Open Journal of Statistics, **2011, 1, 205-211**
doi:10.4236/ojs.2011.13024 Published Online October 2011 (http://www.SciRP.org/journal/ojs)

Bias of the Random Forest Out-of-Bag (OOB) Error for Certain Input Parameters

Matthew W. Mitchell
Metabolon Inc., *Research Triangle Park*, *New Caledonia*, *USA*
E-mail: *mmitchell@metabolon.com*
Received June 8, 2011; *revised July* 10, 2011; *accepted July* 18, 2011

Abstract

Random Forest is an excellent classification tool, especially in the *omics* sciences such as metabolomics, where the number of variables is much greater than the number of subjects, *i.e.*, "$n \ll p$". However, the choices for the arguments for the random forest implementation are very important. Simulation studies are performed to compare the effect of the input parameters on the predictive ability of the random forest. The number of variables sampled, *m-try*, has the largest impact on the true prediction error. It is often claimed that the out-of-bag error (OOB) is an unbiased estimate of the true prediction error. However, for the case where $n \ll p$, with the default arguments, the out-of-bag (OOB) error *overestimates* the true error, *i.e.*, the random forest actually performs *better* than indicated by the OOB error. This bias is greatly reduced by subsampling without replacement and choosing the same number of observations from each group. However, even after these adjustments, there is a low amount of bias. The remaining bias occurs because when there are trees with equal predictive ability, the one that performs better on the in-bag samples will perform worse on the out-of-bag samples. Cross-validation can be performed to reduce the remaining bias.

Keywords: Random Forest, Multivariate Classification, Metabolomics, Small *n* Large *p*

1. Introduction

Random forest [1] is an ensemble method based on aggregating predictions from a large number of decision trees. Some of the advantages of random forest classification are the following: it is invariant to transformation, it is resistant to outliers, it does not overfit the data, it is fairly easy to implement with available software, and it works well when the number of subjects, *n*, is much fewer than the number of variables, *p*, *i.e.*, "$n \ll p$." Breiman discusses the properties of random forest for the various input parameters in his seminal paper [1]. However, in this discussion, the number of samples was larger than the number of variables. Thus, these properties may differ when $n \ll p$. Strobl *et al*. [2] have observed that there is bias in variable selection when subsampling with replacement (the default) is used, but the effect on the out-of-bag (OOB) error is not assessed. It is often stated that the OOB error is an unbiased estimate of the true prediction error. However, we will show that this is not necessarily the case.

In this manuscript, we compare the effect on the prediction error for different choices of the input parameters: 1) for subsampling with versus without replacement (the *replace* parameter); 2) the proportions of observations used for the in-bag samples (the *sampsize* parameter); and 3) various values of the number of variables sampled (the *m-try* parameter). These are compared for various simulated data sets with varying dimensions. Additionally, for each of these scenarios we compare the OOB error to the true prediction error where the choice of these parameters can cause a severe *overestimation* of the true error. This is in contrast to many methods, which can easily overfit the data and *underestimate* the true prediction error.

2. Simulation Study

To compare the effects of various input parameters on the OOB error estimate and the true prediction error, various models are simulated. Three models are compared: 1) a model of random noise; 2) a model with 20 true predictors; and 3) a model with 40 true predictors and correlated variables. More specific details for each model are given below. All models are simulated for p = 400 variables, and for two groups with sizes $n_1 = n_2 = 6$,

$n_1 = n_2 = 10$, and $n_1 = n_2 = 30$. The dimensions of the data sets were chosen to mimic metabolomics data: rat studies typically have 10 or fewer rats per group, and groups of size 30 are more common for human studies. Subsampling with replacement (the default) and subsampling without replacement are compared for various proportions of in-bag observations (the default is approximately 63% of the total observations). The *m-try* parameter was set to 4, 20 (the default, which is the square root of the number of variables), 200 and 400 (all variables).

The value of *m-try* made no difference for the conclusions for the random noise model, Model 1, so the results are shown only for *m-try* = 20, the default. For each random forest, 1000 trees were used. For each combination, 500 simulation runs were performed. For each simulated data set, a test set with the same dimensions was simulated, so that the OOB estimate of the error can be compared to the actual prediction error. All simulations were performed with *R* [3], using the *randomForest* package [4].

The results for Model 1 for $n_1 = n_2 = 6$, $n_1 = n_2 = 10$, and $n_1 = n_2 = 30$ are shown in **Tables 1-3**, while the results for Models 2 and 3 are shown in **Tables 4-6**, and **Tables 7-9**, respectively. Each entry in the table represents the average value across the simulation runs, and in parentheses, the margin of error for a 95% confidence interval is given (*i.e.*, $1.96 \times s/\sqrt{500}$). The bias is the average OOB error minus the average test set prediction error. The "N" column represents the *sampsize* argument.

Simulated Models

Model 1

This model is random noise: there are 400 independent normal random variables with mean zero and standard deviation equal to 0.3 for each group.

