Stability of Bagged Decision Trees (⋆)

Stabilità di alberi di decisione con il bagging

Yves Grandvalet
Heudiasyc, UMR CNRS 6599, Université de Technologie de Compiègne, BP 20.529,
60205 Compiègne cedex, France
e-mail: Yves.Grandvalet@utc.fr

Riassunto: Il Bagging è una tecnica di aggregazione, in cui uno stimatore viene ottenuto come media di predittori calcolati su campioni bootstrap. Gli alberi di decisione con il bagging quasi sempre migliorano il predittore originario, ed è opinione comune che l'efficacia del bagging sia dovuta alla riduzione della varianza. In questo lavoro mostriamo un contro-esempio e diamo evidenza sperimentale al fatto che il bagging stabilizza la predizione bilanciando l'influenza delle unità di training. Unità molto influenti vengono pesate di meno a causa della loro assenza in alcuni dei campioni di bootstrap. Abbiamo quindi testato empiricamente alcune recenti teorie che mettono in relazione stabilità ed errore di generalizzazione. L'ipotesi di stabilità è stata valutata in base a diversi benchmark, e mostriamo come il processo di bilanciamento dell'influenza dei singoli esempi migliora significativamente la stabilità, che a sua volta può migliorare la capacità di generalizzazione.

Keywords: Bagging, Influence, Leverage, Bias/Variance, Margin

1. Introduction

Among the numerous applications of the non-parametric bootstrap, bagging has a very special place, since bootstrapping is used to define a predictor, instead of characterizing its distribution. Bagging, introduced by Breiman (1996a), stands for Bootstrap AGGregatING. It is a very simple ensemble technique, where a bagged estimator is produced by averaging several predictors fitted to bootstrap samples

A bootstrap sample (Efron and Tibshirani 1993) is created by drawing with replacement n examples from a learning set of size n. A bootstrap sample has thus the size of the original sample and it contains several replicates of some examples, while others are not represented. In bagging, bootstrap sampling is repeated many times (typically 25 or 50 times). A learning algorithm is applied on each bootstrap sample, and the bagged estimate is obtained by averaging the resulting estimators.

Numerous experimental studies showed that bagging is very effective at reducing the generalization error of decision trees, stumps, naive Bayes classifiers or neural networks (Bauer and Kohavi 1999, Breiman 1996a,b, Dietterich 2000, Drucker 1997, Maclin and Opitz 1997, Quinlan 1996, Schapire et al. 1998). The almost systematical improvements observed in these studies motivated several theoretical works aiming at accounting for bagging's success. One such line of research is the analysis of the generalization ability of learning algorithms characterized by notions of stability (Bousquet and Elisseeff 2002,

(⋆) This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence IST-2002-506778. This publication only reflects the authors' views.

Poggio et al. 2004, Evgeniou et al. 2004, Elisseeff et al. 2005). However, the stability of bagging global predictors such as decision trees seems not to be amenable to a theoretical analysis.

Some explanations for bagging's success are briefly recalled in Section 2. A short summary of stability analysis is then provided before the experimental results supporting the main point of this paper: bagging stabilizes decision trees. Section 4 describes the effects of bagging in a setup selected to illustrate some limits of previous explanations. Section 5 shows that stabilization also occurs in more typical cases.

2. How does bagging work?

Consider a learning set $\mathcal{L} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where each example is described by a pattern $\mathbf{x}_i \in \mathcal{X}$ and by a class label $y_i \in \Omega = \{\omega_1, \ldots, \omega_K\}$. A learning algorithm $\mathcal{A}$ is a function mapping from $\mathcal{L}$ to a decision rule function $\widehat{f}$, which maps $\mathcal{X}$ to Ω: $\mathcal{A}(\mathcal{L}) = \widehat{f}$ and $\widehat{f}(\mathbf{x}) \in \Omega$.

