Determining Fern species through morphological data

Nabin Kumar Karki

October 15, 2018

Introduction

Ferns are one of the oldest groups of plants on Earth, with a fossil record dating back to the middle Devonian (383-393 million years ago) (Taylor, Taylor, and Krings, 2009). The diversity of ferns we see today evolved relatively recently in geologic time, many of them in only the last 70 million years.Ferns are the second-most diverse group of vascular plants on Earth, outnumbered only by flowering plants.There is consequently only one anatomical feature that unites them, an inconspicuous trait that requires observing the development of vascular tissue in the stem. Most ferns have rhizomes, underground stems from which the leaves are produced.An entire leaf is called a frond, while further subdivisions are referred to as pinnae (first division), which grow along the main stem (called a rachis in ferns), and pinnules (subsequent divisions). The portion of the rachis without pinnae is referred to as the stipe (petiole), which attaches directly to the rhizome.Ferns reproduce by spores, which are generally produced on the bottom (abaxial side) of leaves by specialized structures called sporangia.

This diversity of fern leads to a failure to distinguish different species of fern. DNA analysis and molecular methods are not conclusive in differentiating very close species as they have limitations. It is important to identify different species for the better understanding of biodiversity and their role in functioning ecosystem. Here, we present Principal Component Analysis(PCA) and Random Forest method to determine the fern species through morphological data. Four species of fern collected from different parts of world are albo,dubia,rufa and grisea.

Random Forest

Random forest(RF) fits many classification trees to a data set, and then combines the predicitions from all the trees. RF trains an ensemble of individuals decision trees based on samples,their class designation and variables. Every tree in the forest is built using the random subset of samples and variables,hence called Random forest.The algorithm begins with the selection of many(e.g., 500) bootstrap samples from the data. In a typical bootstrap sample, approximately 63% of the orginal data set occue at least once.The remaining sample in the original data set are the 'Out-of-bag'(OOB) samples. The tree is grown using the bootstrap data set by reversive partitining. For every tree node, only a small number of randomly selected variables(eg., the square root of the number of variables) are available and evaluated for their ability to split the data. The variable resulting in the largest decrease in impurity is chosen to separate the sample at each 'parent node', starting at the top node, into two subsets, ending up in two distinct 'child nodes'. In RF, the impurity measure is the Gini impurity. A decrease in Gini Impurity is related to an increase in the amountof order in the sample classes introduced by split in the decision tree. After the bootstrap data has been split at the top node, the splitting process is repeated. The partitioning is finished when the final nodes, 'terminal nodes' are either (i) 'pure, i.e, they contain only sample belonging to the same class or (ii) contain a specific number of samples. A classification tree is usually grown until the terminal nodes are pure, even if that results in terminal nodes containing a single sample. The tree is thus grown to its largest extent; it is not 'pruned'. After a forest has been fully grown,the training process is completed.The random forest model can subsequently be used to predict the class of new sample. Every classification tree in the forest casts an unweighted vote for the sample after whic majority vote determine the class of the sample.

Accuricies and error rates are computed for each observation using the out-of -bag predictions, and then averaged over all observations. Because the out-of-bag observation were not used in the fitting of the trees, the out-of-bag estimates are essentially cross-validated accuracy estimates. Probabilities of membership in the different classes are estimated by the proportions of out-of-bag predictions in each class.

Application of RF to fern

Dr. Hopper provides potential four fern species namely grisea, albo,dubia and rufa from different parts of worlds.These species cover broad range of morphological characteristics.We are using Random Forest to classify fern species compared to other statistical classifiers includes (1) very high classification accuracy ; (2) a novel method of determining variable importance ; (3) ability to model complex interaction among predictor variables; (4) flexibility to perform several types of statistical data analysis , including classification, regression, survival analysis and unsupervised learning ; (5) an algorithm for imputing missing values. (Might not actually use the algorithm inside randomForest to imputing missing value, because we are using RandomForest to classify our data).

