Statistics and Its Interface Volume 2 (2009) 381–388

Search for the smallest random forest

Heping Zhang[*] and Minghui Wang

Random forests have emerged as one of the most commonly used nonparametric statistical methods in many scientific areas, particularly in analysis of high throughput genomic data. A general practice in using random forests is to generate a sufficiently large number of trees, although it is subjective as to how large is sufficient. Furthermore, random forests are viewed as "black-box" because of its sheer size. In this work, we address a fundamental issue in the use of random forests: how large does a random forest have to be? To this end, we propose a specific method to find a sub-forest (e.g., in a single digit number of trees) that can achieve the prediction accuracy of a large random forest (in the order of thousands of trees). We tested it on extensive simulation studies and a real study on prognosis of breast cancer. The results show that such sub-forests usually exist and most of them are very small, suggesting they are actually the "representatives" of the whole random forests. We conclude that the sub-forests are indeed the core of a random forest. Thus it is not necessary to use the whole forest for satisfying prediction performance. Also, by reducing the size of a random forest to a manageable size, the random forest is no longer a black-box.

Keywords and phrases: Random forest, Classification, Smallest forest.

1. INTRODUCTION

Breiman et al. (1984) popularized classification and regression trees (CART). Besides being a flexible nonparametric classification and regression method, tree structures produced from CART are intuitive to interpret and thus practically appealing (Breiman 1984; Zhang and Singer 1999). However, tree-based methods have two major limitations. First, a resulting tree can be unstable even with minor data perturbations, although this weakness is not unique, because other stepwise model and variable selection procedures have a similar limitation. Second, thanks to the advancement of genomics and informatics, high dimensional data are very common. As a result, a single tree cannot model the rich information in the data. For example, many studies (Zhang and Singer 2003; Ye et al. 2005; Chen et al. 2007) use tens of thousands of gene expressions to predict an outcome using several tens or hundreds of subjects. This is commonly referred to as the "large p (the number of genes) and small n (the number of subjects)" problem (Kosorok and Ma 2007; Zhang et al. 2008). The classic statistical view of "one optimal model" to a given data set may need to be broadened, because there may not be a parsimonious model that can summarize the richness of a data set of a massive size.

To overcome these two weaknesses, the method of forests has emerged as an ideal solution. Here, a forest refers to a constellation of many tree models. Because a forest consists of many trees, it is more stable and less prone to prediction errors as a result of data perturbations (Brieman 1996; 2001). For the same reason, i.e., having many trees, we have an opportunity to utilize more information (i.e., more variables) in the data set, and hence we can seek more insights into and have a deeper understanding of the data. In some applications, different trees may unravel alternative pathways to disease prognosis or development.

Although the method of forests addresses the two challenges that the tree-based methods face, it also loses some of the advantages that the tree-based methods possess. Most importantly, because of so many trees in a forest, it is impractical to present a forest or interpret a forest. This is what Breiman referred to as a "black-box" in his 2002 Wald lectures presented at the annual meeting of the Institute of Mathematical Statistics. Our goal is to explore whether it is possible to find a common ground between a forest and a single tree so that we retain the easy interpretability of the tree-based methods and avoid the problems that the tree-based methods suffer from. In other words, does a forest have to be large, or how small can a forest be? To answer this fundamental question, our key idea is to shrink the forest with two objectives: (a) to maintain a similar (or even better) level of prediction accuracy; and (b) to reduce the number of the trees in the forest to a manageable level.

2. METHODS

To shrink the size of a forest while maintaining the prediction accuracy, we need to consider methods that facilitate the process of selecting trees for removal. To this end, we consider three measures to determine the importance of a tree in a forest in terms of prediction performance in order to find the minimal size of the forest.

The first measure focuses on the prediction: a tree can be removed if its removal from the forest has the minimal impact on the overall prediction accuracy. This is done as follows. First, calculate the prediction accuracy of forest F, denoted by p_F. Second, for every tree, denoted by T, in

[*] Corresponding author.

forest F, calculate the prediction accuracy of forest F_{-T} that excludes T, denoted by $p_{F_{-T}}$. Let Δ_{-T} be the difference in prediction accuracy between F and F_{-T}.

$$(1) \qquad \Delta_{-T} = p_F - p_{F_{-T}}$$

The tree T^p with the smallest Δ_T is the least important one and hence subject to removal.

$$(2) \qquad T^p = \arg\min_{T \in F}(\Delta_{-T})$$

This method will be referred to as "by prediction."

The other two deletion methods are based on the similarity between two trees. The idea is that we can afford to remove a tree if it is "similar" to other trees in the forest. The measure of similarity is defined as follows. For each data point, we have a predicted outcome from any tree T, denoted by P_T. Given two trees T_i and T_j, the correlation of the predicted outcomes by the two trees are defined as:

$$(3) \qquad cor_{i,j} = cor(P_{T_i}, P_{T_j}), \quad i, j = 1, 2, \ldots, N_F$$

where N_F represents the size of the original random forest F. $cor_{i,j}$ gives rise to a similarity between the two trees. For tree T, the average of its similarities with all trees, denoted by ρ_T, in F_{-T} reflects the overall similarity between T and F_{-T}.

$$(4) \qquad \rho_T = \frac{1}{N_F - 1} \sum_{t \in F, t \neq T} cor_{t,T}$$

Then, the tree T^s with the highest rho_T is the most similar to the trees in F_{-T} and hence subject to removal.

$$(5) \qquad T^s = \arg\max_{T \in F}(\rho_T)$$

This method will be referred to as "by similarity."

