of the dataset. This goal can be reached either by modifying the training criterion, or by constructing a predictor with in-built invariance properties which automatically fulfills these constraints. A very simple means to create in-built invariance is to define the predictor as the expectation of predictors trained on perturbed data. This expectation can be approximated by the average on many surrogate samples, made of perturbed replications of the examples. This heuristic has been successfully applied to robustness constraints regarding input perturbations (Raviv & Intrator, 1996), and output perturbations (Breiman, 1996b, 1998). Bagging can be interpreted as another variation on the same theme, connected to robustness regarding case-weight perturbations. Bootstrap sampling then appears as one possibility among others (e.g. leave-one-out or K-fold cross-validation) to provide balanced case-weight perturbations.[4]

5.3. Relation to previous work

Friedman and Hall (2000) already proposed to replace bootstrap sampling by half sampling. Their analysis, confirmed by Buja and Stuetzle (2000), show the equivalence (in terms of asymptotical bias, variance and expected quadratic error) between sampling with replacement a proportion P and sampling without replacement a proportion $P/(P+1)$. For $P < 1$, the present proposal $1-\exp(-P)$ is slightly larger than $P/(P+1)$, with differences increasing with P.

A comparison was conducted in the regression example of Section 4, where the estimation of β was stabilized by a ridge penalizer in order to avoid singular or numerically ill-conditioned design matrices. The number of bootstrap samples was increased to 1000 to approach more closely the idealized bagging studied in theoretical analyses.

Figure 10 reports the differences in prediction error between bagging with replacement a proportion P (for $P = 0.2$ and $P = 1$) and sampling without replacement a proportion $P/(P+1)$ or sampling without replacement a proportion $1 - \exp(-P)$. Regarding mean prediction error, the solutions provided by the three resampling schemes are rather close to each other, but the approximation preserving the probability of absence is slightly more accurate than the one derived from asymptotical analyses.

The relative failure of the theoretically motivated $P/(P+1)$ can be explained by two deviations from its application domain. First, the number of free parameters of the predictor is not overwhelmed by the number of examples: we are far from asymptotics. Second, up to now, no theory applies to subset selection which is an instable estimate with a multimodal distribution; only the OLS predictor is continuous (and linear in response variables).[5]

Note that the experimental comparison of Friedman and Hall (2000) between resampling with and without replacement in non-linear CART regression is also out of the scope of their theoretical analysis. Some of the curves they displayed might be explained by the absence of the examples in 37% of bootstrap samples.

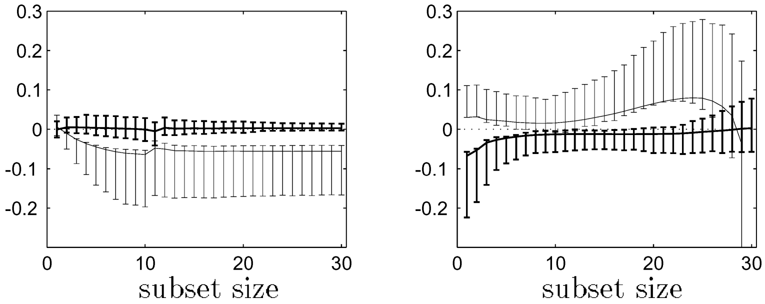

Figure 10. Average difference in prediction error between bagged regressors obtained by sampling with and without replacement vs. number of selected variables. A proportion P is sampled with replacement, and a proportion $P/(P+1)$ (thin line), or $1-\exp(-P)$ (bold line) is sampled without replacement. The bars cover 90% of the distributions of differences. Left: $P = 0.2$, right: $P = 1$.

6. Discussion

The experiments provided in this paper illustrated that bagging equalizes the influence of examples in setting a predictor: fewer points have a small influence, while the highly influential ones are down-weighted. Hence, bagging is useless when all examples have the same influence on the original estimate, is harmful when high impact points improve accuracy, and is otherwise beneficial.

Our analysis supports one of Breiman's (1996a) early statements: the vital element for gaining accuracy thanks to bagging is the instability of the prediction method. However, instability is no more related to the intrinsic variability of the predictor, but to the presence of influential examples in a dataset for a given prediction method. In many situations, highly influential points are outliers, and their down-weighting reduces the variance of the estimator. But our explanation, which is not rooted in an averaging argument, also applies when bagging fails: the stabilization effect of bagging is harmful when estimation accuracy benefits from influential points (the so-called "good" leverage points).