A bootstrap sample $\mathcal{L}_b = \{(\mathbf{x}_i^b, y_i^b)\}_{i=1}^n$ is obtained by independent draws with replacement from $\mathcal{L}$. It produces $\widehat{f}_b = \mathcal{A}(\mathcal{L}_b)$, and the bagged estimate $\widehat{f}_{\text{bag}}$ is defined as the majority vote among the B bootstrap predictors

$$\widehat{f}_{\text{bag}}(\mathbf{x}) = \underset{y \in \Omega}{\text{Argmax}} \sum_{b=1}^{B} \mathbf{I}_{\{y = \widehat{f}_b(\mathbf{x})\}} \quad , \tag{1}$$

where $\mathbf{I}_A$ denotes the indicator of the set A.

Breiman (1996a) presents bagging as a variance reduction procedure *mimicking* averaging over several learning sets. His argument would be definitive if averaging was performed on different learning sets, but it acts on bootstrap replicates of a single learning set. The work of Buja and Stuetzle (2000) on U-statistics details counter-examples for which bagging is proved to asymptotically increase squared bias and variance. Thus, although experimental results often show the expected variance reduction (Bauer and Kohavi 1999, Breiman 1996b, Schapire et al. 1998), several other stances have been explored to explain the success of bagging. A few selected pointers to the literature are listed below, a more comprehensive state of the art on the subject is provided in Grandvalet (2004).

The theoretical analyses of Friedman and Hall (2000) and Bühlmann and Yu (2002) prove that bagging asymptotically stabilizes the estimation process, leading thus to a reduction in variance. This effect clearly occurs also for finite samples, but it might not be the major reason for bagging's success. In particular, these analyses are based on expansions whose validity is limited to smooth estimation processes, while bagging is mostly effective for unstable predictors (Breiman 1996a, Dietterich 2000, Elisseeff et al. 2005).

In the framework of regression, Grandvalet (2004) relates bagging's success to the equalization of the potential influence of examples. Modifications of global characteristics of the estimator follow, stemming mainly from the down-weighting of leverage points, *i.e.* examples that may have a high influence on the predictor. Grandvalet (2004) concludes that bagging is beneficial when influential points are outliers and that it is detrimental otherwise. The present paper adopts the stability viewpoint in order to extend this conclusion to the classification framework.

In regression, under suitable smoothness conditions, the influence of examples can be measured by $\partial\widehat{f}(\mathbf{x}_i)/\partial y_i$. In classification such a measure is not relevant, since the response variable is categorical. Even if the response y_i were mapped to some numerical attribute, $\Delta\widehat{f}(\mathbf{x}_i)/\Delta y_i$ would not measure the actual influence, which is likely to depend on the actual class label (most classifiers are not linear predictors).

In classification, the influence of example i can be measured by the difference between the predictor $\widehat{f}$ fitted to $\mathcal{L}$ and the jackknife predictor $\widehat{f}^{-i} = \mathcal{A}(\mathcal{L}^{-i})$ fitted to $\mathcal{L}^{-i} = \{\mathbf{x}_j, y_j\}_{j\neq i}$, the learning sample deprived of example i. The difference between decision rules may then be evaluated either by a local measure, such as the disagreement in decisions at $\mathbf{x}_i$,

$$\delta(\widehat{f}, \widehat{f}^{-i}) = \mathbf{I}_{\{\widehat{f}(\mathbf{x}_i)\neq\widehat{f}^{-i}(\mathbf{x}_i)\}} \ , \tag{2}$$

or by a global measure, such as the expected disagreement in decisions,

$$\Delta(\widehat{f}, \widehat{f}^{-i}) = \mathbb{P}_X(\widehat{f}(X) \neq \widehat{f}^{-i}(X)) \ . \tag{3}$$

The latter is closely related to the notion of stability described in the following section.

3. Stability and generalization

Besides the bias-plus-variance argument, Breiman (1996a) also 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.

The notion of stability has been recently formalized in a series of papers. Bousquet and Elisseeff (2002) show how the stability of a learning algorithm with respect to changes in the learning set may be related to the generalization error of learning algorithms. Based on this work, Poggio et al. (2004) propose a definition of stability, from which necessary and sufficient conditions for consistency can be derived for a wide range of learning algorithms. They conclude that the stability viewpoint subsumes empirical risk minimization (see *e.g.* Vapnik 1998), enabling to characterize the generalization ability of learning algorithm which do not explore the entire hypothesis space. Additional works on stability, more specifically targeted towards the study of bagging have been developed by Evgeniou et al. (2004) and Elisseeff et al. (2005).

