Out-of-bag Estimation of the Optimal Sample Size in Bagging

Gonzalo Martínez-Muñoz[*,a], Alberto Suárez[a]

[a] *C/Francisco Tomás y Valiente, 11*
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Madrid (28049), Spain

Abstract

The performance of m-out-of-n bagging with and without replacement in terms of the sampling ratio (m/n) is analyzed. Standard bagging uses resampling with replacement to generate bootstrap samples of equal size as the original training set $m_{wor} = n$. Without-replacement methods typically use half samples $m_{wr} = n/2$. These choices of sampling sizes are arbitrary and need not be optimal in terms of the classification performance of the ensemble. We propose to use the out-of-bag estimates of the generalization accuracy to select a near-optimal value for the sampling ratio. Ensembles of classifiers trained on independent samples whose size is such that the out-of-bag error of the ensemble is as low as possible generally improve the performance of standard bagging and can be efficiently built.

Key words: Bagging, subagging, Bootstrap sampling, subsampling, Optimal sampling ratio, Ensembles of Classifiers, Decision Trees

1. Introduction

Empirical studies have established that bagging is a simple and robust method that generally increases the accuracy of a single learner [1, 2, 3, 4, 5, 6, 7]. In standard bagging individual classifiers are trained on independent bootstrap samples that are extracted with replacement from the set of labeled examples available for learning [8]. The size of these samples is generally chosen to coincide with the number of examples in the original training dataset. This prescription is arbitrary and need not be optimal in terms of the generalization accuracy of the ensemble. In this work we carry out an investigation on the dependence of the ensemble performance on the sampling ratios used. The results

[*] Corresponding author
Email addresses: gonzalo.martinez@uam.es (Gonzalo Martínez-Muñoz), aberto.suarez@uam.es (Alberto Suárez)
URL: http://www.eps.uam.es/ gonzalo/ (Gonzalo Martínez-Muñoz)

obtained show that ensembles of decision trees trained on bootstrap samples whose sizes are different from the usual choice can match the performance and sometimes even outperform standard bagging. In most of the classification tasks investigated, the ensembles with the lowest test errors use bootstrap samples that are smaller than the original training data. In fact, combining classifiers that are trained on bootstrap samples (with replacement) containing either 60% or 80% of the original training data exhibit a good overall performance.

A less explored variant of bagging is m-out-of-n bagging without replacement, also known as subagging [9, 10, 11]. Without-replacement (subsampling) methods were used early on to provide estimates of statistical quantities by sampling (see for instance [12] and references therein). A common choice for the sample size in subagging is $m_{wor} = n/2$ [11]. In fact, provided that high-order terms can be neglected, half-sampling without replacement is expected to behave similarly as $m = n$ sampling with replacement [10, 11]. The similarities between the behaviors of bagging and subagging can be viewed as a particular case of a more general correspondence between the statistical properties of with-replacement samples of size m_{wr} and of subsamples of size m_{wor}, obtained without replacement, when

$$\frac{m_{wr}}{n} = \frac{\frac{m_{wor}}{n}}{1 - \frac{m_{wor}}{n}} \quad . \tag{1}$$

This relation has been rigorously derived for U-statistics in [10]. This work also provides empirical support for the validity of this equivalence in bagging ensembles of regression trees. The special case $m_{wr} = n$ and $m_{wor} = n/2$ has also been empirically explored in [11] using ensembles of regression trees. In this work we extend these investigations and show that the correspondence also holds, at least in an approximate manner, in ensembles of decision trees for classification. Similarly to bagging, subagging ensembles trained on smaller subsamples, $m_{wor} < n/2$, often match or improve the performance ensembles built with half-subsampling.

In summary, the sizes of the optimal samples for m-out-of-n bagging both with and without replacement are generally lower than the usual choices of $m_{wr} = n$, when replacement is used, and $m_{wor} = n/2$, if examples are not replaced after having been extracted. However, the optimal sampling ratios can be very different in different problems. This indicates that the amount of resampling should not be fixed beforehand, independently of the classification task to be solved. To estimate the optimal sampling ratios in bagging and subagging, we propose to use out-of-bag estimates of the generalization error [13]. As expected, this procedure tends to select sample sizes that are smaller than the standard choices and that are close to being optimal.

The article is organized as follows: In Section 2 we present a review of previous investigations on the properties of with and without replacement sampling, specially regarding the induction of bagging ensembles from data. The effects of the sampling ratio on the properties of the bagging ensembles generated are discussed in Section 3. Section 4 presents the results of an empirical investigation

of the classification performance of bagging and subagging ensembles generated using different sampling ratios. Finally, the conclusions of this investigation are summarized in Section 5.

2. Previous Work

Subsampling and m-out-of-n bootstrapping with $m < n$ have been proposed in the statistical literature as alternatives to the standard bootstrap [14, 15, 16, 17, 18]. In general, the motivation of these studies is to repair inconsistencies of the bootstrap estimates obtained by sampling with replacement and $m = n$, or to improve the efficiency of these estimates. Extensions of the bootstrap along these lines have also been studied in the context of automatic induction of models from data [19, 9, 20, 10, 11, 21, 22]. One of the first variants of bagging that uses a sampling ratio different from the standard value is *Rvotes* [19]. This algorithm is designed to increase the training speed when very large databases are available for learning. In Rvotes the individual ensemble classifiers are built using very small bootstrap samples, as small as 0.5% of the original training data. As a consequence, the generalization performance of the proposed algorithm is rather poor in problems where the training data is not abundant and redundant in some way.

The properties of m-out-of-n bagging with and without replacement have also been analyzed in detail in [9, 20, 10, 11]. In [9, 20] subagging is proposed as a computationally less intensive variant of bagging that is expected to have similar accuracy in regression and classification problems. The focus of this work is to provide a rigorous framework to understand the variance reduction effect of bagging and subagging. Of particular interest for the current investigation is the observation made in [10, 11] on the equivalence of results between a bagging ensemble that uses samples of size m_{wr} obtained with replacement and an ensemble built using without-replacement samples of size m_{wor} such that relation (1) holds. Although no general proof is given, these investigations provide strong empirical evidence for this correspondence in the context of ensembles of regression trees. In [11] it is shown that the bias and variance of regression bagging ensembles can be reduced by decreasing the size of the bootstrap samples.

There are other studies where improvements in the generalization performance of bagging are obtained by using samples sizes different from the original training data [21, 22]. In particular, the performance of nearest neighbor can be improved with bagging only if the samples used contain on average less than 50% of different examples from the original training set [22]. This corresponds to sampling ratios of $m_{wr}/n < 0.69$ for with-replacement sampling and $m_{wor}/n < 1/2$ for subsampling. In fact, the performance of standard bagged nearest-neighbor, which uses samples of size $m_{wr} = n$, is equal to the performance of nearest-neighbor alone [1]. By contrast, if the sampling ratio tends to 0 as the size of the original training set tends to infinity, the performance of the ensemble of nearest neighbor classifiers can be shown to approach the Bayes limit [22].