We can also modify this method as follows. Assign the initial weight of each tree T in forest F to 1.

$$(6) \qquad w_T = 1, \quad T \in F$$

Then, we use equation (3) to evaluate the pairwise similarity $cor_{i,j}$ of two trees T_i and T_j in forest F, according to their predicted outcomes P_{T_i} and P_{T_j}. Next, we select the pair of trees being most similar, named T_{s1} and T_{s2}. Also, calculate the average of similarity ρ_{s1} and ρ_{s2} for the two trees. The tree T^{rs} with higher $\rho_{T^{rs}}$ is subject to removal.

$$(7) \qquad T^{rs} = \arg\max(\rho_{s1}, \rho_{s2})$$

Finally, considering the pairwise similarity, we calculate the new weights by distributing $w_{T^{rs}}$ to all other trees in $F_{-T^{rs}}$, proportional to the pairwise similarity in.

$$(8) \qquad w'_t = w_t + \frac{cor(T^{rs}, t)}{\rho_{T^{rs}}} * (N_f - 1), \quad t \in F_{-T_{rs}}$$

This method will be referred to as "by restricted similarity."

To select the optimal size sub-forest, we need to track the performance of the sub-forests. Let $h(i)$, $i = 1, \ldots, N_f - 1$ denote the performance trajectory of a sub-forest of i trees, where N_f is the size of the original random forest. Note that $h(i)$ is specific to the method measuring the performance, because there are many sub-forests with the same number of trees. If we have only one realization of $h(i)$, we select the optimal size i_{opt} of the sub-forest by maximizing $h(i)$ over $i = 1, \ldots, N_f - 1$.

$$(9) \qquad i_{opt} = \arg\max_{i=1,\ldots,N_f-1} (h(i))$$

If we have M multiple realizations of $h(i)$, we select the optimal size sub-forest by using the 1-se rule as described by Breiman et al. (1984). That is, we first compute the average $\overline{h}(i)$ and its standard error $\sigma(i)$. As discussed by Breiman et al. (1984) and Zhang and Singer (1999), the 1-se rule tends to yield a more robust and parsimonious model.

$$(10) \qquad \overline{h}(i) = \frac{1}{M} \sum_{j=1,\ldots,M} h_j(i), \quad i = 1, \ldots, N_f - 1$$

$$(11) \qquad \sigma(i) = var(h_1(i), \ldots, h_M(i)), \quad i = 1, \ldots, N_f - 1$$

Then, find the i_m that maximizes the average $\overline{h}(i)$ over $i = 1, \ldots, N_f - 1$.

$$(12) \qquad i_m = \arg\max_{i=1,\ldots,N_f-1} (\overline{h}(i))$$

Finally, we choose the smallest sub-forest such that its corresponding $\overline{h}$ is within one standard error (se) of $\overline{h}(i_m)$ as the optimal sub-forest size i_{opt}.

$$(13) \qquad i_{opt} = \min_{i=1,\ldots,M}(h(i) > (\overline{h}(i_m) - \sigma(i_m)))$$

We call i_{opt} the critical point of the performance trajectory.

3. RESULTS AND DISCUSSION

3.1 Simulation designs

For each data set, we generated 500 observations, each of which has one response variable and 30 predictors from Bernoulli distribution with success probability of 0.5. We randomly chose ν of the 30 variables to determine the response variable. For convenience, these ν predictors are labeled as, $X_1, \ldots, X_\nu$. Then, the response variable is defined as follows:

$$y = \begin{cases} 1, & \text{if } \frac{1}{\nu}\sum_{i=1}^{\nu} X_i + \epsilon > 0.5, \\ 0, & \text{otherwise} \end{cases}$$

where ϵ is a random variable following the normal distribution with mean zero and variance σ^2. We considered two choices for ϵ (5 and 10) and two choices of σ (0.1 and 0.3).

