Guided Random Forest in the RRF Package
Houtao Deng
Intuit, Mountain View, CA 94043, USA.

ABSTRACT
Summary: Random Forest (RF) is a powerful supervised learner and has been popularly used in many applications such as bioinformatics. In this work we propose the guided random forest (GRF) for feature selection.

Similar to a feature selection method called guided regularized random forest (GRRF), GRF is built using the importance scores from an ordinary RF. However, the trees in GRRF are built sequentially, are highly correlated and do not allow for parallel computing, while the trees in GRF are built independently and can be implemented in parallel.

Experiments on 10 high-dimensional gene data sets show that, with a fixed parameter value (without tuning the parameter), RF applied to features selected by GRF outperforms RF applied to all features on 9 data sets and 7 of them have significant differences at the 0.05 level. Therefore, both accuracy and interpretability can be significantly improved. GRF selects more features than GRRF, however, leads to better classification accuracy.

In this work we propose the guided random forest (GRF) for feature selection. In the code, a classification data set with 500 features is simulated, and only 2 features are relevant to the class. While RF uses all the features and misclassifies 54 out of 250 instances, RF applied to features selected by GRF is more accurate than RF.

2 METHODS
Let $gain(X_i)$ denote the Gini information gain of using a feature $X_i$ to split a tree node. The key idea of GRF is weighting $gain(X_i)$ using the importance scores from an RF.

$$gain_{iC}(X_i) = \lambda_i \cdot gain(X_i)$$

where $\lambda_i$ is calculated as

$$\lambda_i = 1 - \gamma + \gamma \frac{Imp}{Imp^*}$$

where $Imp_i$ is the importance score of $X_i$ from an RF, $Imp^*$ is the maximum importance score, $\frac{Imp}{Imp^*} \in [0, 1]$ is the normalized importance score, and $\gamma \in [0, 1]$ controls the weight of the importance scores from RF.

Note the key difference between GRF and GRRF is that the features used in a GRF model are expected to be relevant and non-redundant, while the features used in a GRRF model are expected to be relevant, but not necessarily non-redundant.

3 EXAMPLES AND RESULTS
Code [1] shows an example of using GRF ($\gamma = 1$) for feature selection. In the code, a classification data set with 500 features is simulated, and only 2 features are relevant to the class. While RF uses all the features and misclassifies 54 out of 250 instances, RF uses 196 features selected by GRF and misclassifies 34.
Let "¢" or "£" denote a significant difference between a method and GRF-RF at the 0.05 level, according to the paired t-test. Particularly, "¢" or "£" standards for a higher or lower error rate of a method compared to GRF-RF.

### Table 1. Error rates of GRF-RF (RF applied to the feature subset selected by GRF), GRF (as a classifier), RF, GRRF (as a classifier), GRF-RF (RF applied to the feature subset selected by GRRF), averaged over 100 runs.

| Method          | adenocarcinoma | brain      | breast.2 class | breast.3 class | leukemia | lymphoma | nci | prostate | rectum |
|-----------------|----------------|------------|----------------|----------------|----------|----------|-----|----------|--------|
| Error rate (%)  | 0.168          | 0.178      | 0.184          | 0.205          | 0.196    | 0.196    | 0.216| 0.216    | 0.216  |
| # 34 instances misclassified | 0             | 0          | 0              | 0              | 0        | 0        | 0   | 0        | 0      |

### Table 2. The number of instances, classes and features of the data sets, and the number of features used in GRF and RF.

| Data Set      | Instances | Classes | Features | GRF | RF |
|---------------|-----------|---------|----------|-----|----|
| adenocarcinoma| 96        | 2       | 96       | 472 | 2143 |
| brain         | 42        | 2       | 42       | 5597| 397  |
| breast.2 class| 77        | 2       | 77       | 4869| 385  |
| breast.3 class| 95        | 3       | 95       | 4869| 421  |
| leukemia      | 62        | 3       | 62       | 2090| 1291 |
| lymphoma      | 62        | 3       | 62       | 4026| 295  |
| nci           | 61        | 8       | 61       | 5244| 444  |
| prostate      | 102       | 2       | 102      | 6033| 414  |
| rectum        | 63        | 4       | 63       | 2308| 262  |

### Code 1. Feature Selection and Classification with GRF

```r
library(RRF) # load the RRF package
set.seed(1) # fix the random seed.
# simulate classification data set
X <- matrix(rnorm(500*500, min=-1, max=1), ncol=500)
# class is only relevant to feature 1 and 21
Y <- X[,1] + X[,21]
ix <- which(Y>quantile(Y, 1/2)); ncol <- length(ix)
y <- Y[ix] - 1 # assign class -1 and 1
split data into training and testing sets
trainX <- X[ix,]; testX <- X[-ix,]
# build an ordinary RF on the training instances
RF <- randomForest(trainX, trainY)
# the default gamma = 1.5 in Equation (2) (fixed as 1 here).
# note the difference between GRF and RF is that
# 'coefReg' is related to impRF in GRF, while
# it is constant for all variables in RF.
# build a GRF with gamma = 1
GRF <- GRRF(trainX, trainY, flagReg=0, coefReg=coefReg)
# test RF and GRF on the testing instances
pred <- predict(RF, testX)
predRF <- predict(GRF, testX)
```

In addition, I applied GRF-RF (RF applied to the feature subset selected by GRF), GRF, GRRF (γ = 0.1) and GRRF-RF (RF applied to the feature subset selected by GRRF), all with 1000 trees, to 10 gene data sets used in Diaz-Uriarte and De Andres (2006); Deng and Runger (2013). The references of the data sets are provided in a supplementary file to save space. I obtained the average error rates and average number of features for each method using the same procedure as Deng and Runger (2013), i.e., calculated from 100 replicates of training/testing splits with a ratio of 2:1. The results of RF and GRRF-RF are slightly different from the results of Deng and Runger (2013) due to randomness.

Table 1 shows the average error rates of different methods. GRF-RF outperforms RF on 9 data sets, 7 of them have significant differences at the 0.05 level. The advantage of GRF-RF over GRF and GRRF-RF is also clear. GRF-RF also outperform GRF, and therefore applying RF to features selected by GRF is better than GRF as a classifier.

### 4 CONCLUSIONS

The guided random forest (GRF) is proposed here for feature selection, particularly, for gene classification in this work. Experiments show that GRF-RF not only significantly outperforms RF in accuracy performance, but also uses many fewer features in the model. In this work I discuss the advantages of GRF for high-dimensional gene data sets. It may also be valuable to find other cases where GRF has advantages over other methods, with the option of tuning the parameter γ in Equation (2) (fixed as 1 here).

Furthermore, in this work, λc is determined by the importance score of feature Xi, from an ordinary random forest. However, λc can be specified by other ways too, e.g., F-score or human knowledge.

### REFERENCES

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Deng, H. and Runger, G. (2013). Gene selection with guided regularized random forest. *Pattern Recognition*, to appear.

Diaz-Uriarte, R. and De Andres, S. (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1), 3.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3), 18–22.

Riddick, G., Song, H., Ahn, S., Walling, J., Borges-Rivera, D., Zhang, W., and Fine, H. A. (2011). Predicting in vitro drug sensitivity using random forests. *Bioinformatics*, 27(2), 220–224.

Yuan, Y., Xu, Y., Xu, J., Ball, R. L., and Liang, H. (2012). Predicting the lethal phenotype of the knockout mouse by integrating comprehensive genomic data. *Bioinformatics*, 28(9), 1246–1252.