The different notions of stability allow to prove non-asymptotic bounds on the difference between the empirical risk R_{emp} and the risk R or the leave-one-out risk R_{loo} and the risk R, where R, R_{emp} and R_{loo} are respectively defined as

$$\begin{aligned} R(\widehat{f}) &= \mathbb{P}_{XY}(\widehat{f}(X) \neq Y) \\ R_{\text{emp}}(\widehat{f}) &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \mathbf{I}_{\{\widehat{f}(\mathbf{x}_i)\neq y_i\}} \\ R_{\text{loo}}(\widehat{f}) &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \mathbf{I}_{\{\widehat{f}^{-i}(\mathbf{x}_i)\neq y_i\}} \ . \end{aligned} \tag{4}$$

For classification, the two main notions of stability defined by Bousquet and Elisseeff (2002) are hypothesis stability and pointwise hypothesis stability. The latter is a slight modification of the former allowing to derive upper bounds on $R(\widehat{f}) - R_{\text{emp}}(\widehat{f})$ instead of

$R(\widehat{f}) - R_{\text{loo}}(\widehat{f})$. It is omitted here for brevity. Hypothesis stability is defined as: algorithm $\mathcal{A}$ has hypothesis stability β_n with respect to the $\{0,1\}$-loss if the following holds:

$$\forall i \in \{1,\ldots,n\}, \mathbb{E}_{\mathcal{L},X,Y}\left[|\mathbf{I}_{\{\widehat{f}(X)\neq Y\}} - \mathbf{I}_{\{\widehat{f}^{-i}(X)\neq Y\}}|\right] \leq \beta_n \ . \tag{5}$$

For binary classification, this definition can be rewritten

$$\forall i \in \{1,\ldots,n\}, \mathbb{E}_{\mathcal{L}}\left[\mathbb{P}_X(\widehat{f}(X) \neq \widehat{f}^{-i}(X))\right] \leq \beta_n \ ,$$

that is,

$$\forall i \in \{1,\ldots,n\}, \mathbb{E}_{\mathcal{L}}\left[\Delta(\widehat{f}, \widehat{f}^{-i})\right] \leq \beta_n \ . \tag{6}$$

For the multicategory classification case, the left-hand-side of (6) is an upper bound of the left-hand-side of (5), so that controlling the expected global influence of learning examples by (6) implies hypothesis stability (5).

In the present context, hypothesis stability ensures that, for any algorithm $\mathcal{A}$ with hypothesis stability β_n (5), the following holds

$$R(\widehat{f}) \leq R_{\text{loo}}(\widehat{f}) + \sqrt{\frac{1+6n\beta_n}{2n\delta}} \ , \tag{7}$$

with probability $1-\delta$ over the random draw of the training set $\mathcal{L}$ (theorem 11 of Bousquet and Elisseeff 2002).

Classification algorithms such as CART and C4.5 are the most prominent in the experimental studies reporting the effectiveness of bagging. Elisseeff et al. (2005) show that randomized algorithm can be handled in the framework of stability. However, only local algorithm such as nearest-neighbor seem to be amenable to a theoretical analysis. The following sections provide experimental evidence showing that bagging, applied to decision trees, equalizes the influence of examples, and thus improves stability.

4. An atypical example

We begin by an example where bagging increases variance, to show that stability equalization is systematic, also occuring in this atypical case. According to the bias-plus-variance argument, bagging fails if variance reduction does not compensate for the bias introduced by using the empirical distribution for the true distribution. Failures of bagging in classification have been extremely rarely reported. An example is however provided by Schapire et al. (1998), who report that stumps (single node trees) achieve a test error rate of 40.6%, compared to 41.4% for bagged stumps in the ringnorm benchmark.