Model 2

This model has 20 true predictors with independent errors. More formally for Group 1, $\mathbf{X} = (\mathbf{X_1}, \cdots, \mathbf{X_{400}})'$ is multivariate normal with mean vector $(0.26 \times \mathbf{1_{10}}, \mathbf{0_{390}})$ where $\mathbf{1_p}$ is a vector of p ones and $\mathbf{0_q}$ is a vector of q zeroes. The covariance matrix is $(0.3)^2 \times \mathbf{I_{400}}$ where $\mathbf{I_{400}}$ is the 400 by 400 identity matrix. Group 2 is identical except that the mean vector is $(\mathbf{0_{10}}, 0.26 \times \mathbf{1_{10}}, \mathbf{0_{380}})$.

Table 1. Mean error rates for Model 1: random noise, $n_1 = n_2 = 6$, *mtry* = 20.

REPLACE	N	OOB Err	Test Err	Bias
Y	Dflt	0.808 (0.013)	0.504 (0.012)	+0.304 (0.017)
N	Dflt	0.809 (0.013)	0.504 (0.013)	+0.306 (0.019)
Y	3, 3	0.597 (0.015)	0.497 (0.012)	+0.100 (0.020)
N	3, 3	0.512 (0.015)	0.496 (0.013)	+0.016 (0.020)
Y	4, 4	0.601 (0.015)	0.496 (0.013)	+0.107 (0.020)
N	4, 4	0.489 (0.016)	0.506 (0.013)	–0.018 (0.020)
Y	5, 5	0.633 (0.015)	0.485 (0.012)	+0.148 (0.020)
N	5, 5	0.503 (0.017)	0.496 (0.012)	+0.007 (0.021)

Table 2. Mean error rates for Model 1: random noise, $n_1 = n_2 = 10$, *mtry* = 20.

REPLACE	N	OOB Err	Test Err	Bias
Y	Dflt	0.689 (0.012)	0.493 (0.010)	+0.196 (0.016)
N	Dflt	0.667 (0.013)	0.497 (0.009)	+0.170 (0.015)
Y	5, 5	0.569 (0.012)	0.492 (0.010)	+0.077 (0.015)
N	5, 5	0.513 (0.013)	0.503 (0.010)	+0.010 (0.016)
Y	7, 7	0.558 (0.013)	0.503 (0.009)	+0.055 (0.016)
N	7, 7	0.506 (0.013)	0.501 (0.009)	+0.005 (0.016)
Y	9, 9	0.587 (0.013)	0.507 (0.010)	+0.080 (0.016)
N	9, 9	0.499 (0.012)	0.507 (0.010)	–0.008 (0.016)
N	9, 9	0.499 (0.012)	0.507 (0.010)	–0.008 (0.016)

Table 3. Mean error rates for Model 1: random noise, $n_1 = n_2 = 30$, *mtry* = 20.

REPLACE	N	OOB Err	Test Err	Bias
Y	Dflt	0.570 (0.007)	0.499 (0.005)	+0.071 (0.009)
N	Dflt	0.555 (0.008)	0.501 (0.006)	+0.054 (0.009)
Y	15, 15	0.521 (0.007)	0.500 (0.006)	+0.021 (0.009)
N	15, 15	0.502 (0.007)	0.499 (0.007)	+0.004 (0.009)
Y	21, 21	0.525 (0.007)	0.500 (0.007)	+0.025 (0.009)
N	21, 21	0.505 (0.008)	0.498 (0.008)	+0.006 (0.009)
Y	27, 27	0.525 (0.007)	0.496 (0.007)	+0.029 (0.009)
N	27, 27	0.500 (0.008)	0.499 (0.008)	+0.001 (0.010)

Model 3

This model has 40 true predictors with two clusters of five correlated variables each and two clusters of five correlated noise variables. More formally, for Group 1, $\mathbf{X} = (\mathbf{X_1}, \cdots, \mathbf{X_{400}})'$ is multivariate normal with mean vector $(0.26 \times \mathbf{1_{20}}, \mathbf{0_{380}})$, and for Group 2, the mean vector is $(0_{20}, 0.26 \times \mathbf{1_{20}}, \mathbf{0_{360}})$. For each group, the covariance matrix is $(0.3)^2 \times \mathbf{R}$ with elements r_{ij}, where $r_{ii} = 1$. For i, $j = 1, 2, \cdots, 5$, $r_{ij} = 0.9$ for $i \neq j$. For $i, j = 21, 22, \cdots, 25$, $r_{ij} = 0.9$ for $i \neq j$. For $i, j = 41, 42, \cdots, 45$, $r_{ij} = 0.9$ for $i \neq j$. For $i, j = 51, 52, \cdots, 55$, $r_{ij} = 0.9$ for $i \neq j$. Otherwise, $r_{ij} = 0$.

3. Discussion

From all the tables, we see that the actual prediction error is similar for most of the combinations for each model, and the prediction error decreases with increasing sample sizes for Models 2 and 3. The *sampsize* parameter has little effect, but shows some degradation when 90% of the samples are used for the in-bag samples. The parameter with the largest impact on the true prediction accuracy is *m-try*, with the performance degrading for

Table 4. Mean error rates for Model 2: 20 true predictors, $n_1 = n_2 = 6$.