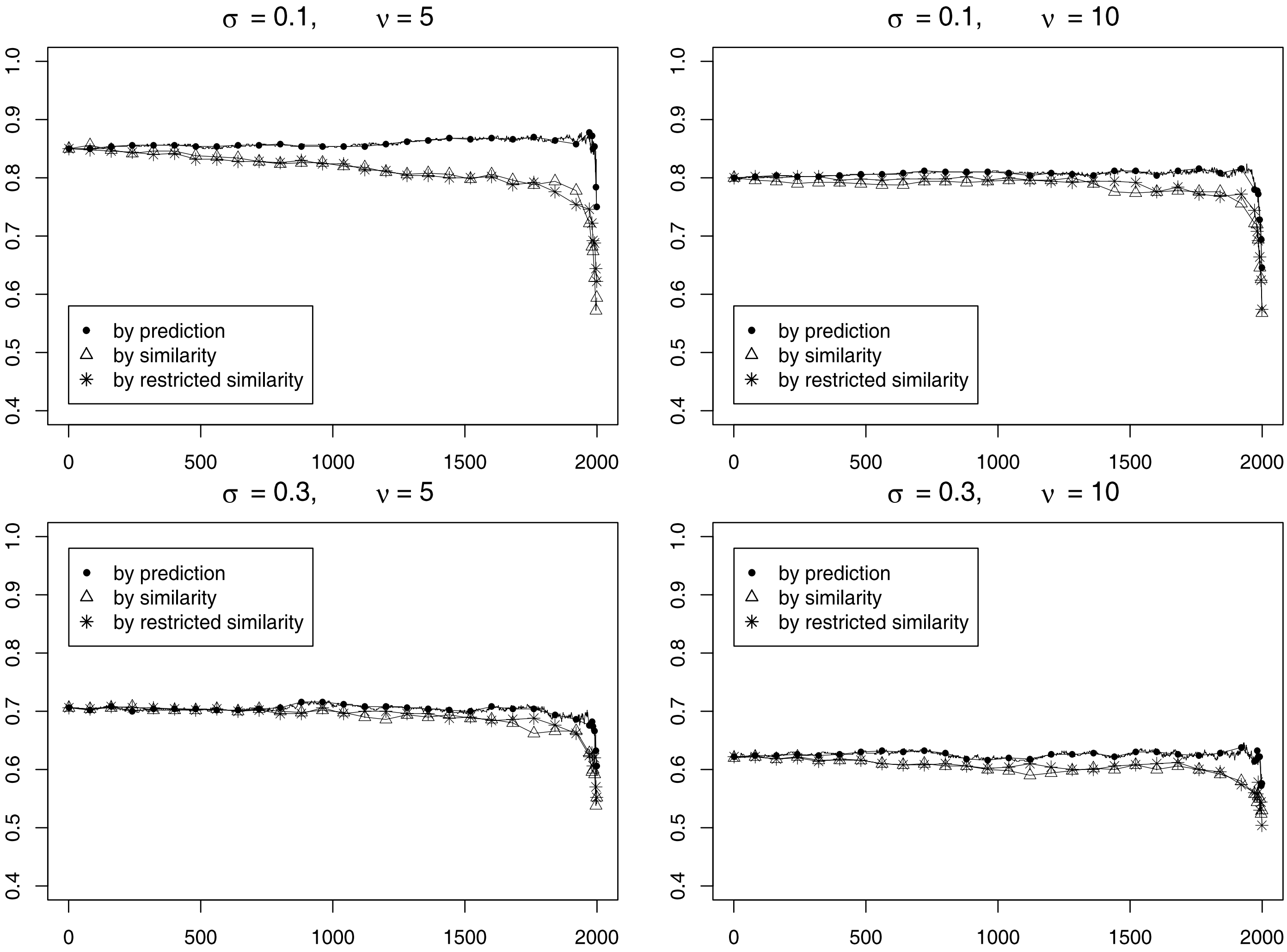

Figure 1. Prediction performance of sub-forests produced from different datasets and methods.

To perform an unbiased comparison of the three tree removal measures introduced in the Methods section below, we simulated independent data sets to train the initial random forest, to delete trees from the initial forest to produce sub-forests, and to evaluate the prediction performance of the sub-forests. These three data sets are referred to as the training set, the execution set, and the evaluation set. The generation and use of these three data sets constituted one run of simulation, and we replicated 100 times.

In practice, however, we generally have one data set only. That is, we may not have the execution and evaluation data sets as in our simulation. A practical question is: how do we select the optimal sub-forest with only one data set? To answer this question, we considered four bootstrap-based approaches and examined them in simulated data sets, leveraging the fact that we have the “golden” standard to be compared with in the simulated data set.

We begin with the construction of an initial forest using the whole data set as the training data set. In the first approach, we use one bootstrap data set for execution and the out-of-bag (oob) samples for evaluation. In the second approach, we use the oob samples for both execution and evaluation. In the third approach, we use the bootstrap samples for both execution and evaluation. Lastly, we re-draw bootstrap samples for execution and again re-draw bootstrap samples for evaluation.

3.2 Simulation results

To decide the size of the original random forest, we began with a random forest with a size of 100 trees. Then we increased the size gradually by step of 100 until the oob error rate was stable. We repeated the procedure on every dataset and ended up with selecting the size of 2000.