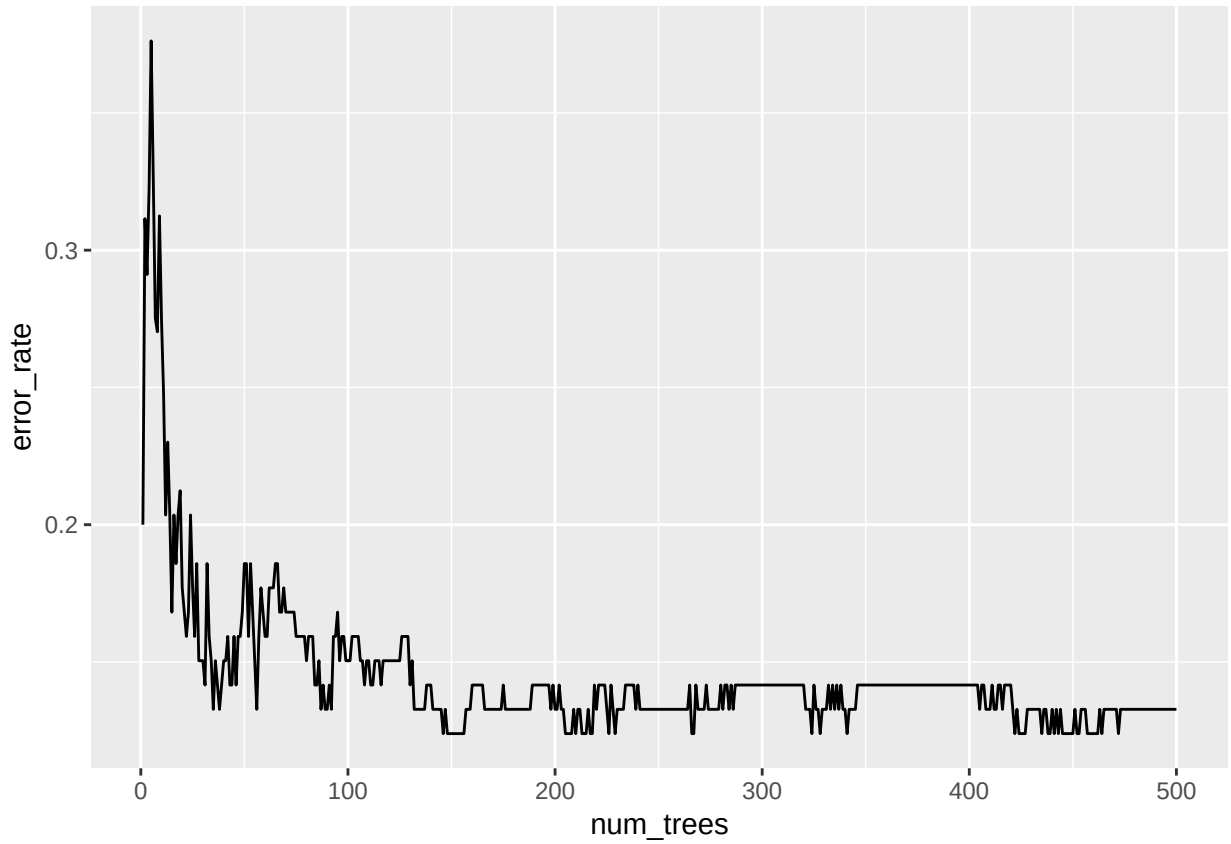
RF Analysis

After all the data cleaning and imputation is done. We conducted our RF analysis in R using randomForest. At the inital step, we used the default setting of the algorithm and run full model.Depending on the variable importance output from the full model we try to include only top 13,8,4 variable and run the reduced RF model respectively.

Full RF Model

so, from the full model we know which variables are important to our model. The full model is complex so we want to build our reduced model containing top 13, 8 and 4 variables and compare our models in better classification and prediction.

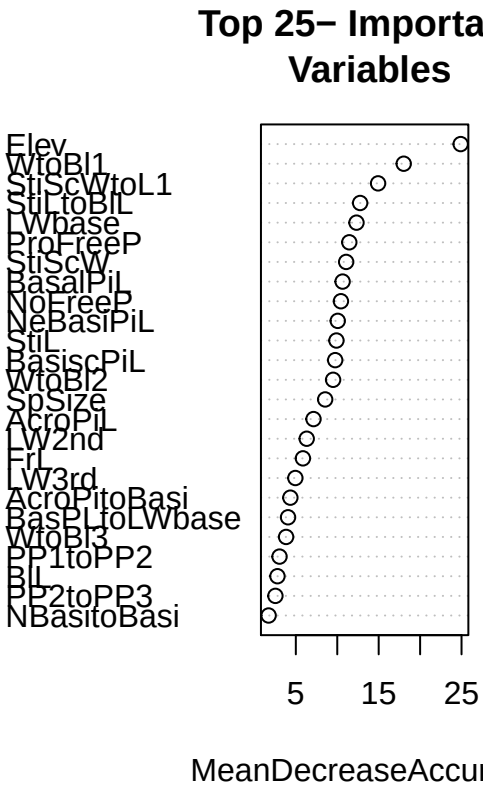
we can also check the error rate of random forest.The out-of-bag estimate of error is the error rate for the trained models,applied to the data left out of the training set for that tree.Using the output from the model,the OOB error can be plotted verses the number of trees in RF. The graph shows the error rate rapidely decreases from over 0.35 before stabilizing around 0.15.