4.1. Experimental setup

The ringnorm benchmark, originally proposed by Breiman (1998), is a difficult task for decision trees, as the Bayes decision boundary is elliptical. There are 20 features and 2 classes. Class 1 is multivariate normal with mean zero and covariance matrix four times the identity. Class 2 is multivariate normal with mean $(a\ a\ a \ldots a)$, $a = 1/\sqrt{20}$ and covariance matrix identity. The Bayes error rate is 1.5%.

Here, as in Breiman's original proposal, the learning sets comprise 300 examples and bagging is computed with 25 bootstrap replications. The prediction error is estimated on test sets of size 10000 and 250 trials are performed.

According to our simulations, the mean prediction error of the bagged estimate is identical to the one of the original estimate (40.4%, with an insignificant difference of 0.05% between the two means). Significant differences between the two estimates are however observed in 35% of experiments (at the 5% level), with a standard error of differences in prediction error of 1.6%. The results are more variable with the bagged predictor, with a standard deviation of prediction error of 1.4% versus 0.7% for unbagged stumps.

4.2. Bias and variance

There is no general agreement on what the bias/variance decomposition should be for classification problems (see James 2003, for a review). For all definitions retaining an additive decomposition where variance represents the variable part of prediction error (including the one of Breiman 1996b), bagging reduces bias and increases variance. We do not detail these results here, since they follow directly from the observation that bagging increases the variance of prediction error without affecting its expectation.

4.3. Influence equalization

The cumulative distribution of $\mathbb{P}_X(\widehat{f}(X) \neq \widehat{f}^{-i}(X))$ is displayed in Figure 1 for stumps and bagged stumps. Histograms do not provide a valuable visualization tool, since the distribution for stumps has an important step at zero and is furthermore heavy-tailed.

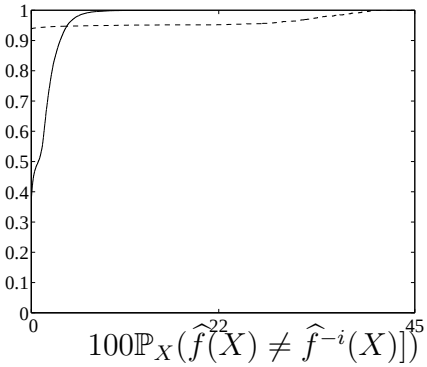

Figure 1: *Cumulative distribution function of* $\mathbb{P}_X(\widehat{f}(X) \neq \widehat{f}^{-i}(X))$ *(in %) for stumps (dashed) and bagged stumps (solid)*

For stumps, more than 93% of examples have no influence on the predictor, in the sense that the classifier output is unchanged when one of them is deleted. The absence of the other examples can cause important changes, and jackknife predictors disagree with the original predictor in up to 45% of test cases.[(1)] Note that reaching such a high influence requires a thorough change of the partition which can only be obtained by modifying

[(1)] These important differences are not reflected by the prediction error of $\widehat{f}^{-i}$: on this benchmark, many different predictors have an error rate of about 40%, due to the symmetry around axis $(1\ 1\ 1 \ldots 1)$ in the definition of classes.

the variable where the split is defined. Overall, the mean disagreement is 1.7%, with a standard deviation of 7.3%.

For bagged stumps, only 37% of examples have no influence on the predictor, and jackknife predictors disagree with the original predictor in at most 16% of test cases. The mean disagreement is slightly reduced to 1.3%, and its standard deviation is only 1.6%: bagging equalizes influence. On the one hand, increasing the proportion of influential points, by raising the number of "active" examples, smoothes the estimation process; on the other hand, decreasing the maximum influence of examples increases stability.

5. More typical situations

Section 4 illustrated a situation where the validity of the "reduce variance by averaging" fails, but bagging often reduces variance. This Section illustrates that influence equalization is also performed in these more typical situations.

5.1. Experimental setup

The computation of influence measures is quite demanding: for the bagged estimate it requires $B \times n$ retraining of the original rule. I therefore used fast-trained decision trees: stumps and CART with best-first induction (without pruning). For the latter, the tree growth is limited by forbidding splits when nodes gather less than 10 learning examples, or when their impurity (measured by the Gini index of Breiman et al. 1984) is below 10^{-2}. We thus cover a wide range of settings, from underfitting to overfitting.