3. Influence of the Sampling Ratio in Bagging

In this section we provide a qualitative analysis of the dependence of the properties of bagging ensembles generated using different sampling ratios. This analysis is straightforward in subsampling, where a subset of m instances is extracted without replacement from the initial set of size n. Samples obtained with values of m close to n are very similar to the original training set. Hence, the members of the bagging ensemble built using these samples should be very similar to the single classifier built using the complete original training data. The differences between the subsample and the original sample become larger as m is lowered. There are exactly $\binom{m}{n}$ different subsamples of size m. The maximum number of different subsamples corresponds to a sampling ratio $m/n = 1/2$. This is one of the reasons why half-sampling has been commonly used in previous investigations [23, 24].

The analysis for with-replacement sampling is slightly more complicated. Assume that a sample of size m is extracted with replacement from a set composed of n examples. For sufficiently large values of n and m, the average number of different examples from the original set included in the bootstrap sample is approximately $1 - \exp\{-m/n\}$. Figure 3 shows the evolution of this average as the sampling ratio $r = m/n$ increases. In the standard version of bagging, these bootstrap samples contain the same number of examples as the original data $m/n = 1$. This means that, on average, a bootstrap sample of size $m = n$ contains $\approx 63.2\%$ different training instances. The remaining $\approx 27.8\%$ are repeated examples. Another interesting point in the curve depicted in Figure 3 corresponds to bootstrap samples that contain on average one half of the different instances in the original training data. This corresponds to samples whose size is $\approx 69.3\%$ of the parent set size. For sampling ratios below this threshold, on average, every instance in the original training set is used for induction in less than half of the classifiers in the ensemble. This means that, for a given training instance, the class assignment made by the ensemble takes into account not only its own class label, but also the class labels of nearby instances.

The influence of the bootstrap sample size on the classification performance of the corresponding bagging ensembles can be analyzed in the limits of very small and very large sampling ratios. In with-replacement bagging, if the size of the bootstrap samples is large compared to the size of the original training data then the number of different instances in the bootstrap samples asymptotically approaches the totality of the original examples: All of the original training examples are selected, and, on average, they are selected the same number of times. Under these conditions, classifiers trained on independent bootstrap samples tend to be very similar to each other and their combination in an ensemble does not significantly modify the accuracy of the individual predictors. In the opposite extreme, the classifier trained on a bootstrap sample that contains just one example would assign the class label of that single example to the whole feature space. Consequently, the combination of the decisions by majority voting of a sufficiently large number of such classifiers would always output the

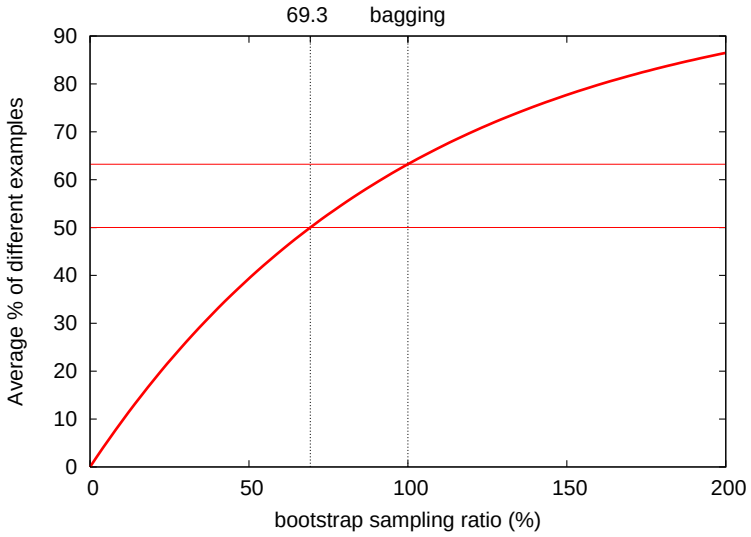

Figure 1: Average percentage of different selected examples with respect to the size of the bootstrap sample.

most common class in the training data, irrespective of the values of the feature vector of the instance that is being classified.

Typically the optimal generalization performance obtains at intermediate sampling ratios, neither too large nor too small. The usual prescription is to use bootstrap samples of the same size as the original training set. In this manner bagging generates a variety of classifiers whose combined decisions generally improve those of single learners and which are also better than the majority class prediction. However, there is no *a priori* reason to expect $m_{wr} = n$ to be the optimal sampling size in terms of the generalization performance of the ensemble. The region $m_{wr} > n$, is, in principle, less attractive than the region $m < n$: On the one hand, for larger samples, the time needed to build the ensemble increases. On the other hand, the different samples become more similar, which implies that the diversity of the classifiers becomes smaller. Hence, the potential for improvement of classification by combining the decisions of the classifiers in the ensemble is reduced.

The experiments presented in the following section show that ensembles that use sampling ratios different from the usual prescriptions, $m_{wr}/n = 1$ in with-replacement sampling and $m_{wor}/n = 1/2$ in without-replacement sampling can have better generalization performance than standard bagging or half-subagging. However, the optimal sampling ratios can be quite different for different problems. We propose to use out-of-bag estimates of the generalization error [13] to determine the sampling ratio that is appropriate for each problem. To estimate the out-of-bag error, the class label assigned by the ensemble to a particular example in the original training data is computed using only the

predictions of classifiers trained on bootstrap samples that did not include that particular instance. The out-of-bag error is then calculated as the fraction of examples in the original training set that are incorrectly labeled using these predictions.

4. Experiments

Extensive experiments on 30 datasets from the UCI repository [25] and synthetic classification problems [26, 27] are carried out to analyze the performance of m-out-of-n with and without-replacement bagging as a function of m, the size of the samples used to train the individual ensemble classifiers. The characteristics of the different datasets used are shown in Table 1. This table includes information on the number of labeled instances available, the protocol used for testing and, in the last two columns, the numbers of attributes and classes, respectively. Given the correspondence between subsampling and the m-out-of-n bootstrap, the experiments are carried out in parallel, using with and without-replacement samples whose sizes are in the relation given by Eq. (1).

The estimates for the classification errors and sampling ratios reported in this article correspond to averages over different realizations of the classification problems. In the synthetic problems, *Led24, Ringnorm, Threenorm, Twonorm* and *Waveform*, 100 versions of the training and testing datasets are generated independently at random. For the remaining problems 10×10-fold cross-validation is used. Each realization involves the following steps:

1. Obtain the train/test sets by random generation or by 10-fold-cv depending on the domain (see 3^{rd} column in Table 1).
2. Using only the examples in the training set generate a sequence of bagging ensembles composed of 200 unpruned CART trees [26] for the different sampling rates considered. Unpruned trees are used as base learners because bagging ensembles of fully developed CART trees generally outperform bagging ensembles of pruned CART trees [4, 5]. The experiments are performed using a range of sampling rates that covers the different regimes of interest with sufficient detail. In sampling with replacement, ensemble classifiers are built from bootstrap samples of sizes that range from 2% to 120% of the size of the original training set, in steps of 2%. This choice of the step size provides sufficient resolution to analyze the dependence of the ensemble performance with the size of the samples. For subagging, the corresponding rates for without-replacement sampling are used. This gives a total of 60 bagging and 60 subagging ensembles composed of classifiers trained on samples of different sizes.
   Standard bagging corresponds to ensembles built using bootstrap samples of the same size as the original training data. For the experiments without replacement, the corresponding sampling ratios, computed with Eq. (1), are used: $1/101, 2/102, \ldots, 100/200, \ldots 120/220$. Half-subagging corresponds to a sampling ratio of 1/2 (100/200 in the series above).