REPLACE	N	mtry	OOB Err	Test Err	Bias
Y	Dflt	4	0.77 (0.012)	0.356 (0.013)	+0.414 (0.018)
N	Dflt	4	0.767 (0.012)	0.339 (0.012)	+0.428 (0.017)
Y	3, 3	4	0.502 (0.014)	0.368 (0.012)	+0.134 (0.018)
N	3, 3	4	0.364 (0.014)	0.352 (0.012)	+0.012 (0.018)
Y	4, 4	4	0.525 (0.014)	0.357 (0.012)	+0.169 (0.018)
N	4, 4	4	0.364 (0.015)	0.34 (0.011)	+0.024 (0.019)
Y	5, 5	4	0.531 (0.015)	0.342 (0.012)	+0.189 (0.018)
N	5, 5	4	0.383 (0.014)	0.337 (0.012)	+0.046 (0.018)
Y	Dflt	20	0.683 (0.017)	0.348 (0.013)	+0.336 (0.019)
N	Dflt	20	0.676 (0.016)	0.336 (0.012)	+0.34 (0.02)
Y	3, 3	20	0.472 (0.015)	0.344 (0.012)	+0.128 (0.019)
N	3, 3	20	0.358 (0.015)	0.336 (0.012)	+0.022 (0.019)
Y	4, 4	20	0.469 (0.015)	0.342 (0.012)	+0.127 (0.019)
N	4, 4	20	0.363 (0.016)	0.342 (0.012)	+0.021 (0.02)
Y	5, 5	20	0.495 (0.017)	0.354 (0.012)	+0.141 (0.02)
N	5, 5	20	0.378 (0.016)	0.343 (0.012)	+0.035 (0.021)
Y	Dflt	200	0.61 (0.018)	0.358 (0.013)	+0.252 (0.022)
N	Dflt	200	0.583 (0.019)	0.354 (0.012)	+0.23 (0.023)
Y	3, 3	200	0.456 (0.015)	0.356 (0.012)	+0.099 (0.018)
N	3, 3	200	0.358 (0.014)	0.334 (0.012)	+0.024 (0.018)
Y	4, 4	200	0.468 (0.015)	0.342 (0.012)	+0.126 (0.019)
N	4, 4	200	0.396 (0.018)	0.359 (0.013)	+0.037 (0.022)
Y	5, 5	200	0.452 (0.018)	0.353 (0.012)	+0.099 (0.021)
N	5, 5	200	0.415 (0.021)	0.399 (0.013)	+0.016 (0.024)
Y	Dflt	400	0.622 (0.019)	0.361 (0.013)	+0.261 (0.023)
N	Dflt	400	0.582 (0.019)	0.347 (0.013)	+0.235 (0.022)
Y	3, 3	400	0.475 (0.015)	0.347 (0.012)	+0.128 (0.018)
N	3, 3	400	0.356 (0.014)	0.34 (0.012)	+0.016 (0.018)
Y	4, 4	400	0.453 (0.016)	0.338 (0.013)	+0.115 (0.02)
N	4, 4	400	0.371 (0.017)	0.34 (0.012)	+0.031 (0.021)
Y	5, 5	400	0.438 (0.018)	0.342 (0.012)	+0.096 (0.021)
N	5, 5	400	0.43 (0.022)	0.398 (0.013)	+0.032 (0.025)

Table 5. Mean error rates for Model 2: 20 true predictors, $n_1 = n_2 = 10$.

REPLACE	N	mtry	OOB Err	Test Err	Bias
Y	Dflt	4	0.565 (0.013)	0.295 (0.009)	+0.27 (0.016)
N	Dflt	4	0.542 (0.012)	0.296 (0.009)	+0.246 (0.014)
Y	5, 5	4	0.402 (0.011)	0.32 (0.009)	+0.082 (0.014)
N	5, 5	4	0.324 (0.01)	0.305 (0.009)	+0.018 (0.013)
Y	7, 7	4	0.408 (0.011)	0.299 (0.009)	+0.109 (0.014)
N	7, 7	4	0.329 (0.011)	0.294 (0.009)	+0.036 (0.014)
Y	9, 9	4	0.427 (0.011)	0.292 (0.009)	+0.135 (0.014)
N	9, 9	4	0.365 (0.011)	0.289 (0.009)	+0.076 (0.015)
Y	Dflt	20	0.481 (0.013)	0.271 (0.01)	+0.21 (0.016)
N	Dflt	20	0.462 (0.014)	0.267 (0.009)	+0.194 (0.017)
Y	5, 5	20	0.35 (0.013)	0.286 (0.009)	+0.063 (0.015)
N	5, 5	20	0.301 (0.012)	0.275 (0.009)	+0.026 (0.014)
Y	7, 7	20	0.359 (0.012)	0.278 (0.009)	+0.081 (0.014)
N	7, 7	20	0.298 (0.011)	0.272 (0.01)	+0.026 (0.014)
Y	9, 9	20	0.361 (0.012)	0.268 (0.009)	+0.093 (0.014)
N	9, 9	20	0.316 (0.013)	0.267 (0.009)	+0.048 (0.016)
Y	Dflt	200	0.389 (0.017)	0.309 (0.01)	+0.08 (0.02)
N	Dflt	200	0.399 (0.016)	0.308 (0.01)	+0.091 (0.02)
Y	5, 5	200	0.331 (0.012)	0.272 (0.009)	+0.059 (0.014)
N	5, 5	200	0.301 (0.013)	0.285 (0.01)	+0.016 (0.015)
Y	7, 7	200	0.317 (0.014)	0.293 (0.01)	+0.024 (0.017)
N	7, 7	200	0.323 (0.014)	0.317 (0.011)	+0.006 (0.018)
Y	9, 9	200	0.349 (0.014)	0.302 (0.01)	+0.048 (0.017)
N	9, 9	200	0.372 (0.017)	0.348 (0.011)	+0.024 (0.021)
Y	Dflt	400	0.402 (0.016)	0.308 (0.01)	+0.094 (0.02)
N	Dflt	400	0.399 (0.017)	0.323 (0.012)	+0.076 (0.022)
Y	5, 5	400	0.332 (0.013)	0.28 (0.009)	+0.052 (0.016)
N	5, 5	400	0.301 (0.013)	0.289 (0.01)	+0.012 (0.016)
Y	7, 7	400	0.333 (0.015)	0.303 (0.011)	+0.031 (0.018)
N	7, 7	400	0.344 (0.014)	0.323 (0.011)	+0.021 (0.019)
Y	9, 9	400	0.356 (0.015)	0.317 (0.01)	+0.039 (0.018)
N	9, 9	400	0.391 (0.018)	0.359 (0.011)	+0.032 (0.021)