A faithful approximation of the expected disagreement $\mathbb{P}_X(\widehat{f}(X) \neq \widehat{f}^{-i}(X))$ requires a large test set, while a steady estimation of influence distribution requires either several learning sets or a large one. As bagging is mostly efficient for small to medium-sized datasets Dietterich (2000), I chose mainly artificial benchmarks, for which several independent learning samples and large test sets can be generated.

Besides ringnorm, the two other artificial benchmarks proposed by Breiman (1998) were selected. Results are averaged on 100 independent learning samples of size 300. Twonorm is a binary classification problem in $\mathbb{R}^{20}$. Each class is drawn from a normal distribution. Class 1 has mean zero and unit covariance. Class 2 has mean $(a\ a\ a \ldots a)$, $a = 2/\sqrt{20}$ and unit covariance matrix. The Bayes error rate is 2.3%. Threenorm is also a binary classification problem in $\mathbb{R}^{20}$. Class 1 is drawn with equal probability from a unit multivariate normal with mean $(a\ a\ a \ldots a)$, and from a unit multivariate normal with mean $(-a\ -a\ -a \ldots -a)$. Class 2 is drawn from a unit multivariate normal with mean $(a\ -a\ a \ldots -a)$, $a = 2/\sqrt{20}$. The Bayes error rate is 10.5%. I also selected a real dataset from the UCI repository: satimage, since its number of learning and test examples (respectively 4435 and 2000) follows to the requirements. It is a six classes classification task in $\mathbb{R}^{36}$.

5.2. Results

The results for stumps in satimage are not reported here, since bagging does not modify the solution (the number of degrees of freedom of the classifier is overwhelmed by the number of examples). As shown in Table 1, in all the remaining experiments, bagging reduces test error.

Table 1: *Expected error rates (in %), for stumps, trees, bagged stumps and bagged trees; Opt. stands for the optimism of the observed error* $\widehat{\mathbb{E}}_{\mathcal{L}}[R_{\text{emp}}(\widehat{f}) - R(\widehat{f})]$ *the stability* $\widehat{\beta}_n$ *is estimated by the mean expected disagreement (3), and* $1 - \delta_0$ *is the confidence such that* $\sqrt{\frac{1+6n\widehat{\beta}_n}{2n\delta_0}} = 1$.

		$\widehat{\mathbb{E}}_{\mathcal{L}}[R(\widehat{f})]$	$\widehat{\mathbb{E}}_{\mathcal{L}}[R_{\text{emp}}(\widehat{f})]$	Opt.	$\widehat{\beta}_n$	$1 - \delta_0$
Ringnorm	stump	40.4	34.8	5.6	1.7	0.950
	bagged stump	40.5	35.3	5.2	1.3	0.961
	tree	22.4	5.1	17.4	4.1	0.876
	bagged tree	12.8	0.9	11.8	2.2	0.934
Twonorm	stump	33.0	26.5	6.5	3.3	0.902
	bagged stump	21.0	14.4	6.6	2.4	0.929
	tree	22.6	4.2	18.4	4.5	0.865
	bagged tree	8.9	0.4	8.6	1.9	0.942
Threenorm	stump	41.7	35.0	6.7	3.6	0.892
	bagged stump	34.4	25.8	8.6	3.4	0.899
	tree	32.6	6.1	26.5	6.8	0.796
	bagged tree	21.0	0.8	20.2	4.2	0.873
Satimage	tree	14.3	5.3	9.0	0.1	0.998
	bagged tree	11.5	3.6	7.9	0.1	0.997

The expected difference in error rates between test and train ($\widehat{\mathbb{E}}_{\mathcal{L}}[R_{\text{emp}}(\widehat{f}) - R(\widehat{f})]$) is not always lowered by bagging, but the estimated stability $\widehat{\beta}_n$ is systematically improved. This stability is always considerably lowered by bagging for trees, approaching the stability of unbagged stumps. Our experiments finally support that the bound provided by the stability analysis (7) may be tight. Indeed, it is not uncommon to obtain bounds which are far from providing any practical guarantees (with error rates much greater than 1). Here, we observe that the levels of confidence corresponding to a 100% optimism of the leave-one-out risk are quite high. [(2)] Unfortunately, our estimation of the stability β_n is computed by the mean of expected disagreement, a quantity which is very expensive to compute, and which furthermore necessitate numerous test data. Thus, we should not expect to be able to estimate β_n in real-life applications.