Dataset	Instances	Test	Attrib.	Classes
Audio	226	10-fold-cv	69	24
Australian	690	10-fold-cv	14	2
Balance	625	10-fold-cv	4	3
Breast W.	699	10-fold-cv	9	2
Diabetes	768	10-fold-cv	8	2
Echo	74	10-fold-cv	6	2
Ecoli	336	10-fold-cv	7	8
German	1000	10-fold-cv	20	2
Glass	214	10-fold-cv	9	6
Heart	270	10-fold-cv	13	2
Hepatitis	155	10-fold-cv	19	2
Horse-Colic	368	10-fold-cv	21	2
Ionosphere	351	10-fold-cv	34	2
Iris	150	10-fold-cv	4	3
Labor	57	10-fold-cv	16	2
Led24	200	5000 cases	24	10
Liver	345	10-fold-cv	6	2
New-thyroid	215	10-fold-cv	5	3
Ringnorm	300	5000 cases	20	2
Segment	2310	10-fold-cv	19	7
Sonar	208	10-fold-cv	60	2
Soybean	683	10-fold-cv	35	19
Threenorm	300	5000 cases	20	2
Tic-tac-toe	958	10-fold-cv	9	2
Twonorm	300	5000 cases	20	2
Vehicle	846	10-fold-cv	18	4
Votes	435	10-fold-cv	16	2
Vowel	990	10-fold-cv	10	11
Waveform	300	5000 cases	21	3
Wine	178	10-fold-cv	13	3

Table 1: Characteristics of the classification problems and testing method

3. Estimate the generalization error of the ensembles using the test set and the out-of-bag instances [13].

The curves that trace the dependence of the training, test and out-of-bag error estimates as a function of sampling ratio for m-out-of-n bagging with and without replacement are displayed in Figures 2 and 3, respectively, for a representative collection of classification problems: *Glass, Ionosphere, Labor*, and *Vowel*. As a reference, the average test error for standard bagging (sampling ratio = 100%) and half-subagging are marked with horizontal and vertical lines in the plots. A paired t-test is performed to determine whether the differences in test error with respect to standard bagging are statistically significant. Errors

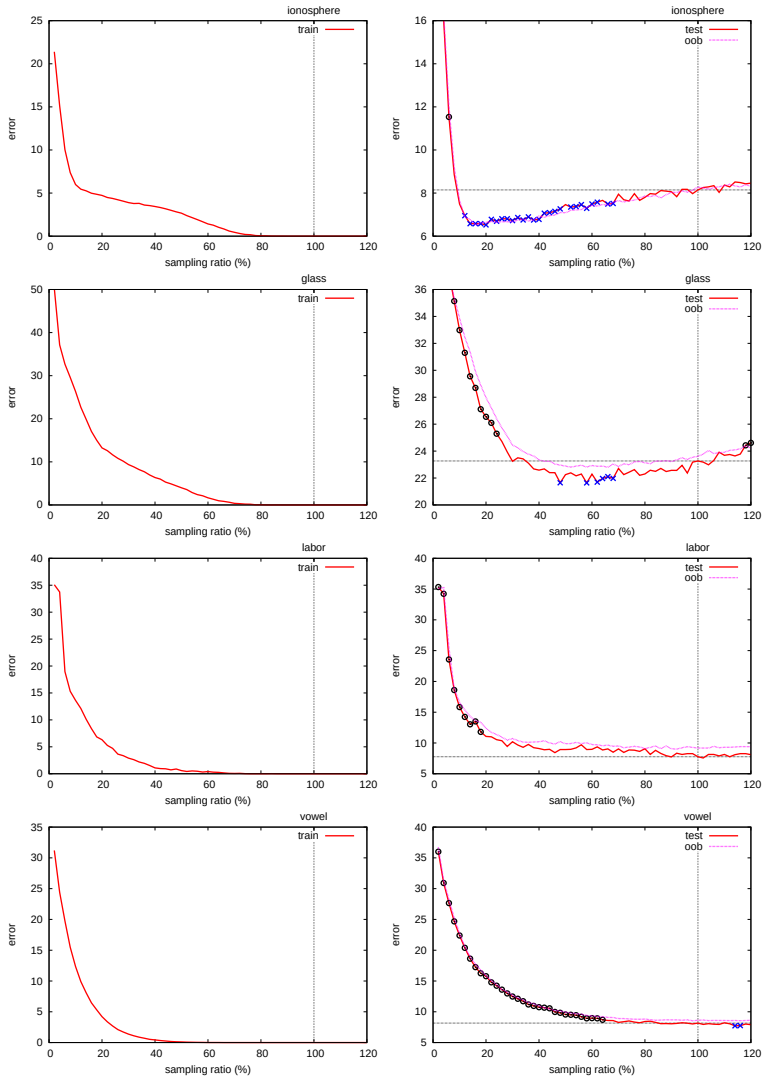

Figure 2: Training, test and out-of-bag error estimates for bagging ensembles (with-replacement sampling) as a function of the size of the bootstrap samples.

that are significantly lower than standard bagging at a significance level $\alpha = 0.01$ are marked with a cross ('x'). Similarly, errors that are significantly worse than bagging ($p - value < 0.01$), are marked with a circle ('o').

In the problems investigated, the test error curves follow one of four distinct patterns:

- In *Balance, Ecoli, German, Heart, Ionosphere, Liver* and in the synthetic datasets (*Led24, Ringnorm, Threenorm, Twonorm* and *Waveform*) the