With the influence equalization viewpoint, bagging is interpreted as a perturbation technique aiming at improving the robustness against outliers. Indeed, averaging predictors trained on perturbed training samples is a means to favor invariance to these perturbations. Bagging applies this heuristic to case-weight perturbations. Bootstrap sampling is a central element of bagging according to the "reduce variance by averaging" argument, but influence equalization is mainly due to the absence of influential examples in 37% of bootstrap samples. Nearly identical results are obtained by replacing bootstrap sampling by resampling without replacement 63% of the training set. The effectiveness of other resample and combine schemes (Friedman & Hall, 2000; Buja & Stuetzle, 2000; Bühlmann & Yu, 2000) can be understood within this perspective.

It should be however stressed that highly influential examples are down-weighted but not suppressed: a single influence point can still have a high impact on a bagged estimate. Furthermore, clusters of leverage points are not efficiently down-weighted. Finally, the influence increase of inactive points induces a decrease of robustness regarding the number of contaminated examples which is tolerated by the predictor. The question thus arises whether bagging should be a recommended alternative to robust estimation in noisy settings. For linear predictors, standard robust estimation is well developed and should be preferred. On the other hand, for non-linear predictors, bagging is an appealing procedure, as it proceeds automatically without requiring to flag influential points by intensive computation.

Appendix

A. Smoothing matrix for bagged predictors

The smoothing matrix of a bagged estimator is easily computed from the smoothing matrix of the original estimate. All smoothing matrices have the same general structure, since linear smoothers explicitly or implicitly minimize a penalized weighted residual sum of squares

$$\sum_{i=1}^{n} v_i \left(\sum_{j=1}^{m} c_j g_j(\mathbf{x}_i) - y_i \right)^2 + \lambda \sum_{j=1}^{m} \sum_{k=1}^{m} \Omega_{jk} c_j c_k, \tag{4}$$

where v_i are positive real coefficients, $\{g_j\}_{j=1}^m$ is a set of functions whose choice is independent of the training set $\mathcal{T}_n = \{\mathbf{x}_i, y_i\}_{i=1}^n$, λ is a positive coefficient or hyper-parameter, and Ω_{jk} are the coefficients of a positive definite matrix $\boldsymbol{\Omega}$.

The n-dimensional vector of fitted values at $\{\mathbf{x}_i\}_{i=1}^n$ is given by:

$$\begin{aligned} \hat{\mathbf{f}} &= \mathbf{G}\hat{\mathbf{c}} \\ &= \mathbf{G}(\mathbf{G}^T\mathbf{V}\mathbf{G} + \lambda\boldsymbol{\Omega})^{-1}\mathbf{G}^T\mathbf{V}\mathbf{y} \\ &= \mathbf{S}\mathbf{y}, \end{aligned} \tag{5}$$

where $G_{ij} = g_j(\mathbf{x}_i)$ and $\mathbf{V}_{ij} = v_i\delta_{ij}$. Hence the smoothing matrix is $\mathbf{S} = \mathbf{G}(\mathbf{G}^T\mathbf{V}\mathbf{G} + \lambda\boldsymbol{\Omega})^{-1}\mathbf{G}^T\mathbf{V}$.

Identically, the smoothing matrix $\mathbf{S}^b$ of the estimator on a bootstrap sample $\mathcal{T}_n^b$ is $\mathbf{S}^b = \mathbf{G}(\mathbf{G}^T\mathbf{V}\mathbf{W}^b\mathbf{G} + \lambda\boldsymbol{\Omega})^{-1}\mathbf{G}^T\mathbf{V}\mathbf{W}^b$, where $\mathbf{W}^b$ is a diagonal matrix whose ith element counts the number of occurrence of point i in $\mathcal{T}_n^b$. As bagging performs an average of predictors trained on bootstrap samples, its smoothing matrix $\mathbf{S}^{\text{bag}}$ is the average of the smoothing matrices $\mathbf{S}^b$ obtained on these samples:

$$\mathbf{S}^{\text{bag}} = \mathbf{G}\frac{1}{B}\sum_{b=1}^{B}(\mathbf{G}^T\mathbf{V}\mathbf{W}^b\mathbf{G} + \lambda\boldsymbol{\Omega})^{-1}\mathbf{G}^T\mathbf{V}\mathbf{W}^b. \tag{6}$$

Formulae for $\mathbf{S}$, $\mathbf{S}^b$, and $\mathbf{S}^{\text{bag}}$ explicitly state that for linear smoother, the original, bootstrapped and bagged estimates all map $\mathbf{y}$ to the space spanned by $\mathbf{G}$, but differ in how $\mathbf{y}$ is projected onto $\mathbf{G}$.