Figure 2 details how stability is improved, by displaying the cumulative distributions of influence statistics for unbagged and bagged predictors. We see that, for all unbagged predictors, a large proportion of examples has no influence on prediction, while a few examples have a high impact. Bagging reduces drastically the proportion of inactive example and the maximum influence. These two phenomena result in a systematic decrease of the span of influence statistics. Note that, for trees as for stumps, the very high influences are caused by thorough modifications of the partition. These changes systematically result from instabilities in the choice of the splitted variable at the root

[(2)] Of course, this optimism is trivially guaranteed, but the figures in Table 1 suggest that the order of magnitude of the bound (7) is correct, even for small sample sizes.

node.

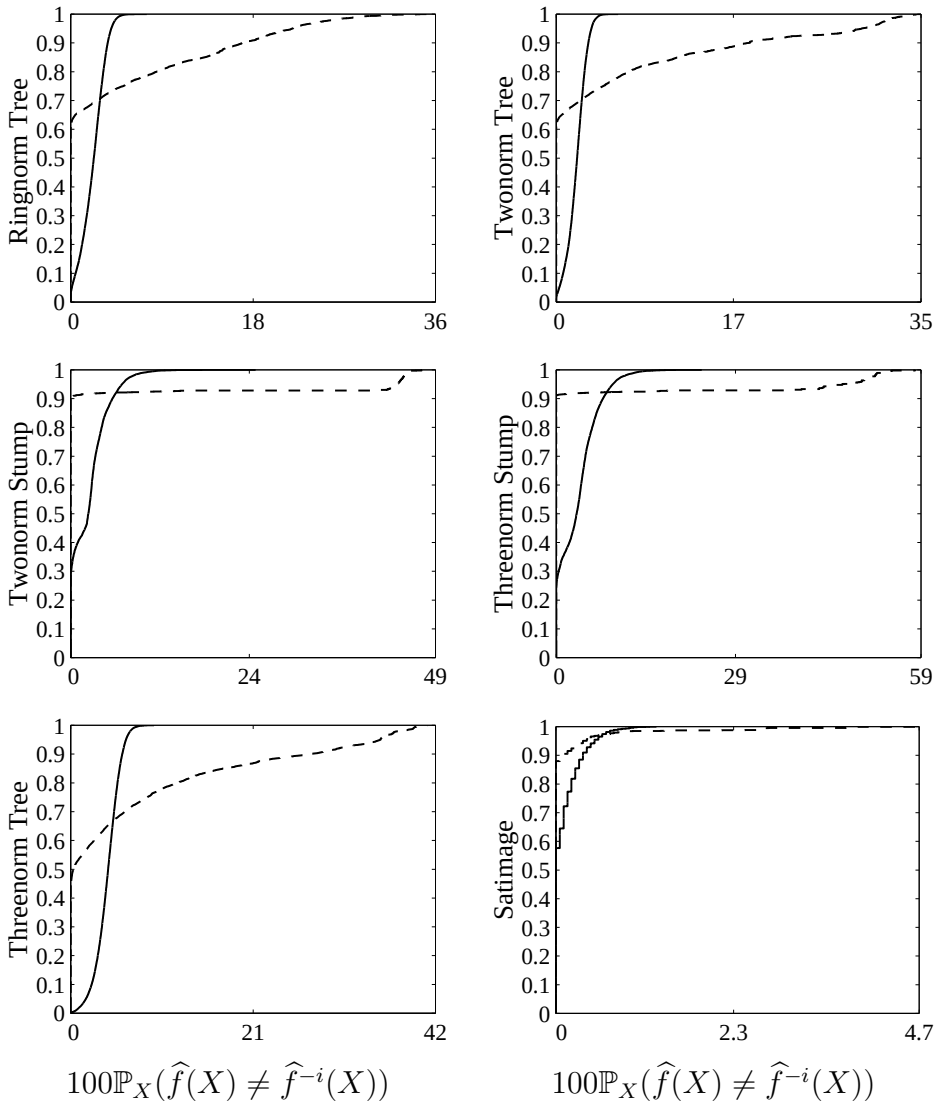

Figure 2: *Empirical cumulative distribution function of* $\mathbb{P}_X(\widehat{f}(X) \neq \widehat{f}^{-i}(X))$ *(in %) for unbagged predictors (dashed) and bagged predictors (solid)*

Again, these graphs suggest that two phenomena occur. Bagging smoothes the estimation process by reducing the proportion of "inactive" examples, while it decreases the maximum influence of examples, providing important stability improvements, which *usually* result in less variable predictors.

6. Discussion

Our experiments with stumps and trees illustrated that bagging equalizes the influence of examples on these predictor: more patterns have an influence on the estimation process, and the highly influential examples are down-weighted. These results corroborate the earlier observations in point estimation and regression reported by Grandvalet (2004).

We conjecture that the up-weighting of points with little influence may be responsible for the smoothing effect which was analyzed asymptotically (Friedman and Hall 2000, Bühlmann and Yu 2002). The example of indicator functions examined by Bühlmann and Yu (2002) is typical of this up-weighting, due to near-boundary examples entering the estimation of split location. On the other hand, the theory developed in these studies cannot explain the down-weighting of highly influential examples, because the latter cause discontinuities in the estimation process, stemming from modified split directions at the root node of a tree.

We present an experiment showing that bagging may increase the variance of decision trees. Breiman (1996a) notifies limits on the scope of his bias-plus-variance argument, which have been reached in asymptotical studies in point estimation, proving that bagging could increase variance (Buja and Stuetzle 2000). We hope that the present counter-example, which was preceded by other ones already exhibited in point estimation and regression (Grandvalet 2004), will contribute to amend the folk belief that bagging simply reduces variance in its averaging process.