the very large or very small values. However, the optimal choice of *m-try* will depend on the number of true predictors and their relationships.

For Model 1, the random noise model, the expected prediction error should be 50%. The true prediction errors, as measured by the average test set prediction errors, are indeed very close to 50%; however, the OOB error rates always overestimate the true error. With the default values of the inputs for the random forest function, the average OOB error is approximately 81% for $n_1 = n_2 = 6$ (see the first row of **Table 1**)! For the groups of sizes $n_1 = n_2 = 10$ and $n_1 = n_2 = 30$ the average OOB errors are 69% and 57%, respectively. This positive bias (*i.e.*, the OOB error is too *pessimistic*) is also seen for Models 2 and 3: for Model 2 the approximate biases are 34%, 21%, and 5%, for the groups of sizes $n_1 = n_2 = 6$, $n_1 = n_2 = 10$, and $n_1 = n_2 = 30$, respectively; while for Model 3, the positive biases are approximately equal to 32%, 15%, and 3%, when $n_1 = n_2 = 6$, $n_1 = n_2 = 10$, and $n_1 = n_2 = 30$, respectively. These results mirror those seen for the data sets with the group labels shuffled in [5], where the predictive ability of the variations of random forest are compared by using the OOB error rates for each. The OOB errors for the scrambled cases are much worse than random chance in most cases, with some OOB error rates equal to 75% and 83%. The data sets (gene arrays) used

Table 6. Mean error rates for Model 2: 20 true predictors, $n_1 = n_2 = 30$.

RE	N	mtry	OOB Err	Test Err	Bias
Y	Dflt	4	0.270 (0.006)	0.171 (0.005)	+0.100 (0.007)
N	Dflt	4	0.269 (0.006)	0.174 (0.004)	+0.094 (0.007)
Y	15, 15	4	0.225 (0.006)	0.191 (0.005)	+0.034 (0.007)
N	15, 15	4	0.206 (0.005)	0.183 (0.005)	+0.024 (0.007)
Y	21, 21	4	0.230 (0.006)	0.177 (0.005)	+0.053 (0.007)
N	21, 21	4	0.218 (0.006)	0.175 (0.005)	+0.043 (0.007)
Y	27, 27	4	0.235 (0.006)	0.174 (0.004)	+0.061 (0.007)
N	27, 27	4	0.266 (0.006)	0.171 (0.005)	+0.094 (0.007)
Y	Dflt	20	0.166 (0.005)	0.117 (0.004)	+0.048 (0.006)
N	Dflt	20	0.160(0.005)	0.117 (0.004)	+0.043 (0.006)
Y	15, 15	20	0.153 (0.005)	0.126 (0.004)	+0.026 (0.006)
N	15, 15	20	0.136 (0.005)	0.12 (0.004)	+0.016 (0.006)
Y	21, 21	20	0.145 (0.005)	0.121 (0.004)	+0.025 (0.006)
N	21, 21	20	0.135 (0.005)	0.119 (0.004)	+0.015 (0.006)
Y	27, 27	20	0.151 (0.005)	0.117 (0.004)	+0.034 (0.006)
N	27, 27	20	0.169 (0.005)	0.122 (0.004)	+0.047 (0.006)
Y	Dflt	200	0.183 (0.004)	0.166 (0.006)	+0.016 (0.007)
N	Dflt	200	0.195 (0.005)	0.183 (0.006)	+0.012 (0.008)
Y	15, 15	200	0.155 (0.004)	0.15 (0.005)	+0.005 (0.006)
N	15, 15	200	0.17 (0.004)	0.16 (0.005)	+0.01 (0.007)
Y	21, 21	200	0.166 (0.004)	0.156 (0.005)	+0.01 (0.007)
N	21, 21	200	0.195 (0.005)	0.188 (0.006)	+0.007 (0.008)
Y	27, 27	200	0.174 (0.004)	0.163 (0.006)	+0.011 (0.007)
N	27, 27	200	0.234 (0.006)	0.224 (0.006)	+0.01 (0.009)
Y	Dflt	400	0.214 (0.005)	0.197 (0.007)	+0.017 (0.009)
N	Dflt	400	0.227 (0.005)	0.21 (0.007)	+0.017 (0.009)
Y	15, 15	400	0.175 (0.004)	0.168 (0.006)	+0.007 (0.008)
N	15, 15	400	0.193 (0.004)	0.192 (0.007)	+0.001 (0.008)
Y	21, 21	400	0.189 (0.004)	0.175 (0.006)	+0.014 (0.008)
N	21, 21	400	0.23 (0.005)	0.227 (0.006)	+0.003 (0.009)
Y	27, 27	400	0.199 (0.004)	0.196 (0.006)	+0.003 (0.008)
N	27, 27	400	0.286 (0.007)	0.273 (0.006)	+0.013 (0.012)