For the smoothing spline experiment of Section 5, g_j was chosen to be the j^{th} element of the natural spline basis, $\mathbf{V}$ is the identity matrix, and $\boldsymbol{\Omega}$ is the penalizer of second-order derivatives.

B. A generalized leverage statistic

The S_{ii} statistics can be only computed for linear smoothers. Related statistics can be computed by a data perturbation approach (Burgess, 1997): as in the linear case S_{ii} is equal to $\partial \hat{f}(\mathbf{x}_i)/\partial y_i$, $\Delta \hat{f}(\mathbf{x}_i)/\Delta y_i$ can generalize S_{ii} to non-linear smoothers. However the result depends on the sign and magnitude of perturbations ($\Delta \hat{f}(\mathbf{x}_i)$ is a non-linear function of Δy_i). A possible choice for the latter can be derived from another property. For linear smoothers, S_{ii} can also be computed from the so-called jackknife estimates $\hat{f}^{-i}(\mathbf{x}_i)$:

$$S_{ii} = (\hat{f}(\mathbf{x}_i) - \hat{f}^{-i}(\mathbf{x}_i))/(y_i - \hat{f}^{-i}(\mathbf{x}_i)), \tag{7}$$

where $\hat{f}^{-i}$ is the predictor obtained from the training sample $\mathcal{T}_n^{-i} = \{\mathbf{x}_j, y_j\}_{j \neq i}$.

Equation (7) defines a particular perturbation: when y_i is replaced by $\hat{f}^{-i}(\mathbf{x}_i)$ ($\Delta y_i = y_i - \hat{f}^{-i}(\mathbf{x}_i)$) then the predictor evaluated at $\mathbf{x}_i$ is $\hat{f}^{-i}(\mathbf{x}_i)$ ($\Delta \hat{f}(\mathbf{x}_i) = \hat{f}(\mathbf{x}_i) - \hat{f}^{-i}(\mathbf{x}_i)$). This relationship is easily obtained from the general form (4) of the criteria minimized by linear smoothers. It suggests a generalized leverage statistic, based on the data perturbation approach where $\Delta y_i = y_i - \hat{f}^{-i}(\mathbf{x}_i)$. The algorithm is as follows:

1. Compute the jackknife estimate $\hat{f}^{-i}(\mathbf{x}_i)$;
2. Create the perturbed training sample $\{(\mathbf{x}_j, y_j)\}_{j \neq i} \cup (\mathbf{x}_i, \hat{f}^{-i}(\mathbf{x}_i))$, from which the perturbed solution $\tilde{f}^{-i}$ is computed;
3. Compute the generalized leverage statistic $\tilde{S}_{ii}$ as $(\hat{f}(\mathbf{x}_i) - \tilde{f}^{-i}(\mathbf{x}_i))/(y_i - \hat{f}^{-i}(\mathbf{x}_i))$.

For a linear smoother $\tilde{S}_{ii}$ is equal to S_{ii}. Note that many estimates (including subset selection but generally excluding bagged estimates) are invariant to the addition of new examples lying on the regression surface. In this case, step 2 can be avoided as $\tilde{f}^{-i}$ is identical to $\hat{f}^{-i}$.

Notes

[1] For a deterministic estimation procedure, the prediction error PE is a function of the training set $\mathcal{T}_n$. Formally, example i is a good leverage point if $\text{PE}(\mathcal{T}_{n-1}^{-i}) \gg \text{PE}(\mathcal{T}_n)$, where $\mathcal{T}_{n-1}^{-i}$ is the training set $\mathcal{T}_n$ deprived of example i; example i is a bad leverage point if $\text{PE}(\mathcal{T}_{n-1}^{-i}) \ll \text{PE}(\mathcal{T}_n)$.