Breiman (1996a) also stated that the vital element for gaining accuracy thanks to bagging is the instability of the prediction method. Hypothesis stability, which is quantified by the expected influence of examples, was shown to be systematically improved for stumps and for the deeper decision trees. Quantifying hypothesis stability supplies bounds on the generalization ability of learning algorithms (Bousquet and Elisseeff 2002). The estimation of stability is however extremely demanding, which makes it an unlikely practical means to control the generalization ability of learning algorithms such as trees. Nevertheless, it is worth stating that our experiments suggest that these bounds are quite tight for unbagged and bagged on decision trees.

References

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

Bousquet O. Elisseeff A. (2002) Stability and generalization. *Journal of Machine Learning Research*, 2:499–526.

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

Breiman L., Friedman J. H., Olshen R., Stone C. J. (1984) *Classification and Regression Trees*. Wadsworth, Belmont.

Bühlmann P., Yu B. (2002) Analyzing bagging. *The Annals of Statistics*, 30:927–961.

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

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.

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

Efron B., Tibshirani R. J. (1993) *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman & Hall.

Elisseeff A., Evgeniou T., and Pontil M. (2005) Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6:55–79.

Evgeniou T., Pontil M., Elisseeff A. (2004) Leave one out error, stability, and generalization of voting combinations of classifiers. *Machine Learning*, 55(1):71–97.

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

Grandvalet, Y. (2004) Bagging equalizes influence. *Machine Learning*, 55(3):251–270.

James G. M. (2003) Variance and bias for general loss functions. *Machine Learning*, 51 (2):115–135.

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

Poggio T., Rifkin R., Mukherjee S., Niyogi P. (2004) General conditions for predictivity in learning theory. *Nature*, 428:419–422.

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

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.

Vapnik V. N. (1998) *Statistical Learning Theory*. Adaptive and learning systems for signal processing, communications, and control. Wiley, New York.