First, to gain insight into tree removal in a forest, in Fig. 1, we randomly selected one run of simulation and presented the stepwise change in the prediction performance as

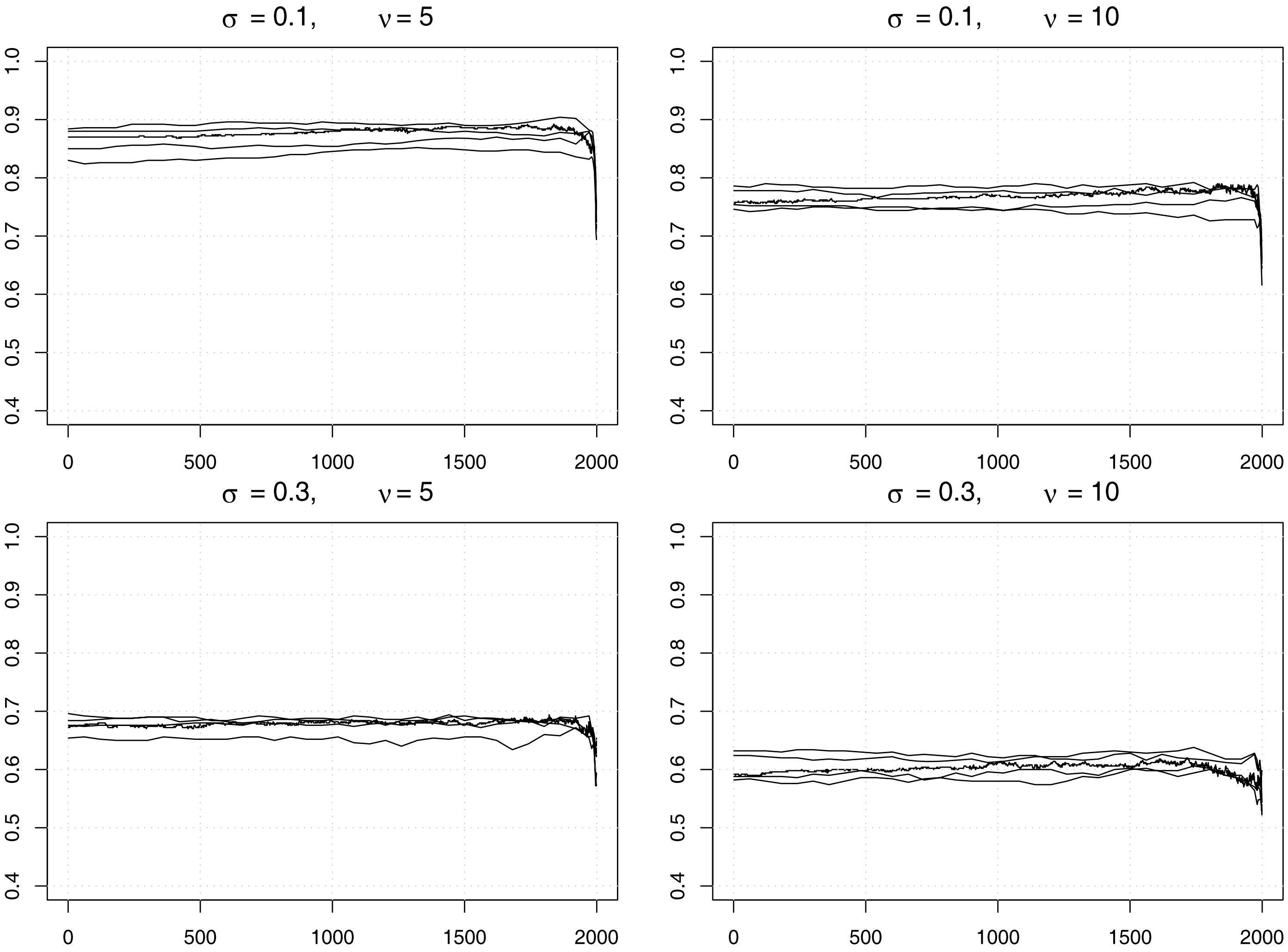

Figure 2. Performance trajectory of the "by prediction" method using the results in five randomly selected runs for four data sets.

a result of removing one least "favorable" tree at a time. Based on Fig. 1, the "by prediction" method is preferable because it can identify a critical point during the tree removal process in which the performance of the sub-forest deteriorates very rapidly. Figure 1 indicates that the performance of the sub-forests may begin to improve before the critical point. Therefore, the "by prediction" method can reduce the size of the forest to a manageable level while maintaining (or even improving) the prediction accuracy.

All of the simulations suggest consistently that the "by prediction" method is the preferable choice in achieving our goal. Thus, we will focus on the "by prediction" method from now on, even though we performed our simulations for all methods.

Second, Fig. 2 displays a summary plot of prediction performance using the results in five randomly selected runs. This figure provides some insights into how the prediction trajectories vary from one simulation run to another. Although the variation of the trajectories is notable, the sizes of the optimal sub-forests are within a reasonable range (11–36) for the "by prediction" method.