[2] Balanced bootstrap (Efron & Tibshirani, 1993), which is used throughout this paper, ensures that all observations appear exactly the same number of times over all bootstrap samples. The usual bootstrap sampling is only approximately balanced and may introduce artificial bias in our influence estimates.

[3] For example, consider two leverage points such that $\mathbf{x}_i = -\mathbf{x}_j$, $y_i = \beta^T \mathbf{x}_i + \varepsilon$, $y_j = \beta^T \mathbf{x}_j + \varepsilon$. If ε is large, each one of these two examples is badly influential for OLS, but jointly they have a good influence on the predictor.

4. This point of view suggests that input, output and case-weight perturbations could be applied together, but this is not tested in this paper.
5. The $P/(P+1)$ approximation is poorer for OLS than for subset selection predictors, but this may be due to the increase in degrees of freedom which moves the setup further away from asymptotics.

References

Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning, 36:1/2*, 105–139.

Breiman, L. (1996a). Bagging predictors. *Machine Learning, 24:2*, 123–140.

Breiman, L. (1996b). *Bias, variance, and arcing classifiers*. Technical Report 460, Statistics Department, University of California at Berkeley.

Breiman, L. (1996c). Heuristics of instability and stabilization in model selection. *The Annals of Statistics, 24:6*, 2350–2383.

Breiman, L. (1998). Arcing classifiers. *The Annals of Statistics, 26:3*, 801–849.

Breiman, L. (1999), Prediction games and arcing algorithms. *Neural Computation, 11:7*, 1493–1517.

Bühlmann, P., & Yu, B. (2000). *Explaining Bagging*. Technical Report 92, Seminar für Statistik, ETH, Zürich.

Buja, A., & Stuetzle, W. (2000). *The effect of bagging on variance, bias and mean squared error*. Technical Report, AT&T Labs-Research.

Burgess, A. N. (1997). Estimating equivalent kernels For neural networks: A data perturbation approach. In M. Mozer, M. Jordan, & T. Petsche (Eds.), *Advances in Neural Information Processing Systems 9* (pp. 382–388). MIT Press.

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization. *Machine Learning, 40:2*, 1–19.

Domingos, P. (1997). Why does bagging work? A Bayesian account and its implications. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining* (pp. 155–158). AAAI Press.

Drucker, H. (1997). Improving regressors using boosting techniques. In *Proceedings of the Fourteenth International Conference on Machine Learning* (pp. 107–115). Morgan Kaufmann.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*, Vol. 57 of *monographs on statistics and applied probability*. Chapman & Hall.

Friedman, J. H., & Hall, P. (2000), *On bagging and non-linear estimation*. Technical Report, Stanford University, Stanford, CA.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*, Vol. 43 of *monographs on statistics and applied probability*. Chapman & Hall.

Huber, P. J. (1981). *Robust statistics*. Wiley.

Maclin, R., & Opitz, D. (1997). An empirical evaluation of bagging and boosting. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence* (pp. 546–551). AAAI Press.

Quinlan, J. R. (1996). Bagging, boosting, and C4.5. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence* (pp. 725–730). AAAI Press.

Rao, J. S., & Tibshirani, R. J. (1997). *The out-of-bootstrap method for model averaging and selection*. Technical Report, University of Toronto.

Raviv, Y., & Intrator, N. (1996). Bootstrapping with noise: An effective regularization technique. *Connection Science, 8:3*, 355–372.

Rousseeuw, P. J. (1997). Robust regression, positive breakdown. In S. Kotz, C. Read, & D. Banks (Eds.), *Encyclopedia of statistical sciences*. Wiley.

Saporta, G. (1990). *Probabilités, analyse des données et statistique*. Paris: Editions Technip.

Schapire, R., Freund, Y., Bartlett, P., & Lee, W. S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics, 26:5*, 1651–1686.

Taniguchi, M., & Tresp, V. (1997). Averaging regularized estimators. *Neural Computation, 9:7*, 1163–1178.

Tibshirani, R. J., & Knight, K. (1999). The covariance inflation criterion for adaptive model selection. *Journal of the Royal Statistical Society, B, 61:3*, 529–546.

Wolpert, D. H., & Macready, W. G. (1996). *Combining stacking with bagging to improve a learning algorithm*. Technical Report SFI-TR-96-03-123, Santa Fe Institute.

Received June 6, 2001
Revised March 8, 2002
Accepted August 3, 2002
Final manuscript October 2, 2002