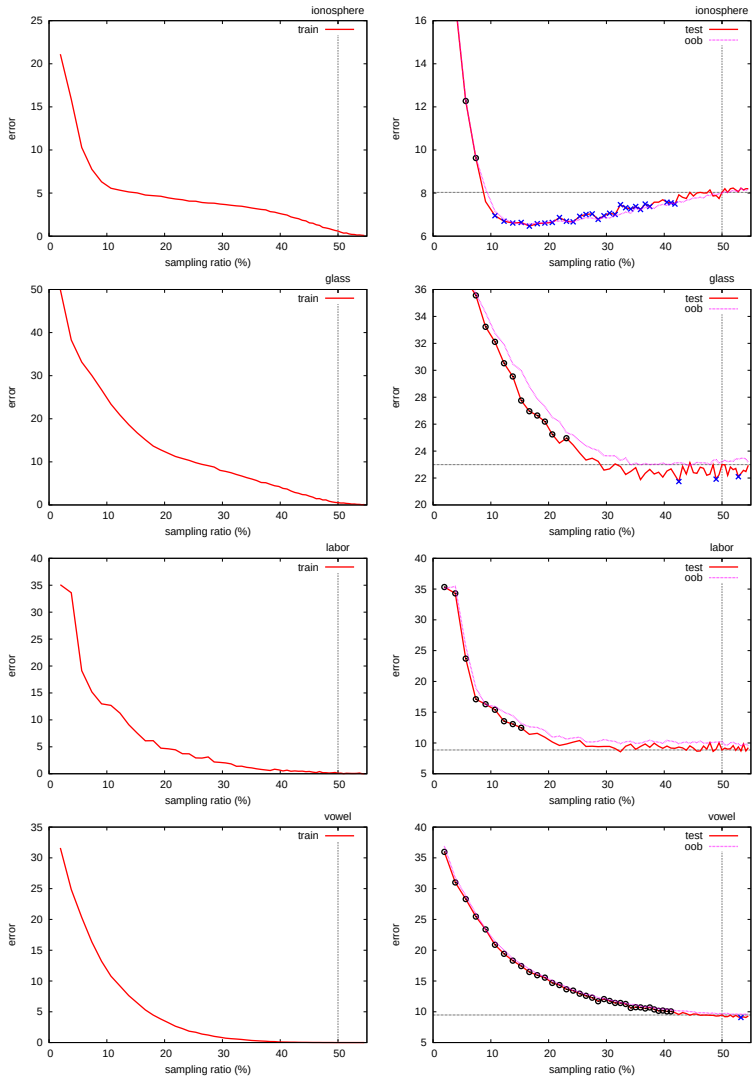

Figure 3: Training, test and out-of-bag error estimates for subagging ensembles (without-replacement sampling) as a function of the size of the samples.

error curves exhibit broad minima at intermediate sampling ratios. For these problems, it is easy to identify sampling ratios that significantly improve the performance of standard bagging.

- The error curves for *Australian, Diabetes, Echo, Glass, Hepatitis* and *Iris* are very irregular. There are several values of the sampling ratio, not necessarily close to each other, for which the corresponding ensembles sig-

Dataset	1/5	2/5	3/5	4/5	1	6/5
audio	29.3±8.6	24.6±8.1	22.1±8.0	20.4±7.8	19.8±7.3	19.6±7.9
australian	13.5±3.7	13.0±3.8	**12.7±3.6**	12.9±3.9	13.1±3.7	13.6±3.9
balance	**10.1±1.9**	**13.9±3.3**	**17.1±3.8**	**18.7±3.9**	19.2±3.9	19.8±3.9
breast	4.2±2.5	3.8±2.4	3.5±2.4	3.6±2.4	3.7±2.3	3.5±2.2
diabetes	24.0±3.8	24.0±4.0	24.1±3.8	24.4±4.0	24.5±4.1	24.9±4.0
echo	25.0±12.5	24.0±13.3	22.6±14.5	22.6±14.6	22.8±14.3	23.3±15.2
ecoli	14.9±5.3	**14.6±6.1**	**14.8±6.0**	15.4±6.2	15.7±6.1	16.7±6.0
german	**23.2±3.5**	23.4±3.5	**23.4±3.6**	23.7±3.5	24.1±3.9	24.0±3.9
glass	26.5±7.1	22.6±7.0	22.3±7.6	22.3±8.0	23.3±7.9	24.6±8.5
heart	**17.1±6.6**	**18.0±6.5**	**18.5±6.7**	**18.6±7.0**	19.6±7.0	20.1±6.5
hepatitis	16.9±7.3	17.0±7.7	16.5±7.5	17.3±7.9	17.4±7.9	18.1±8.0
horse-colic	17.0±5.3	16.0±5.6	15.4±5.5	14.9±5.4	15.0±5.4	15.4±5.3
ionosphere	**6.5±4.4**	**6.8±4.4**	**7.5±4.5**	7.8±4.7	8.1±5.1	8.5±4.8
iris	4.6±5.2	5.1±5.6	5.5±5.7	5.4±5.7	5.3±6.0	5.4±5.9
labor	11.1±13.2	9.1±11.1	9.0±10.5	9.1±10.7	7.8±9.9	8.1±10.2
led24	**28.4±1.6**	**29.8±1.7**	**31.4±1.8**	**32.8±1.9**	34.0±2.1	35.0±2.0
liver	**26.3±6.6**	**26.4±6.8**	**27.0±7.1**	**27.7±7.6**	29.0±7.7	29.6±8.0
new-thyroid	5.8±5.3	5.6±4.8	5.0±4.6	5.5±4.5	5.6±4.6	5.6±4.7
ringnorm	9.9±1.3	**8.7±1.7**	**9.0±2.0**	**9.3±1.9**	9.9±2.1	10.4±1.9
segment	3.5±1.2	2.7±1.1	2.4±0.9	2.3±0.9	2.3±0.9	2.3±0.9
sonar	22.8±9.6	22.2±9.0	20.7±8.3	20.4±8.0	20.7±8.8	19.8±7.8
soybean	7.7±2.7	6.4±2.4	6.8±2.4	6.5±2.4	6.7±2.5	6.5±2.5
threenorm	**18.4±1.4**	**18.2±1.4**	**18.6±1.6**	**18.9±1.6**	19.3±1.8	19.6±1.8
tic-tac-toe	1.9±1.6	1.3±1.2	1.2±1.1	1.2±1.1	1.2±1.1	1.3±1.1
twonorm	**4.9±0.8**	**5.3±1.0**	**5.8±1.2**	**6.3±1.4**	6.7±1.6	7.0±1.6
vehicle	25.5±3.7	24.9±3.8	25.0±3.7	24.8±3.9	24.7±4.0	24.9±3.9
votes	4.4±3.1	4.5±3.0	4.4±3.0	4.5±3.0	4.4±3.0	4.6±3.2
vowel	15.8±4.2	10.7±3.4	9.0±3.0	8.4±2.8	8.2±2.5	7.9±2.5
waveform	**17.8±1.2**	**18.3±1.2**	**18.7±1.3**	**19.1±1.3**	19.5±1.4	19.9±1.4
wine	3.0±3.9	2.4±3.7	2.6±3.9	2.7±4.0	2.9±4.1	3.2±4.2
w/d/l	9/12/9	10/16/4	12/16/2	8/22/0	-	0/22/8

Table 2: Average test error (in %) using different sample ratios in bagging (with replacement sampling).

nificantly outperform standard bagging. In these classification problems, using smaller bootstrap samples has the potential of improving the performance of bagging. However, the exact value of the optimal sampling ratio is difficult to estimate.