Table 1. The medians of the numbers of trees in the optimal sub-forests in 100 replications. The 1st quartile and 3rd quartile are provided in the parentheses

ν / σ	5	10
0.1	20(13,29)	31(20,40)
0.3	22(15,32)	18(11,37)

Lastly, Table 1 provides a summary from the 100 simulation runs. For a lower noise level ($\sigma = 0.1$), the performance of the sub-forests shows a kick-back before the critical point, and is better than that of the initial forest. When the noise level is relatively high ($\sigma = 0.3$), the performance of all

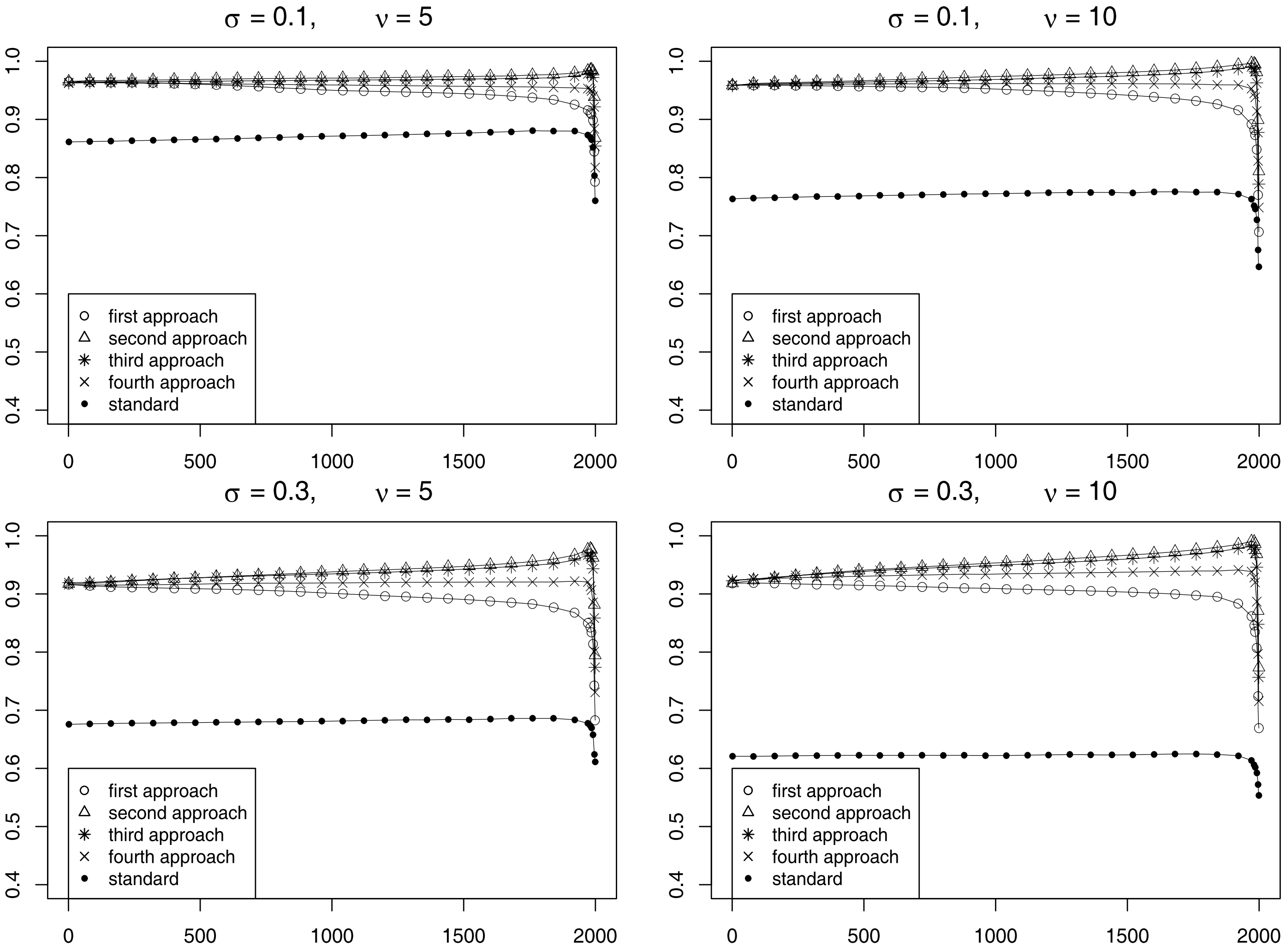

Figure 3. A performance summary plot of the "by prediction" method.

sub-forests remains stable until after the critical point is passed.

Figure 3 compares the performance of the four bootstrap-based approaches in the four simulation data sets. The comparison is based on the average performance in 100 runs.

It is expected that the performance trajectories of the four bootstrap-based approaches may not overlap with the "golden" standard. However, for the selection of the optimal sub-forest, the similarity among the trajectories is most relevant, because it could lead to the same or similar sub-forest. For this consideration, in Fig. 4, we examined the correlation between the original (the "golden" standard) trajectory and each of the four bootstrap approaches.

Figures 3 and 4 indicate that the first approach, i.e., using the bootstrap samples for execution and the oob samples for evaluation, is an effective sample-reuse approach to selecting the optimal sub-forest, when there is only one data set, i.e., in the real data application.

3.3 Prediction for breast cancer prognosis

It is documented that adjuvant systemic therapy substantially improves disease-free and overall survival in women with breast cancer up to the age of 70 years (Early Breast Cancer Trialists' Collaborative Group 1998), especially in patients with poor prognostic features benefit (National Institutes of Health Consensus Development Panel 2001). The main prognostic factors in breast cancer are age, tumor size, status of axillary lymph nodes, histologic type of the tumor, pathological grade, and hormone-receptor status (van de Vijver 2002). More recently, microarrays have been used to analyze breast-cancer tissues (Perou et al. 2000), to distinguish cancers associated with BRCA1 or BRCA2 mutations (Hedenfalk 2001; van't Veer 2002; Zhang 2002) and to determine estrogen-receptor status (van't Veer 2002) and lymph-node status (West 2001).