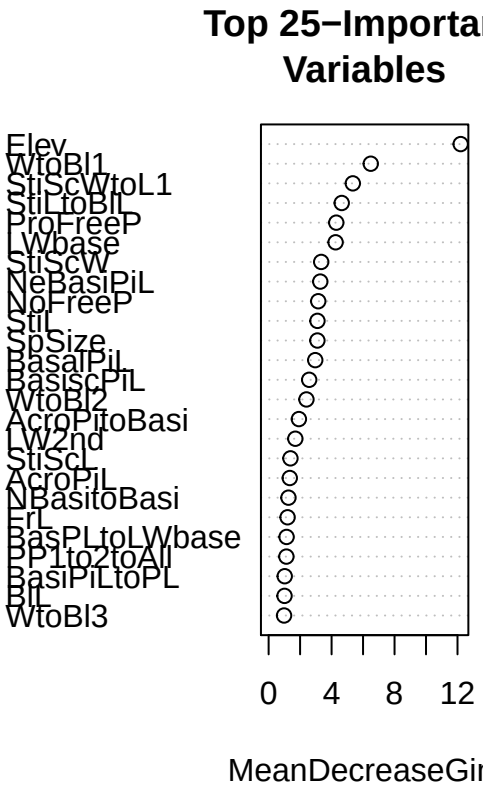


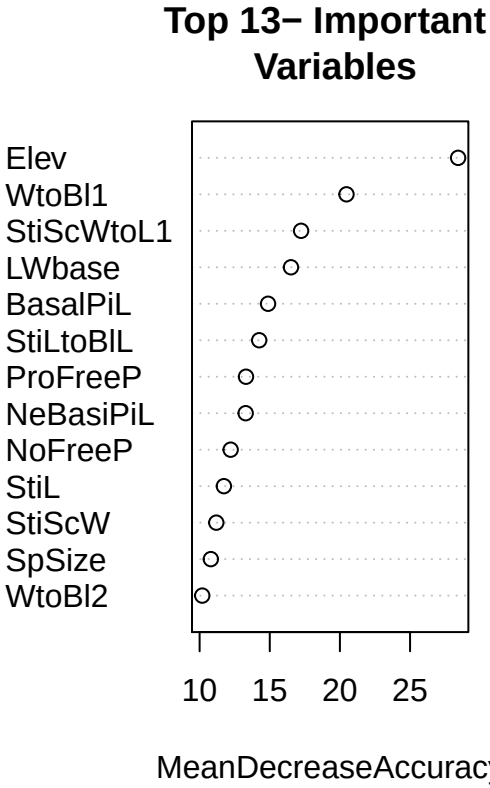
Variable importance

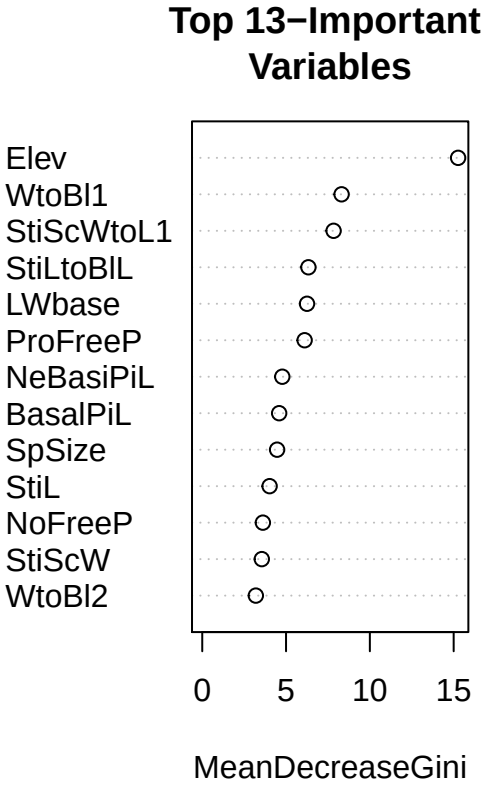
The power of the random forest algorithm shows itself when you build predictive models for data with many features and records. It has the ability to automatically determine which predictors are important and discover complex relationships between predictors corresponding to interaction terms.Importance estimates can be very useful to interpret the relevance of variables for the data set under study. Two frequently used types of the RF variable importance measure exist. The mean decrease in classification is based on permutation. For each tree, the classification accuracy of the OOB samples is determined both with and without random permutation of the values of the variable.The prediction accuracy after permutation is subtracted from the prediction accuracy before permutation and averaged over all trees in the forest to give the permutation importance value. The second importance measure is the Gini importance of variable and is calculated as the sum of the Gini impurity decrease of every node in the forest for which that variable was used for splitting.This measures how much improvement to the purity of the nodes that variable contributes.