Table 7. Mean error rates for Model 3: 40 true predictors with correlations, $n_1 = n_2 = 6$.

RE	N	mtry	OOB Err	Test Err	Bias
Y	Dflt	4	0.565 (0.013)	0.295 (0.009)	+0.27 (0.016)
N	Dflt	4	0.542 (0.012)	0.296 (0.009)	+0.246 (0.014)
Y	3, 3	4	0.402 (0.011)	0.32 (0.009)	+0.082 (0.014)
N	3, 3	4	0.324 (0.01)	0.305 (0.009)	+0.018 (0.013)
Y	4, 4	4	0.408 (0.011)	0.299 (0.009)	+0.109 (0.014)
N	4, 4	4	0.329 (0.011)	0.294 (0.009)	+0.036 (0.014)
Y	5, 5	4	0.427 (0.011)	0.292 (0.009)	+0.135 (0.014)
N	5, 5	4	0.365 (0.011)	0.289 (0.009)	+0.076 (0.015)
Y	Dflt	20	0.481 (0.013)	0.271 (0.01)	+0.21 (0.016)
N	Dflt	20	0.462 (0.014)	0.267 (0.009)	+0.194 (0.017)
Y	3, 3	20	0.35 (0.013)	0.286 (0.009)	+0.063 (0.015)
N	3, 3	20	0.301 (0.012)	0.275 (0.009)	+0.026 (0.014)
Y	4, 4	20	0.359 (0.012)	0.278 (0.009)	+0.081 (0.014)
N	4, 4	20	0.298 (0.011)	0.272 (0.01)	+0.026 (0.014)
Y	5, 5	20	0.361 (0.012)	0.268 (0.009)	+0.093 (0.014)
N	5, 5	20	0.316 (0.013)	0.267 (0.009)	+0.048 (0.016)
Y	Dflt	200	0.389 (0.017)	0.309 (0.01)	+0.08 (0.02)
N	Dflt	200	0.399 (0.016)	0.308 (0.01)	+0.091 (0.02)
Y	3, 3	200	0.331 (0.012)	0.272 (0.009)	+0.059 (0.014)
N	3, 3	200	0.301 (0.013)	0.285 (0.01)	+0.016 (0.015)
Y	4, 4	200	0.317 (0.014)	0.293 (0.01)	+0.024 (0.017)
N	4, 4	200	0.323 (0.014)	0.317 (0.011)	+0.006 (0.018)
Y	5, 5	200	0.349 (0.014)	0.302 (0.01)	+0.048 (0.017)
N	5, 5	200	0.372 (0.017)	0.348 (0.011)	+0.024 (0.021)
Y	Dflt	400	0.402 (0.016)	0.308 (0.01)	+0.094 (0.02)
N	Dflt	400	0.399 (0.017)	0.323 (0.012)	+0.076 (0.022)
Y	3, 3	400	0.332 (0.013)	0.28 (0.009)	+0.052 (0.016)
N	3, 3	400	0.301 (0.013)	0.289 (0.01)	+0.012 (0.016)
Y	4, 4	400	0.333 (0.015)	0.303 (0.011)	+0.031 (0.018)
N	4, 4	400	0.344 (0.014)	0.323 (0.011)	+0.021 (0.019)
Y	5, 5	400	0.356 (0.015)	0.317 (0.01)	+0.039 (0.018)
N	5, 5	400	0.391 (0.018)	0.359 (0.011)	+0.032 (0.021)

in the paper have tens of thousands of variables with a very low number of subjects (10 - 35 total). In that paper, the default bootstrap sampling (sampling with replacement and not necessarily the same number chosen from each group) was used. Hence comparing the OOB error rates was not appropriate as these can be severely biased.

For a given *sampsize* argument, subsampling without replacement generally has a lower bias than subsampling with replacement (the default). Furthermore, when subsampling without replacement, forcing the same proportion to be sampled from each group reduces the bias over the default (no restriction of proportion for each group, only the total). This bias occurs because this subsampling will oversample from one of the two groups, so that an individual tree is weighted towards predicting that group over the other. However, the other group is more represented in the out-of-bag samples, resulting in predictions that *appear* worse than random chance. For example, suppose for two groups when $n_1 = n_2 = 6$ that the in-bag samples have five observations from Group 1 and one observation from Group 2. This tree will tend to predict most observations as belonging to Group 1, but the out-of-bag samples have one from Group 1 and five from Group 2. Such trees are created whenever the proportion of samples from each group is different for the in-bag samples, which results from subsampling with replace-

Table 8. Mean error rates for Model 3: 40 true predictors with correlations, $n_1 = n_2 = 10$.

