Report

Zexuan Zhao

2022-12-11

Contents

1 Background

Species diversity is a key component of bio-diversity. Understanding why there are numerous species on the Earth requires to understand speciation, that is the process by which a single species of organism splits into two distinct species. This process has occurred over the history of Earth and is driven by various mechanisms such as geographical isolation, reproductive isolation, hybridization, and natural selection. Two major compatible theories about the drivers of speciation have gained popularity: isolation-by-distance (IBD) model predicts that geographical distance can create difficulties for individuals of the same species in exchanging genetic materials (e.g. mating) and thus genetic distance should be positively correlated with geographical distance; isolation-by-environment(IBE) model treats differences in environmental factors more seriously than barely geographical distances by emphasizing the role of local adaptation of individuals living in a different environment than its original habitat, and thus genetic distance should be more strongly correlated with environmental distance.

IBD and IBE are both valuable in the theory of speciation, but they inform little to conservation. Governments and organizations can only perform conservation given important factors that drives speciation on a board scale, and protecting such factors would help to preserve bio-diversity. In this project, I try to address the question of what factors better explain genetic distances among various species using machine learning. This work is built based on the work of Pelletier and Carstens. In their paper, they first did matrix correlation tests of genetic distances versus geographical distances(IBD) and environmental distances(IBE) in 8955 species and then used random forest to uncover important predictors that contribute to the accuracy

in predicting IBD and/or IBE. Here, I will first replicate their random forest analysis using their labeled data set generated by statistical tests and then try to use neural network as a substitute of the random forest. I will also contribute some ideas to the few miniature oversights in their comprehensive and scrupulous work.

2 Material and methods

2.1 Data retrieval and preprocessing

Scripts and data tables were retrieved from here on Dec 1 2022. Scripts are stored in the directory `scripts_from_original_paper` and the data table is saved as `Rscripts/AppendixS1.csv`. Unfortunately the data set is corrupted by EXCEL as `#NAME?` appears 444 times in the file. Therefore, I am not able to replicate their results precisely numerically, but the conclusions they drew should be relatively robust to deleting these corrupted rows.

The newest script `scripts_from_original_paper/7_Random_forest/random_forest_NEW_3.r` indicated by the last modified data is used for reference. For replicating random forest analysis, the data table is pre-processed the same way as in their scripts. Redundant variables, such as statistics generated by Mantel tests and low-level taxonomies, are removed, and p-values `pmtp` are transformed to `Yes/No` label by a threshold of 0.05. No correction is done. `n` is the number of individuals whose genetic sequences were used to infer genetic structure. Low `n` suggests the inference of genetic structure is not reliable. In their paper, it's shown that `n>=20` guarantees reliable results, which is also implemented in my scripts. The labeled data set is highly unbalanced because of the intrinsic property of p-values, a balanced data set is generated by sampling from positive and negative cases with equal number.

I further excluded all taxonomic variables in the data set for my neural network because keeping just a few board taxonomic labels does not reflect the phylogenetic relationships anyway. `metabolism`and `habit` are hot encoded as they are string type. I applied an empirical threshold of `n > 5` because the distribution of p-value is converging to stability when `n > 5` (Figure 2), and excluded `n` for further analysis, because `n` contributes nothing to real biological significance. Applying a threshold of statistical significance is at risk of losing information and introducing human bias because the threshold is mainly an artificial number. To fully unleash the potential of neural network, I keep the p-values as response variables to indicate the level of IBD and IBE. P-values for both IBD and IBE are included, instead of only IBD, which is what the original script did. Therefore, the dimension of response variable is 2.

2.2 Random forest

Random forest uses decision trees to predict the response label by predictors. Each tree is grown on a bootstrapped data set from the input, which only contains 150 data points with 5 predictors. Importance of each predictor is evaluated by a permutation approach. The value of a predictor is permuted, and the more important a predictor is, the more mean accuracies will decrease due to permutation. A confusion matrix is also reported, which is not included in the original paper. Accuracies of random forests trained on original data, unbalanced data filtered by `n>=20`, unfiltered balanced data and filtered balanced data are compared. As in the original paper, I used `randomForest` package in `R` for random forest and importance analysis.