After running the model we have graphed the top 25 and 13 variables based on its accuracy and gini index. It seems like the top 25 and 13 variables are similar but not every variable are in the same order. It might be due to the correlation between the predecting variables.


Top 25− Important Variables
Elev
WtoBl1
StiScWtoL1
StiLtoBlL
LWbase
ProFreeP
StiScW
BasalPiL
NoFreeP
NeBasiPiL
StiL
BasiscPiL
WtoBl2
SpSize
AcroPiL
LW2nd
FrL
LW3rd
AcroPitoBasi
BasPLtoLWbase
WtoBl3
PP1toPP2
BlL
PP2toPP3
NBasitoBasi
5
15
25
MeanDecreaseAccuracy


Top 25−Important Variables
Elev
WtoBl1
StiScWtoL1
StiLtoBlL
ProFreeP
LWbase
StiScW
NeBasiPiL
NoFreeP
StiL
SpSize
BasalPiL
BasiscPiL
WtoBl2
AcroPitoBasi
LW2nd
StiScL
AcroPiL
NBasitoBasi
FrL
BasPLtoLWbase
PP1to2toAll
BasiPiLtoPL
BlL
WtoBl3
0
4
8
12
MeanDecreaseGini


Top 13− Important Variables
Elev
WtoBl1
StiScWtoL1
LWbase
BasalPiL
StiLtoBlL
ProFreeP
NeBasiPiL
NoFreeP
StiL
StiScW
SpSize
WtoBl2
10
15
20
25
MeanDecreaseAccuracy


Top 13−Important Variables
Elev
WtoBl1
StiScWtoL1
StiLtoBlL
LWbase
ProFreeP
NeBasiPiL
BasalPiL
SpSize
StiL
NoFreeP
StiScW
WtoBl2
0
5
10
15
MeanDecreaseGini

Results

Table: identify matrix summaries generated from random forest analysis of morphological traits

species	N Specimens	30 Traits %misclassified	13 Traits %misclassified	8 Traits % misclassified	4 Traits %misclassified
albo	41	0.073	0.122	0.146	0.171
dubia	12	0.583	0.417	0.250	0.250
grisea	40	0.075	0.025	0.050	0.225
rufa	20	0.100	0.100	0.100	0.150
Accuracy(%)		86.73	89.5	89.5	82.3

After running the full model and reduced model. We compare the accuracy of all four model. The output shows that the reduced model with 13 and 8 prediction variable have better prediction compared to other two model.These two model have the same accuracy of 89.5 % where as the full model has the error rate of 13.27% and four variable model has error rate of 19.47%.Even thought the model with 13 and 8 variable have the same accuracy but there is a difference in classification.

Confusion matrix of RFM13 show the model correctly classify 36/41 albo and misclassify 1 albo with dubia and 4 with grisea.For dubia,model correctly classify 7/12 and misclassify 3 with albo and 2 with rufa. For grisea, model correctly classify 39/40 and 1 misclassified with albo.For, rufa, model correctly calssify 18/20 and 2 misclassified with rufa.

Confusion matrix of RFM8 show the model correctly classify 35/41 albo and misclassify 1 albo with dubia and 5 with grisea.For dubia,model correctly classify 9/12 and misclassify 3 with albo . For grisea, model correctly classify 38/40 and 2 misclassified with albo.For, rufa, model correctly calssify 18/20 and 1 misclassified with rufa and other with dubia.

We see the largest error in classifying the species dubia which makes sense as it might be hybrid between the albo and rufa fern species. Otherwise RFM has done good job of classifying the fern species based on the morphological traits.

References

Moffat, C. E., et al. "Morphology Delimits More Species than Molecular Genetic Clusters of Invasive Pilosella." American Journal of Botany, vol. 102, no. 7, 2015, pp. 1145–1159., doi:10.3732/ajb.1400466.

Cutler, D. Richard, et al. "Random Forests For Classification In Ecology." Ecology, vol. 88, no. 11, 2007, pp. 2783–2792., doi:10.1890/07-0539.1.

Touw, Wouter G., et al. "Data Mining in the Life Sciences with Random Forest: A Walk in the Park or Lost in the Jungle?." Briefings in Bioinformatics, vol. 14, no. 3, May 2013, pp. 315-326. EBSCOhost, search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=87826382&site=ehost-live.

About ferns. (n.d.). Retrieved from https://www.amerfernsoc.org/about-ferns/