RE	N	mtry	OOB Err	Test Err	Bias
Y	Dflt	4	0.418 (0.012)	0.184 (0.008)	+0.234 (0.014)
N	Dflt	4	0.399 (0.012)	0.183 (0.008)	+0.216 (0.015)
Y	5, 5	4	0.281 (0.01)	0.199 (0.008)	+0.081 (0.012)
N	5, 5	4	0.227 (0.01)	0.191 (0.008)	+0.037 (0.012)
Y	7, 7	4	0.287 (0.011)	0.188 (0.008)	+0.099 (0.013)
N	7, 7	4	0.212 (0.009)	0.184 (0.008)	+0.028 (0.012)
Y	9, 9	4	0.295 (0.011)	0.183 (0.008)	+0.112 (0.012)
N	9, 9	4	0.262 (0.01)	0.176 (0.007)	+0.085 (0.013)
Y	Dflt	20	0.327 (0.013)	0.175 (0.008)	+0.152 (0.015)
N	Dflt	20	0.326 (0.013)	0.184 (0.008)	+0.141 (0.016)
Y	5, 5	20	0.234 (0.01)	0.182 (0.008)	+0.052 (0.012)
N	5, 5	20	0.186 (0.009)	0.171 (0.008)	+0.016 (0.012)
Y	7, 7	20	0.231 (0.01)	0.174 (0.008)	+0.057 (0.013)
N	7, 7	20	0.205 (0.01)	0.172 (0.008)	+0.033 (0.012)
Y	9, 9	20	0.242 (0.011)	0.172 (0.008)	+0.07 (0.014)
N	9, 9	20	0.238 (0.01)	0.194 (0.008)	+0.044 (0.013)
Y	Dflt	200	0.304 (0.013)	0.242 (0.009)	+0.063 (0.017)
N	Dflt	200	0.33 (0.014)	0.236 (0.009)	+0.094 (0.018)
Y	5, 5	200	0.229 (0.011)	0.187 (0.008)	+0.042 (0.013)
N	5, 5	200	0.224 (0.01)	0.203 (0.009)	+0.021 (0.014)
Y	7, 7	200	0.258 (0.011)	0.207 (0.009)	+0.051 (0.015)
N	7, 7	200	0.272 (0.012)	0.254 (0.009)	+0.018 (0.017)
Y	9, 9	200	0.264 (0.012)	0.23 (0.009)	+0.034 (0.017)
N	9, 9	200	0.333 (0.016)	0.302 (0.01)	+0.031 (0.019)
Y	Dflt	400	0.318 (0.014)	0.258 (0.01)	+0.06 (0.019)
N	Dflt	400	0.324 (0.015)	0.26 (0.01)	+0.063 (0.019)
Y	5, 5	400	0.24 (0.011)	0.185 (0.008)	+0.055 (0.013)
N	5, 5	400	0.227 (0.011)	0.212 (0.008)	+0.015 (0.014)
Y	7, 7	400	0.254 (0.012)	0.228 (0.009)	+0.026 (0.016)
N	7, 7	400	0.292 (0.013)	0.265 (0.01)	+0.027 (0.018)
Y	9, 9	400	0.288 (0.012)	0.249 (0.01)	+0.038 (0.017)
N	9, 9	400	0.355 (0.017)	0.316 (0.011)	+0.039 (0.022)

Table 9. Mean error rates for Model 3: 40 true predictors with correlations, $n_1 = n_2 = 30$.

RE	N	mtry	OOB Err	Test Err	Bias
Y	Dflt	4	0.142 (0.004)	0.08 (0.003)	+0.062 (0.005)
N	Dflt	4	0.14 (0.005)	0.085 (0.003)	+0.055 (0.006)
Y	15, 15	4	0.117 (0.004)	0.092 (0.004)	+0.025 (0.005)
N	15, 15	4	0.105 (0.004)	0.085 (0.003)	+0.02 (0.005)
Y	21, 21	4	0.114 (0.004)	0.087 (0.003)	+0.027 (0.005)
N	21, 21	4	0.111 (0.004)	0.087 (0.004)	+0.024 (0.005)
Y	27, 27	4	0.122 (0.004)	0.087 (0.003)	+0.035 (0.005)
N	27, 27	4	0.154 (0.005)	0.085 (0.004)	+0.069 (0.006)
Y	Dflt	20	0.103 (0.004)	0.078 (0.004)	+0.025 (0.005)
N	Dflt	20	0.101 (0.004)	0.08 (0.004)	+0.021 (0.005)
Y	15, 15	20	0.089 (0.003)	0.074 (0.004)	+0.014 (0.005)
N	15, 15	20	0.079 (0.003)	0.073 (0.004)	+0.006 (0.004)
Y	21, 21	20	0.087 (0.003)	0.076 (0.004)	+0.011 (0.005)
N	21, 21	20	0.085 (0.003)	0.079 (0.004)	+0.006 (0.005)
Y	27, 27	20	0.091 (0.003)	0.074 (0.004)	+0.017 (0.004)
N	27, 27	20	0.108 (0.004)	0.081 (0.004)	+0.028 (0.005)
Y	Dflt	200	0.139 (0.004)	0.126 (0.006)	+0.013 (0.007)
N	Dflt	200	0.153 (0.004)	0.141 (0.006)	+0.012 (0.007)
Y	15, 15	200	0.109 (0.003)	0.104 (0.005)	+0.005 (0.006)
N	15, 15	200	0.128 (0.003)	0.121 (0.005)	+0.007 (0.006)
Y	21, 21	200	0.124 (0.003)	0.114 (0.005)	+0.01 (0.006)
N	21, 21	200	0.156 (0.004)	0.159 (0.006)	–0.002 (0.007)
Y	27, 27	200	0.131 (0.004)	0.123 (0.005)	+0.008 (0.006)
N	27, 27	200	0.209 (0.005)	0.194 (0.006)	+0.015 (0.009)
Y	Dflt	400	0.166 (0.004)	0.156 (0.007)	+0.01 (0.008)
N	Dflt	400	0.187 (0.004)	0.166 (0.007)	+0.021 (0.009)
Y	15, 15	400	0.123 (0.003)	0.12 (0.005)	+0.004 (0.006)
N	15, 15	400	0.151 (0.003)	0.147 (0.007)	+0.004 (0.007)
Y	21, 21	400	0.146 (0.003)	0.142 (0.006)	+0.004 (0.007)
N	21, 21	400	0.195 (0.004)	0.192 (0.007)	+0.004 (0.009)
Y	27, 27	400	0.159 (0.004)	0.148 (0.007)	+0.011 (0.008)
N	27, 27	400	0.251 (0.007)	0.25 (0.007)	+0.001 (0.012)