- For *Breast, Labor, New-thyroid, Soybean, Tic-tac-toe, Votes* and *Wine* the curves drop rather rapidly and quickly stabilize at the error of standard bagging. In this case, ensembles with medium to large sampling ratios have very similar classification performance. However, in terms of efficiency and

Dataset	1/6	2/7	3/8	4/9	1/2	6/11
audio	29.7±8.3	26.3±8.4	23.3±7.9	22.2±8.0	21.1±7.7	20.5±7.7
australian	13.6±3.9	13.2±3.7	**12.8±3.8**	13.0±3.9	13.2±3.8	13.1±3.8
balance	**10.3±2.0**	**12.8±2.9**	**15.9±3.6**	**17.1±3.5**	18.2±3.9	18.7±3.9
breast	4.3±2.5	4.0±2.5	3.7±2.5	3.6±2.5	3.6±2.3	3.6±2.4
diabetes	24.0±4.2	23.9±4.0	24.1±4.0	24.3±3.8	24.6±3.9	24.7±3.9
echo	25.8±14.1	24.3±12.7	23.3±14.0	21.6±14.7	22.4±15.1	22.4±14.8
ecoli	15.2±5.2	14.5±5.7	14.8±5.8	15.1±5.7	15.3±6.0	15.8±6.0
german	23.5±3.5	23.5±3.5	23.6±3.3	23.8±3.6	23.7±3.4	23.8±3.5
glass	27.0±7.2	23.2±7.6	22.6±8.0	23.1±7.6	23.0±7.6	22.9±8.2
heart	17.7±6.8	18.0±6.6	18.4±7.0	18.5±6.9	18.8±7.0	19.3±7.1
hepatitis	16.4±6.9	17.1±7.1	17.1±8.0	17.1±7.9	17.5±7.9	17.8±8.5
horse-colic	17.2±5.6	16.3±5.4	15.5±5.4	15.3±5.3	14.8±5.3	14.7±5.4
ionosphere	**6.5±4.0**	**6.8±4.3**	**7.4±4.6**	8.0±4.9	8.0±4.6	8.2±4.7
iris	4.8±5.3	4.9±5.6	4.9±5.7	5.1±5.7	5.2±5.8	5.2±5.6
labor	11.4±13.9	9.4±11.6	9.3±10.6	9.5±10.8	8.9±10.4	9.2±10.7
led24	**28.3±1.7**	**29.2±1.7**	**30.5±1.7**	**31.6±1.9**	32.4±1.9	33.2±1.9
liver	27.1±6.5	**26.1±7.3**	**26.5±7.1**	26.9±7.3	27.6±7.4	28.1±7.3
new-thyroid	6.1±5.4	5.7±5.2	5.8±4.8	5.3±4.5	5.5±4.6	5.7±4.5
ringnorm	10.3±1.5	**8.9±1.5**	**9.1±1.6**	**9.5±1.8**	10.0±1.9	10.3±2.0
segment	3.5±1.1	2.9±1.1	2.6±1.0	2.5±1.0	2.5±0.9	2.5±0.9
sonar	23.0±8.7	22.1±9.4	21.1±8.8	21.1±8.9	20.5±9.0	20.1±8.6
soybean	8.1±2.8	6.6±2.4	6.6±2.5	6.8±2.4	6.6±2.6	6.7±2.4
threenorm	**18.4±1.3**	**18.5±1.5**	**18.8±1.6**	**19.1±1.7**	19.5±1.8	19.8±1.9
tic-tac-toe	2.1±1.5	1.3±1.2	1.2±1.1	1.2±1.1	1.2±1.1	1.2±1.1
twonorm	**5.0±0.8**	**5.3±1.0**	**5.8±1.1**	**6.2±1.4**	6.7±1.6	7.0±1.7
vehicle	26.1±3.4	25.4±3.8	25.5±4.0	25.2±3.6	25.1±3.8	25.1±3.7
votes	4.4±3.1	4.5±3.1	4.4±3.0	4.5±3.0	4.5±3.0	4.5±3.2
vowel	16.5±4.4	11.7±3.6	10.7±3.3	9.4±3.0	9.5±2.9	9.3±2.9
waveform	**17.8±1.2**	**18.3±1.2**	**18.8±1.3**	**19.1±1.4**	19.5±1.4	19.8±1.5
wine	3.4±4.1	3.3±4.1	2.7±3.9	3.0±4.0	3.0±4.1	3.0±4.0
w/d/l	6/13/11	8/16/6	9/17/4	6/23/1	-	0/24/6

Table 3: Average test error (in %) using different sampling ratios in subagging (without replacement sampling).

speed of classification smaller sampling sizes should be preferred.

- The descent of the error curves in the problems *Audio, Horse-colic, Segment, Sonar, Vehicle* and *Vowel* is slow. The classification error is significantly worse than bagging for most of the smaller and intermediate sampling ratios. In these tasks the best performance is obtained for large sampling ratios.

A general conclusion that can be drawn from the analysis of these curves is that the standard choices $m_{wr} = n$ and $m_{wor} = n/2$ are in many classification

problems suboptimal. The error of ensembles built by training individual classifiers on very small bootstrap samples is generally very poor. In fact, as the sampling ratio tends to zero, the accuracy of the ensemble becomes close to the fraction of examples that belong to the majority class. This limiting behavior is confirmed in the datasets with the lowest number of instances, such as *Labor* and *Echo*. In particular, in *Labor* only one instance is selected for the sampling ratio of 2%. In this case, the trees generated are formed only by the root node. In consequence, the ensemble assigns the whole feature space to the majority class in the training set. Hence, the average accuracy of this ensemble is similar to the proportion of the instances of the majority class ($\approx 65\%$ for *Labor*). For medium sampling sizes the generalization capability of the ensemble increases. In many problems and for a large range of intermediate values for the sampling ratio, the ensembles generated outperform standard bagging. For larger sampling ratios (larger than 100% in with-replacement sampling, larger than 50% in without-replacement sampling), the test error curves generally present a slight upwards tendency. This pattern is more clearly marked in *Glass* and *Ionosphere*. In fact, if the with-replacement sampling ratio were sufficiently large, all of the original examples would be selected in every bootstrap sample. As a consequence, all the trees trained on these samples would actually be very similar.

Tables 2 and 3 display the average and standard deviation of the test errors for the different classification problems investigated and for different values of the sampling ratio. In all tables the estimates of the test error that are significantly different from bagging or subagging (using a t-test with $\alpha = 0.01$) are marked. They are highlighted in boldface if they improve the results of bagging and underlined if they are significantly worse than bagging. The last rows of these tables summarize the significant wins/draws/losses with respect to standard bagging and half-subagging using a t-test with a significance level $\alpha = 0.01$.