To provide a more accurate estimate of the risks of metastases associated with the two gene-expression signatures and

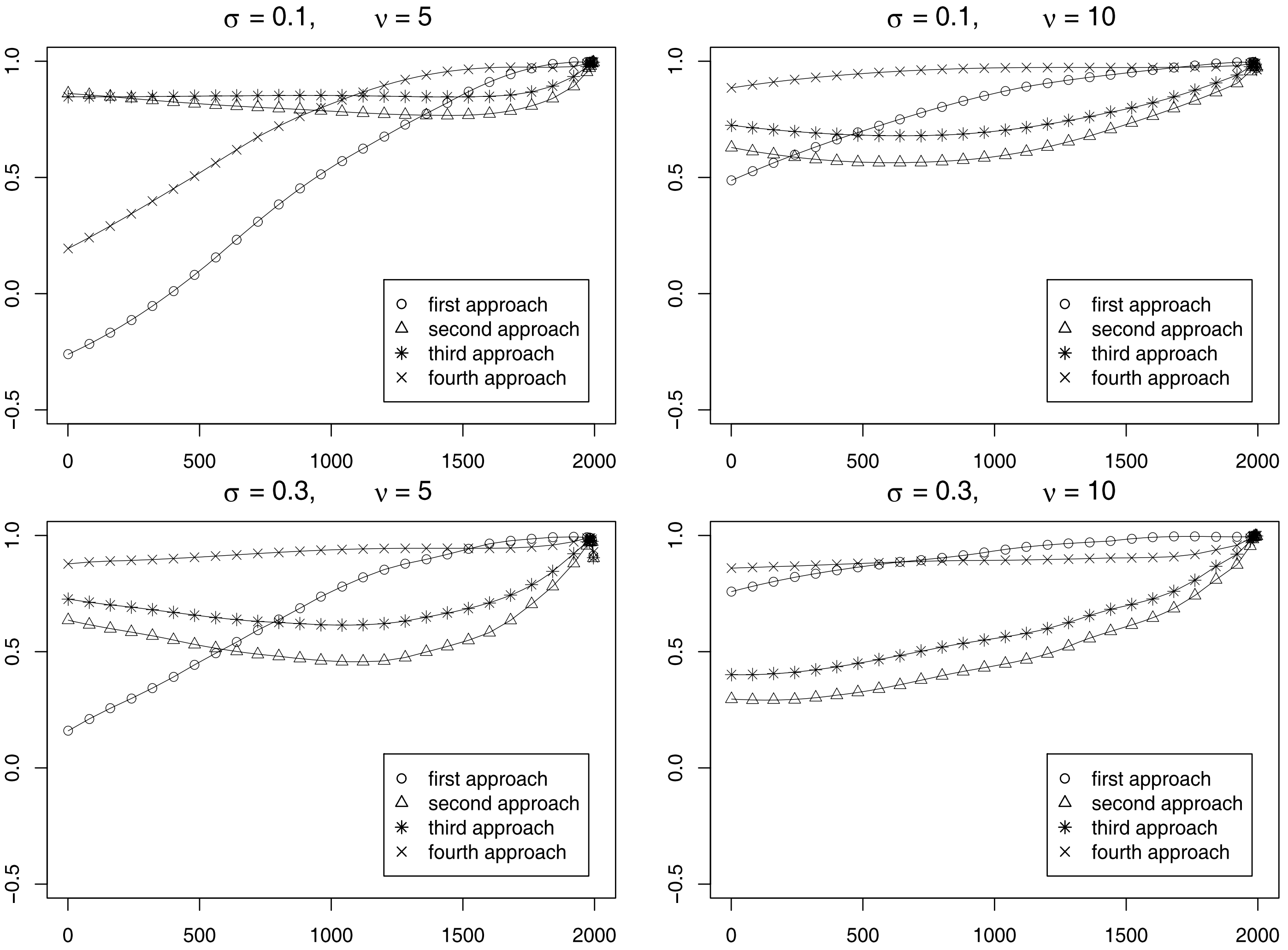

Figure 4. The correlation between the performance trend by each of the four bootstrap strategies and the "standard" curve.

to substantiate that the gene-expression profile of breast cancer is a clinically meaningful tool, van de Vijver et al. (2002) studied a cohort of 295 young patients with breast cancer, some of whom were lymph-node-negative and some of whom were lymph-node-positive. Using expression profiles from 70 previously selected genes, they concluded that the gene-expression profile in their study is a more powerful predictor of the outcome of disease in young patients with breast cancer than standard systems based on clinical and histologic criteria (van de Vijver 2002).