ment or subsampling without replacement but not forcing the proportion sampled from each group to be the same.

From **Tables 1-9** we see that the lowest biases occur when subsampling is performed without replacement and the proportion sampled from each group for the in-bag samples is the same. However, there is still bias remaining: the overwhelming majority of bias estimates are positive (each row represents an independent simulation run). The reason for this bias is more subtle: when there are variables of equal predictive ability on the whole data set, the one that performs worse on the out-of-bag samples for some combinations of in-bag observations will be chosen. To illustrate, two variables of equal predictive ability for two groups of size five are shown in **Table 10**. We see that the decision trees X1 > 1 or X2 > 1 have equal predictive ability (20% error). Now, suppose the *sampsize* argument is set to (4, 4). For many combinations of in-bag samples, the performance of these trees is identical (e.g., in-bag samples 02 - 09). However, when samples 01 - 04 and 06 - 09 are chosen, X1 will be chosen over X2. This tree will be 100% accurate on the in-bag samples. However, samples 05 and 10 will be predicted incorrectly. Likewise, if samples 02 - 05 and 07 - 10 are chosen, the scenario for X2 is identical. (Note: because, of the *m-try* argument, these two variables will not necessarily always be compared.) It can never be the case

Table 10. Two sample variables with equal predictive ability.

SAMPLE	GROUP	X1	X2
01	A	1.1	0.8*
02	A	1.2	1.1
03	A	1.3	1.2
04	A	1.4	1.3
05	A	0.8*	1.4
06	B	0.8	1.2*
07	B	0.9	0.8
08	B	0.7	0.9
09	B	0.6	0.7
10	B	1.2*	0.6

The entries with * indicate those misclassified for the trees X1 > 1, then A or X2 > 1, then A.

that with two variables of equal predictive ability that the one that performs worse on the in-bag samples but better on the out-of-bag samples will be chosen.

The modification of the random forest proposed by Strobl *et al*. [2] was also compared to see if the same behavior is exhibited. This implementation uses subsampling without replacement by default, so this source of bias is eliminated. Simulations were performed with the *cforest* function from the *party* library [2,6,7]. Since *m-try* = 20 performed well, this value was the only one used for these simulation runs. For sample sizes of six and ten per group, *cforest* failed to choose any variables and fit only the mean (resulting in 50% prediction accuracy on the test sets for every run for all three models). The results for the three models for the group size of 30 each are shown in **Table 11**, and we see a similar pattern to the bias as seen with the standard random forest. This is not surprising as the issue discussed in the previous paragraph applies here for the same reason.

To address the issue of the remaining bias, one can simply report the OOB error estimate an expected upper bound to the true prediction error, as the bias is fairly low when sampling without replacement and the same proportion from each group are chosen. Otherwise, cross-validation can be performed to further refine the error estimate. For Model 2, with $n_1 = n_2 = 30$, *mtry* = 20, and 15 from each group was sampled without replacement, we also estimate the true prediction error using the average error obtained using leave-one-out cross-validation (LOO-CV). The results are shown in **Table 12**, and we see that the bias is zero as desired using LOO-CV

Table 11. *Cforest* results, *mtry* = 20, $n_1 = n_2 = 30$.

MODEL	OOB Err	Test Err	Bias
Random noise	0.608 (0.008)	0.501 (0.006)	+0.107 (0.009)
20 True Predictors	0.153 (0.004)	0.102 (0.004)	+0.051 (0.006)
40 Predictors with Correlation	0.104 (0.004)	0.079 (0.004)	+0.025 (0.005)

Table 12. Leave-one-out Cross-Validation Results for Model 2, $n_1 = n_2 = 30$, *mtry* = 20, subsampling 15 from each group without replacement.