Table 2 shows that ensembles that use a fixed sampling ratio of 80% (column 4/5) never perform significantly worse than bagging in the classification problems investigated. This strategy is significantly better than standard bagging in eight of the classification tasks analyzed. Ensembles that use a sampling ratio of 120% (6/5), which is above the conventional choice, never perform significantly better than bagging in the problems investigated. Using sampling ratios under the 69.3% frontier—which means that, on average, less than half of the original training examples are selected in each bootstrap sample—increases the number of significant wins (twelve, ten and nine for 3/5, 2/5 and 1/5 respectively) but also has a larger number of significant losses with respect to bagging (two, five and nine, respectively). In some domains the accuracy improvements are remarkable. In *Heart* the error of standard bagging (19.6%) is reduced to 17.1% when each classifier in the ensemble is trained on bootstrap samples whose size is only 20% of the original data. For the problems *German, Ionosphere* and *Liver*, a 20% sampling ratio also significantly improves the performance. In *Ecoli* and in the synthetic domains a substantial improvement is observed for a wide range of sampling ratios, smaller than the conventional choice. A fairly

large increase in accuracy is also observed in the *Balance-scale* dataset: the error rate is reduced from 19.2% of standard bagging to 10.1% using a sampling ratio of 20%. However, the improvement in this domain is a side-effect of the difficulties that CART trees have to represent one particular class in this problem. The *Balance-scale* task consists in classifying each example as either having the balance scale tip to the right, tip to the left, or balanced. The attributes are the weight on the left arm, the left distance, the weight on the right arm and the corresponding distance. The classification rule consists in selecting the greater of the quantities (left-distance * left-weight) and (right-distance * right-weight). If they are equal, the scale is balanced. The architecture of decision trees is not suitable for learning the *balanced* class: from a training instance such as {1, 5, 1, 5} it is very difficult to mark examples such as {5, 1, 5, 1} or {3, 3, 3, 3} as *balanced*. In fact, in the experiments performed, none of the examples in the test set that belong to the *balanced* class are correctly classified by the ensembles, irrespective of the sampling ratio used. Similarly, none of the examples labeled by the ensembles as *balanced* are of this type. Because the examples of the balanced class are surrounded only by non-balanced instances the reduction of the sampling size in this problem has the effect of effectively erasing the balanced class from the possible outputs of the ensemble. The effect of not attempting to label any instance as *balanced* is that the instances belonging to the other two classes are classified more accurately.

Table 3 shows that the trends in subagging are similar to those in with-replacement bagging. In particular, the number of significant losses increases as the sampling ratio is reduced; the number of significant wins is largest at sampling rates 3/8 (which corresponds to 3/5 in with-replacement sampling) and the sampling ratio 6/11 never performs significantly better than subagging.

The trends in Figures 2 and 3 are very similar. These results illustrate that the statistical equivalence between sampling with and without replacement [9, 11] is also valid in bagging ensembles for classification. To analyze this correspondence in detail the training and test error curves are plotted in Figure 4 using comparable scales for the sizes of the samples obtained with and without replacement. For this purpose, instead of using m_{wor}/n in the abscissa, the without-replacement error rate is drawn as a function of the corresponding with-replacement sampling ratio given by Eq. (1). Using this transformation the curves for m-out-of-n bagging with and without-replacement appear almost superimposed. There are sizable differences for intermediate sampling ratios in the training error curves for the problems *Glass* and *Ionosphere*. To a lesser extent, this behavior is also present in the training error curves for *Labor* and *Vowel*. The origin of these discrepancies is the different proportion of distinct instances from the original training set that appear in the equivalent samples obtained with and without replacement. Given that unpruned trees are used, the training error approaches zero for sufficiently large sampling ratios and ensemble sizes. As the sampling ratio is lowered, this error becomes different from zero. The transition takes place around the point where the samples contain on average half of the different instances of the original complete training set: $m_{wr}/n = 69.3\%$ for with replacement sampling and $m_{wor}/n = 50\%$ (or, equiva-

lently, $m_{wr}(m_{wor})/n = 100\%$) for without-replacement sampling. As discussed in Section 3, when, on average, less than half different instances from the original training set are sampled, the class label assigned by the ensemble to a given training example also depends on the class labels of the surrounding training instances. Consequently, for sampling rates lower than the aforementioned values, the ensemble training error is typically larger than zero.

In all cases, the out-of-bag and the test error curves exhibit similar qualitative and quantitative behavior (see Figures 2 and 3). This means that the out-of-bootstrap error can be used to estimate sampling ratios that are near-optimal. To reduce the computational cost of determining the optimal sample size the out-of-bag error is estimated in only twelve of the 60 ensembles available. When sampling with replacement is used, the 12 different ensembles considered are those with sampling ratios between 1/10 and 12/10 in steps of 1/10. In subsampling, the out-of-bag selection is made from the equivalent ensembles, that is, those with sampling ratios: 1/11, 2/12, 3/13, 4/14, 5/15, 6/16, 7/17, 8/18, 9/19, 10/20 (half-subagging), 11/21 and 12/22. Note that a precise determination of the optimal sampling ratio is not possible because of the fluctuations in the out-of-bag estimates. In fact, using a finer grid for the search of the optimal sampling ratio does not lead to a significant improvement of the generalization performance. Table 4 shows the average generalization error for standard bagging, half-subagging and of ensembles built using out-of-bag estimates of the optimal sampling sizes. The average values of these sampling ratios are displayed in the fourth and seventh columns of this table. In most classification problems, the selected sampling ratios are smaller than the conventional choices. Since fewer training examples are used, the individual classifiers can be built faster. Furthermore, the trees generated with smaller surrogate training datasets also tend to be smaller, which implies that the corresponding ensembles need less storage and classify faster. Nonetheless these improvements are only moderate in the problems investigated.

Table 4 shows that the performance of ensembles built using with-replacement sampling in which the sample size is estimated by minimizing the out-of-bag error (*oob wr* ensembles) is significantly better than bagging in 9 problems and has equivalent accuracy in 21. For without-replacement sampling estimating the optimal sample size using out-of-bag instances significantly improves the classification accuracy of subagging in 7 classification problems and is equivalent in 23. In the problems investigated, using oob estimates of the sample size never leads to a significantly worse generalization performance than the standard choices. The out-of-bag estimates of the sampling ratios are near-optimal in terms of the classification accuracy of the corresponding ensembles. In *Audio, Segment* or *Vowel* datasets, where the best average result from Table 4 corresponds to standard bagging, out-of-bag selects bootstrap sizes around 100% for with-replacement and 50% for without-replacement sampling. In the other extreme, in datasets where the optimal sample ratios are around 20% (1/6 in without-replacement sampling), such as the synthetic problems, the out-of-bag estimates tend to select smaller sampling ratios, in the range 20-40%.