In this study we used the microarray data of that cohort to evaluate the performance of our method. The responses of all patients are defined by whether the patients remained disease-free five years after their initial diagnoses or not, as described in van't Veer et al. (2002). Seven patients in the cohort stopped follow-ups in less than 5 years and developed no metastases during this period. As in van't Veer et al. (2002), we removed those subjects from the cohort in the present analysis for consistency.

We used the "by prediction" measure as the optimization criterion for sub-forest, and then the original data set to construct an initial forest, a bootstrap data set for execution, and the oob samples for evaluation. We replicated the procedure for a total of 100 times. In each run, we used the oob samples to compare the performance of the initial random forest and the optimal sub-forest. The sizes of the optimal sub-forests fall in a relatively narrow range, of which the 1st quartile, the median, and the 3rd quartile are 13, 26 and 61, respectively.

Since our main goal is to shrink the initial random forest as small as possible to enable us to examine and understand the tree structures in the forest, we chose the smallest optimal sub-forest in the 100 repetitions with the size of 7.

To compare the performance of the initial random forest with this optimal sub-forest, we used the two forests as classifiers in the original data set. Table 2 presents the misclassification rates based on the oob samples. It should be noted that 19 out of the 288 samples were in the training

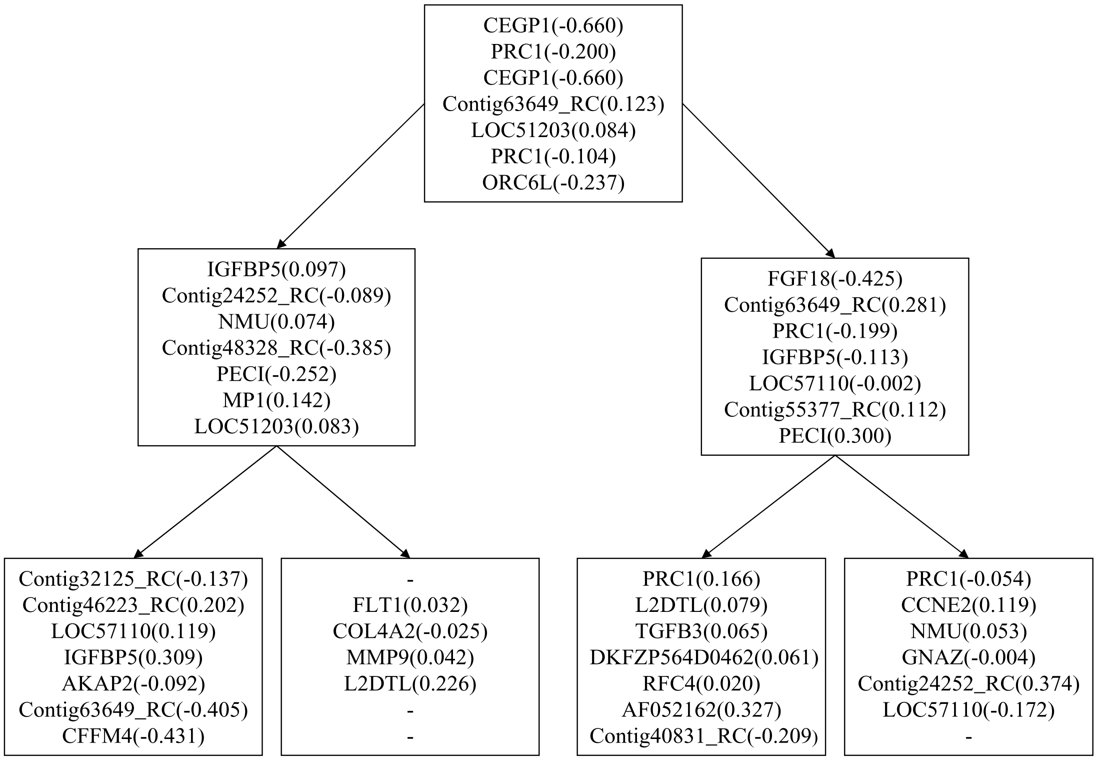

Figure 5. The top three layers of the optimal sub-forest consisting of seven trees.

Table 2. Comparison of prediction performance of the initial random forest, the optimal sub-forest, and a previously established 70-gene classifier

Method	Error rate	Predicted / True	Good	Poor
Random	26.0%	Good	141	17
Forest		Poor	53	58
Sub	26.0%	Good	146	22
forest		Poor	48	53
70-gene	35.3%	Good	103	4
classifier		Poor	91	71

data sets of all trees in the optimal sub-forest, and hence were not considered in the calculation of the oob misclassification rate for the optimal sub-forest in order to avoid bias. The initial forest and the optimal sub-forest achieve almost the same level of performance accuracy.

As a benchmark, we used the classifier proposed by Vijver et al. (2002), which was based on the 70-gene profile selected by van't Veer et al. (2002) with 78 samples. In our analysis, the number of samples (288 patients) is larger. As shown in Table 2, the 70-gene classifier has an out-of-bag error rate of 35.3%. Thus, its accuracy is much lower than those of the forests.

The results above demonstrate that performance of the optimal sub-forest is consistent with the initial random forest, and may offer a better approach than the classifier proposed by van de Veer et al. (2002). Thus, we effectively reduced the forest size and maintained the prediction accuracy of the initially large forest.