METHOD	Error Estimate	Test Error	Bias
OOB Error	0.134 (0.005)	0.122 (0.004)	+0.012 (0.006)
LOO-CV Error	0.122 (0.005)	0.122 (0.004)	0.000 (0.006)

(although the bias is low initially, so may make very little practical difference).

Finally we compare the OOB error for several of the genomics data sets shown in [5]. These data sets are available at http://www.rci.rutgers.edu/~cabrera/DNAMR, and consist of the "Astrocytoma" [8], "BreastCancer" [9], "Epilepsy" [10], and "HIV" [11] data sets. For each data set, subsampling without replacement was performed with *sampsize* set to 50% of the smaller group size (*i.e.*, if the group sizes are 6 and 8, *sampsize* was set to (3,3), the default value of *m-try* was used, and 10,000 trees were used for each random forest. Each of these data sets has more than 10,000 variables. We compare the OOB error rates reported in [5], which were based on the default bootstrap sampling to those obtained using the recommended arguments (which are actually upper bounds as described above). In each case, the error rate was reduced over those given in [5]—many by 50%. The results are shown in **Table 13**.

4. Conclusions

Various models were simulated for a variety of combinations of input parameters (*replace*, *sampsize*, and *m-try*) and sample sizes for random forest in order to assess the performance of the out-of-bag (OOB) error estimate and the actual prediction error. The *m-try* parameter had the largest effect on the actual predictive ability, while the other parameters had little effect on the actual predictive ability in most cases. However, these parameters have a large effect on the OOB error estimate, which for certain parameters causes a severe positive bias. This bias is greatly reduced by subsampling without replacement and choosing the same proportion of observations from each group for the in-bag samples. There is still a small remaining positive bias that results from the variable selec-

Table 13. Comparison of OOB Error for genomics data sets given in [5] to those obtained using subsampling without replacement and sampling the name number from each group.

Data Set	OOB Error Reported in [5]	OOB Error with Proposed
Astrocytoma	0.214	0.071
Breast Cancer	0.029	0.000
Epilepsy	0.154	0.077
HIV	0.357	0.179

tion, and performing cross-validation can further refine the error estimate. However, since the bias is low, one may simply prefer to report the OOB error as an expected upper bound to the actual error.

5. References

[1] L. Breiman, "Random Forests," *Machine Learning*, Vol. 45, No. 1, 2001, pp. 5-32. doi:10.1023/A:1010933404324

[2] C. Strobl, A. L. Boulesteix, A. Zeileis and T. Hothorn, "Bias in Random Forest Variable Importance Measures: Illustrations, Sources, and Solution," *BMC Bioinformatics*, Vol. 8, Article 25, 2007. doi:10.1186/1471-2105-8-25 http://www.biomedcentral.com/1471-2105/ 8/25

[3] R Development Core Team, "R: A Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, 2008. http://www.R-project.org

[4] A. Liaw and M. L. Weiner, "Classification and Regression Trees by RandomForest," *R News* 2002, Vol. 2, No. 3, 2002, pp. 18-22. http://www.webchem.science.ru.nl/PRiNS/ rF.pdf

[5] D. Amaratunga, J. Cabrera and Y.-S. Lee, "Enriched Random Rorests," *Bioinformatics*, Vol. 24, No. 18, 2008, pp. 2010-2014. doi:10.1093/bioinformatics/btn356

[6] T. Hothorn, P. Buehlmann, S. Dudoit, A. Molinaro and M. Van Der Laan, "Survival Ensembles," *Biostatistics*, Vol. 7, No. 3, 2006, pp. 355-373. doi:10.1093/biostatistics/kxj011

[7] C. Strobl, A. L. Boulesteix, T. Kneib, T. Augustin and A. Zeileis, "Conditional Variable Importance for Random Forests," *BMC Bioinformatics*, Vol. 9, Article 307, 2008. doi:10.1186/1471-2105-9-307 http://www.biomedcentral.com/1471-2105/9/307

[8] T. MacDonald, "Human Glioblastoma," 2001. http://pepr.cnmcrsearch.org/browse.do?action=list_prj_exp&proje- cted=65

[9] M. Chan, X. Lu, F. Merchant, J. D. Iglehart and P. Miron, "Gene Expression Profiling of NMU-Induced Rat Mammary Tumors: Cross Species Comparison with Human Breast Cancer," *Carcinogenesis*, Vol. 26, No. 8, 2005, pp. 1343-1353. doi:10.1093/carcin/bgi100

[10] D. N. Wilson, H. Chung, R. C. Elliott, E. Bremer, D. George and S. Koh, "Microarray Analysis of Postictal Transcriptional Regulation of Neuropeptides," *Journal of Molecular Neuroscience*, Vol. 25, No. 3, 2005, pp. 285-298. doi:10.1385/JMN:25:3:285

[11] E. Masliah, E. S. Roberts, D. Langford, I. Everall, L. Crews, A. Adame, E. Rockenstein and H. S. Fox, "Patterns of Gene Dysregulation in the Frontal Cortex of Patients with HIV Encephalitis," *Journal of Neuroimmunology*, Vol. 157, No. 1-2, 2004, pp. 163-175. doi:10.1016/j.jneuroim.2004.08.026