To determine whether these improvements are statistically significant, the

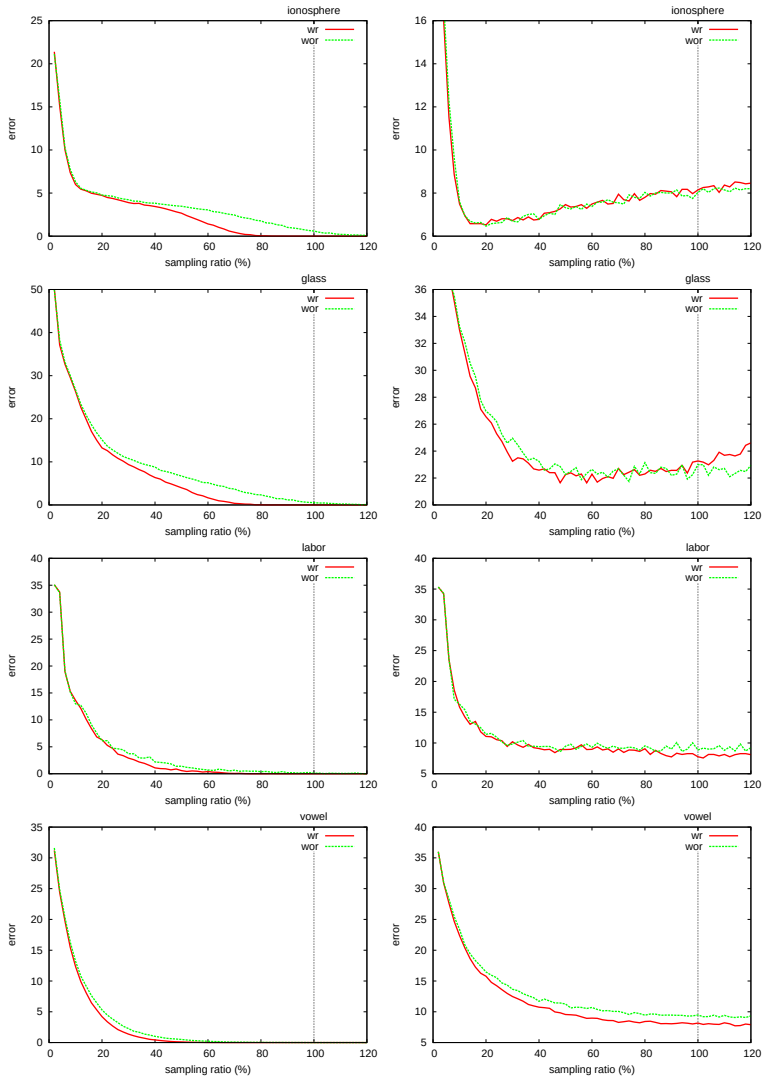

Figure 4: Comparison of training and test error for m-out-of-n with and without-replacement bagging as a function of the size of the samples. To make the comparison possible, in the curves for without-replacement sampling, the sample size m_{wor}, is scaled according to Eq. (1), and the corresponding m_{wr} is used.

performance of the different classification systems in the set of problems investigated are compared using the methodology proposed in [28]. For a given classification problem, the algorithms are ranked according to their performance on the test set. Then, a Friedman test with $\alpha = 0.05$ is applied to the average ranks to determine whether it is possible to reject the hypothesis that there

	with-replacement sampling			without-replacement sampling		
	bagging	oob wr		subag.	oob wor	
Dataset	error	error	m_{wr}^{oob}/n	error	error	m_{wor}^{oob}/n
audio	19.8±7.3	20.1±8.0	100.6±16.4	21.1±7.7	21.5±7.9	51.6±3.3
australian	13.1±3.7	13.0±3.9	65.9±22.8	13.2±3.8	13.2±3.7	41.5±9.5
balance	19.2±3.9	**9.8±1.9**	12.9±4.6	18.2±3.9	**9.9±1.8**	12.1±3.8
breast	3.7±2.3	3.7±2.3	79.3±24.7	3.6±2.3	3.7±2.3	45.4±7.8
diabetes	24.5±4.1	24.3±4.1	53.4±29.0	24.6±3.9	24.8±4.1	33.6±13.5
echo	22.8±14.3	22.2±14.2	75.5±23.5	22.4±15.1	22.6±14.2	41.8±11.0
ecoli	15.7±6.1	15.0±6.1	47.0±17.1	15.3±6.0	14.8±5.3	35.9±9.4
german	24.1±3.9	23.4±3.6	58.8±27.5	23.7±3.4	23.8±3.4	36.3±11.5
glass	23.3±7.9	22.9±7.9	68.7±22.0	23.0±7.6	22.8±7.7	44.6±7.3
heart	19.6±7.0	**17.8±6.9**	25.0±20.6	18.8±7.0	17.7±6.5	23.6±13.7
hepatitis	17.4±7.9	17.1±7.5	51.2±28.4	17.5±7.9	17.7±7.3	33.4±12.4
horse-colic	15.0±5.4	15.4±5.4	82.7±23.7	14.8±5.3	15.0±5.4	49.7±5.7
ionosphere	8.1±5.1	**7.1±4.0**	32.5±13.3	8.0±4.6	**6.6±4.1**	25.0±8.5
iris	5.3±6.0	5.4±5.5	70.8±40.7	5.2±5.8	5.3±5.4	36.4±15.3
labor	7.8±9.9	8.6±9.8	94.7±23.1	8.9±10.4	9.3±10.2	47.9±8.8
led24	34.0±2.1	**28.9±1.6**	24.6±12.2	32.4±1.9	**28.9±1.9**	19.7±9.3
liver	29.0±7.7	**26.8±7.1**	41.2±17.2	27.6±7.4	26.7±6.8	31.2±11.0
new-thyroid	5.6±4.6	5.5±4.5	64.3±23.8	5.5±4.6	5.7±4.6	41.4±9.0
ringnorm	9.9±2.1	**9.0±2.0**	64.7±30.5	10.0±1.9	**9.0±1.5**	35.8±9.2
segment	2.3±0.9	2.4±0.9	98.1±19.7	2.5±0.9	2.5±0.9	51.2±3.9
sonar	20.7±8.8	20.6±8.4	88.8±24.8	20.5±9.0	21.5±8.5	46.7±7.2
soybean	6.7±2.5	6.6±2.3	80.7±31.8	6.6±2.6	6.6±2.4	41.6±8.1
threenorm	19.3±1.8	**18.5±1.5**	45.9±26.1	19.5±1.8	**18.7±1.5**	30.8±11.2
tic-tac-toe	1.2±1.1	1.1±1.1	81.1±24.0	1.2±1.1	1.1±1.1	45.4±7.0
twonorm	6.7±1.6	**5.1±0.8**	33.2±21.7	6.7±1.6	**5.0±0.8**	20.2±9.0
vehicle	24.7±4.0	25.1±3.9	82.6±28.9	25.1±3.8	25.5±3.7	42.4±9.6
votes	4.4±3.0	4.7±3.0	65.2±31.6	4.5±3.0	4.7±3.0	37.2±13.2
vowel	8.2±2.5	8.1±2.5	100.6±15.9	9.5±2.9	9.4±3.0	50.7±3.6
waveform	19.5±1.4	**18.1±1.3**	34.9±26.8	19.5±1.4	**18.0±1.3**	21.1±11.3
wine	2.9±4.1	2.9±4.0	74.5±30.8	3.0±4.1	2.8±3.8	41.0±12.1
w/d/l		9/21/0			7/23/0	

Table 4: Average test error (in %) of ensembles that use samples whose sizes are determined using the out-of-bag error (oob) and the corresponding sampling ratios (also in %) for with and without-replacement sampling. The average test error for standard bagging and half-subaging are given for reference.

are no differences in performance among the different classification methods for the problems investigated. If this hypothesis is rejected, a Nemenyi test is applied to determine whether the differences in average ranks between pairs of algorithms are statistically significant at a 95% confidence level. Figure 5 displays the average ranks of standard bagging, subagging and of the ensembles in