As we discussed earlier, a main motivation for us to seek the smallest possible forest is to enable us to examine the forest. To illustrate this, Fig. 5 displays the most critical part (the top three layers) of the optimal sub-forest consisting of the seven trees. It is interesting to note that the selected genes are quite diverse and unique.

4. CONCLUSIONS

Random forest has become a very useful tool for analyzing high dimensional data, particularly high-throughput genomic data including single nucleotide polymorphisms and gene expression profiling. While random forests tend to produce classifiers that are more accurate than many existing methods, as is also demonstrated in our analysis of breast cancer prognosis, they generally consist of so many trees that they are regarded as a "black-box" by its own inventor. The objective of this work is to reduce the forest size to the minimal level while maintaining the prediction accuracy at the comparable level with the initially large forest. We observed from our simulation studies and real data analysis that the size of the optimal sub-forest is in the range of tens and that some sub-forests can even over-perform the original forest in terms of prediction accuracy, likely due to their parsimonious property. Therefore, we have demonstrated that it is possible to construct a highly accurate random forest consisting of a manageable number of trees to allow in-depth examination of the trees and splits in the forests.

The key advantage of our proposed sub-forest is the ability to examine and present the forests. In the existing work (Zhang 2003), the examination of the forests is possible only at the descriptive level, namely, the frequency of the genes that are selected to split nodes. As demonstrated in Fig. 5, it is possible to present a sub-forest succinctly. We have noted above that the selected genes in Fig. 5 are quite diverse and unique. This may not be accidental because the sub-forest is selected to be very efficient in representing all genes considered. However, further studies are warranted to examine this efficiency.

The limitation of this work is that while we may have improved the existing classifier, future samples and studies are needed to evaluate the performance of the forest-based classifiers as highlighted in Fig. 5. In addition, even though the genes displayed in Fig. 5 are selected for the prediction purpose, it is also possible some of those genes are in the pathways to the breast cancer. This can only be confirmed through independent studies.

ACKNOWLEDGEMENTS

This research was supported in part by grants K02DA017713, R01DA016750, and U01HD050062 from the National Institutes of Health.

Received 19 July 2008

REFERENCES

[1] Breiman, L. (1996). Bagging predictors. *Machine Learning* **26** 123–140.

[2] Breiman, L. (2001). Random forests. *Machine Learning* **45** 5–32.

[3] Breiman, L., Friedman, J., Stone, C., and Olshen, R. (1984). *Classification and Regression Trees*. Chapman and Hall, New York. MR0726392

[4] Chen, X., Liu, C., Zhang, M., and Zhang, H. (2007). A forest-based approach to identifying gene and gene-gene interactions. *PNAS* **104** 19199–19203.

[5] Early Breast Cancer Trialists' Collaborative Group (1998). Polychemotherapy for early breast cancer: an overview of the randomised trials. *Lancet* **352** 930–942.

[6] Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M. et al. (2001). Gene-expression profiles in he-reditary breast cancer. *N Engl J Med* **344** 539–548.

[7] Kosorok, M. R. and Ma, S. (2007). Marginal asymptotics for the "large p, small n" paradigm: With applications to microarray data. *Ann Statist* **35** 1456–1486. MR2351093

[8] National Institutes of Health Consensus Development Panel (2001). National Institutes of Health Consensus Development Conference Statement: adjuvant therapy for breast cancer, November 1–3, 2000. *J Natl Cancer Inst* **93** 979–989.

[9] Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S. et al. (2000). Molecular portraits of human breast tumours. *Nature* **406** 747–752.

[10] van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A. M. et al. (2002). A Gene-Expression Signature as a Predictor of Survival in Breast Cancer. *N Engl J Med* **347** 1999–2009.

[11] van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M. et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415** 530–536.

[12] West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S. et al. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *PNAS* **98** 11462–11467.

[13] Ye, Y., Zhong, X., and Zhang, H. (2005). A genome-wide tree- and forest-based association analysis of comorbidity of alcoholism and smoking. *BMC genetics* **6(Suppl1)** S135.

[14] Zhang, H. and Singer, B. (1999). *Recursive Partitioning in the Health Sciences*. Springer, New York. MR1683316

[15] Zhang, H. and Yu, C. (2002). Tree-based analysis of microarray data for classifying breast cancer. *Frontiers in Bioscience* **7** c63–67.

[16] Zhang, H., Yu, C., and Singer, B. (2003). Cell and tumor classification using gene expression data: Construction of forests. *PNAS* **100** 4168–4172.

[17] Zhang, M., Zhang, D., and Wells, M. (2008). Variable selection for large p small n regression models with incomplete data: Mapping QTL with epistases. *BMC Bioinformatics* **9** 251.

Heping Zhang
Department of Epidemiology and Public Health
Yale University School of Medicine
New Haven, CT 06520-8034, USA
E-mail address: heping.zhang@yale.edu

Minghui Wang
Department of Epidemiology and Public Health
Yale University School of Medicine
New Haven, CT 06520-8034, USA
E-mail address: minghui.wang@yale.edu