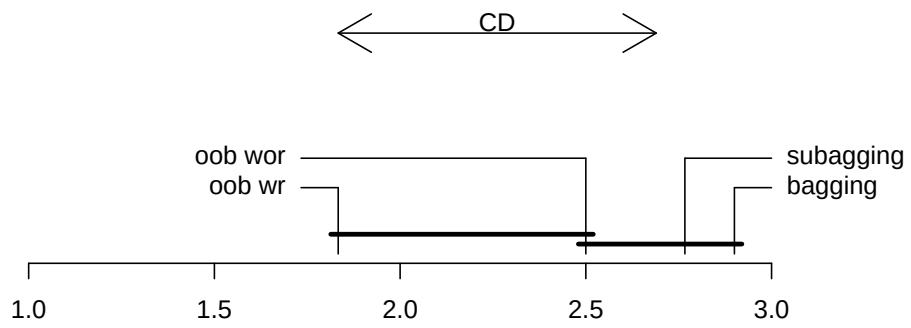

Figure 5: Comparison of the different methods using the Namenyi test. Classifiers not significantly different (p-value=0.05) are connected.

which the sampling ratio is estimated using out-of-bag data. In this diagram methods whose average ranks are not significantly different according to the Nemenyi test (p-value<0.05) appear connected by a horizontal line. The critical difference is shown for reference (CD=0.86 for 4 methods, 30 dataset and $\alpha = 0.05$). The best performance corresponds to the *oob wr* ensembles, which employ samples extracted with replacement, followed by *oob wor* ensembles, where without-replacement sampling is used. The out-of-bag with-replacement ensembles are significantly better than both standard bagging and subagging. In the case of subagging there is not enough evidence to determine whether the differences between sampling ratios estimated using out-of-bag data and the standard choice (half subsampling) are statistically significant.

5. Conclusions

In this article we have carried out an empirical analysis of the dependence of the generalization performance of m-out-of-n bagging and subagging with the size of the samples used to train the individual ensemble classifiers. This investigation shows that there is a correspondence between the results of with-replacement and without-replacement sampling for ensembles of classification trees when the corresponding sampling ratios are related by (1). In most of the problems investigated the sampling ratios that are optimal in terms of the generalization performance of the ensemble do not coincide with the conventional choices. The optimal value of the sampling ratios is problem dependent and can be estimated using the out-of-bag estimate of the generalization error of the ensemble.

In most cases, sampling ratios smaller than the standard choices are selected by the out-of-bag method proposed. Using smaller training samples has some advantages: First, the ensembles are generated faster. Second, classification trees trained on smaller samples also tend to be simpler. These properties entail a reduction in storage needs for the ensemble and an acceleration of the classification process, both of which are desirable properties, especially in on-

line applications. In terms of classification performance, using smaller bootstrap samples tends to increase the diversity of the classifiers and, therefore, the likelihood that the errors made by the different classifiers are uncorrelated. In this manner, incorrect classifications can be compensated by pooling the decisions of the classifiers in the ensemble. Another reason why using smaller samples can be useful to improve the generalization performance of the ensemble is that this procedure effectively smooths the feature space by reducing the influence of isolated examples whose class label is different from the class label of neighboring examples: As the sampling ratio becomes smaller, some training examples are included in less than half of the bootstrap samples. Consequently, the regions in the neighborhood of these examples are classified by the ensemble taking into account the class labels of nearby instances as well.

References

[1] L. Breiman, Bagging predictors, Machine Learning 24 (2) (1996) 123–140.

[2] J. R. Quinlan, Bagging, boosting, and C4.5, in: Proc. 13th National Conference on Artificial Intelligence, Cambridge, MA, 1996, pp. 725–730.

[3] D. Opitz, R. Maclin, Popular ensemble methods: An empirical study, Journal of Artificial Intelligence Research 11 (1999) 169–198.

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

[5] T. G. Dietterich, An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization, Machine Learning 40 (2) (2000) 139–157.

[6] G. I. Webb, Multiboosting: A technique for combining boosting and wagging, Machine Learning 40 (2) (2000) 159–196.

[7] R. Caruana, A. Niculescu-Mizil, An empirical comparison of supervised learning algorithms, in: ICML '06: Proceedings of the 23rd international conference on Machine learning, ACM Press, New York, NY, USA, 2006, pp. 161–168. doi:http://doi.acm.org/10.1145/1143844.1143865.

[8] B. Efron, R. J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall/CRC, 1994.

[9] P. Bühlmann, B. Yu, Analyzing bagging, Annals of Statistics 30 (2002) 927–961.

[10] A. Buja, W. Stuetzle, Observations on bagging, Statistica Sinica 16 (2006) 323–351.

[11] J. H. Friedman, P. Hall, On bagging and nonlinear estimation, Journal of Statistical Planning and Inference 137 (3) (2007) 669–683.

[12] J. Hartigan, Using subsample values as typical values, Journal of the American Statistical Society 64 (1969) 1303–1317.

[13] L. Breiman, Out-of-bag estimation, Tech. rep., Statistics Department, University of California (1996).

[14] J. W. H. Swanepoel, A note on proving that the (modified) bootstrap works, Communications in Statistics - Theory and Methods 15 (1986) 3193–3203.

[15] P. J. Bickel, F. Gtze, W. R. van Zwet, Resampling fewer then n observations: gains, losses, and remedies for losses, Statistica Sinica 7 (1997) 1–31.

[16] K.-H. Chung, S. M. S. Lee, Optimal bootstrap sample size in construction of percentile confidence bounds, Scandinavian journal of statistics 28 (2001) 225–239.

[17] D. Politis, J. P. Romano, M. Wolf, Subsampling, Springer Series in Statistics, Springer, 1999.

[18] A. C. Davison, D. V. Hinkley, G. A. Young, Recent developments in bootstrap methodology, Statistical Science 18 (2003) 141–157.

[19] L. Breiman, Pasting small votes for classification in large databases and on-line, Machine Learning 36 (1-2) (1999) 85–103.

[20] P. Bühlmann, Bagging, subagging and bragging for improving some prediction algorithms, in: Recent Advances and Trends in Nonparametric Statistics (eds. Akritas, M.G. and Politis, D.N.), Elsevier, 2003, pp. 19–34.

[21] M. Terabe, T. Washio, H. Motoda, The effect of subsampling rate on s3bagging performance, in: Proc. of ECML2001/PKDD2001 Workshop on Active Learning, Database Sampling, and Experimental Design: Views on Instance Selection, 2001, pp. 48–55.

[22] P. Hall, R. J. Samworth, Properties of bagged nearest neighbour classifiers, Journal of the Royal Statistical Society Series B 67 (3) (2005) 363–379.

[23] P. J. McCarthy, Replication: an approach to the analysis of data from complex surveys, Vital Health Statistics. Public Health Service Publication 14.

[24] B. Efron, The jackknife, the bootstrap, and other resampling plans, Society of Industrial and Applied Mathematics CBMS-NSF Monographs 38.

[25] A. Asuncion, D. Newman, UCI machine learning repository (2007).
URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`

[26] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, Classification and Regression Trees, Chapman & Hall, New York, 1984.

[27] L. Breiman, Arcing classifiers, The Annals of Statistics 26 (3) (1998) 801–849.

[28] J. Demšar, Statistical comparisons of classifiers over multiple data sets, Journal of Machine Learning Research 7 (2006) 1–30.