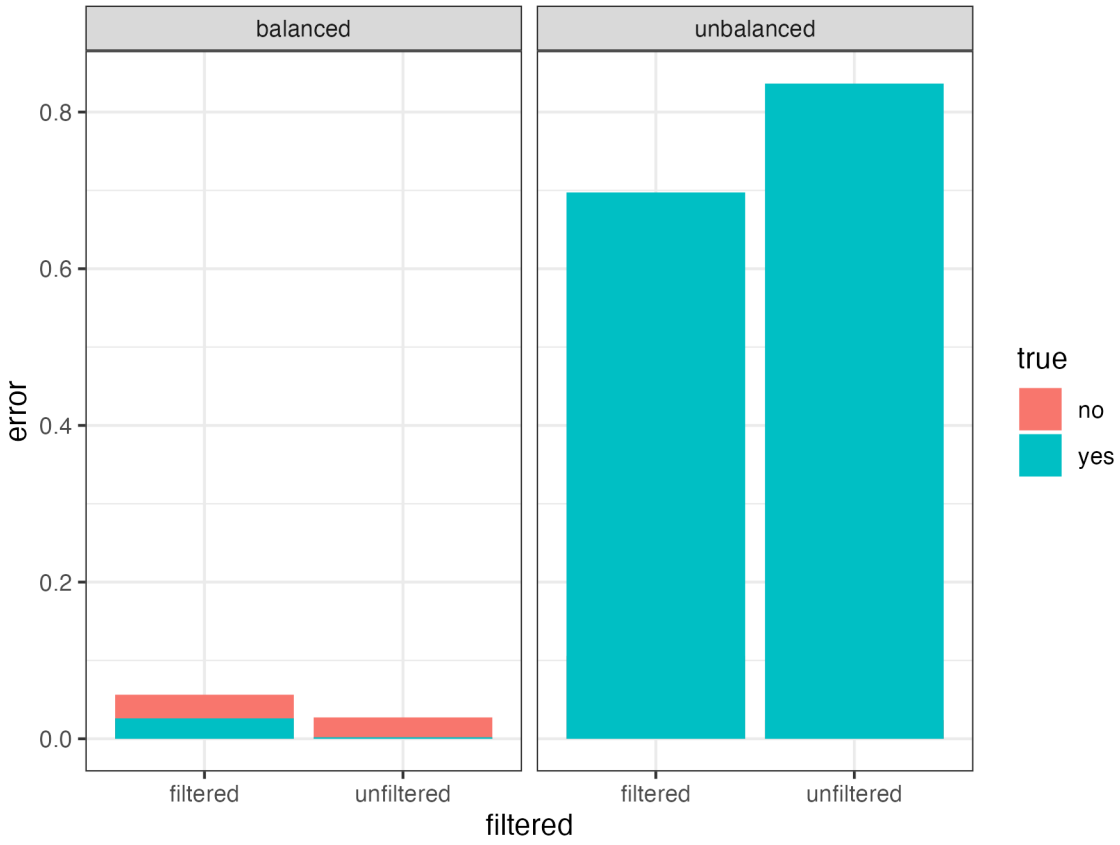

Figure 1: Erro rates of random forests built on filtered/unfiltered balanced/unbalanced data sets. Filtering is to reduce noice in genetic structure inference, and balancing positive/negative training cases is to reduce bias of training accuracy. Colors denote true labels, where red means IBD or IBE and blue means no genetic structure.

2.3 Neural network

3 Result and Discussion

3.1 Random forest

After removing redundant variables, the original data set has 41 predictors, 1 response variable and 12252 data points. However, the proportion of positives is low (14.9%), and therefore training on original data set will likely be biased towards minimizing false positives and render low sensitivity. Besides, `n` as the number of individuals of a species used to infer genetic structure will also affect the reliability of labels (training noise), and noisy data would also create difficulty in training random forests.

Here I applied filtering and balancing techniques to the training set, and compared their accuracy based on true labels. Despite decent overall accuracy (85.5%), random forests perform poorly on predicting positive on true positives, if training data set is unbalanced (Figure 1). After resampling the original data set to generate new data set with equal number of positives and negatives, the overall error rate is reduced and the bias towards false negative is eliminated.

Filtering has opposite effect on balanced and unbalanced data set. In unbalanced data set, after apply filtering `n` the error rate is reduced, mainly on false negatives, while in balanced data set, filtering increased false negative rates, although false negative rates are relatively similar to false positive rates, which means the bias is also reduced. Filtering can reduce noise in the original data set, but it also reduce the amount of data. Therefore, the net effect of filtering is dependent on the trade off between the number of data we have and the noise we want to reduce. It seems reasonable to apply a more soft filtering criteria to keep more data, which is done in the next section.

Table 1: Mean decrease accuracy of predictors.

type	meanDecreaseAccuracy
elevation_sd	83.12273
n	71.94724
abs_mid_lat	70.34212
elevation_mean	65.49152
length	64.48341
area	59.66498

A similar result, though not precisely the same as in the original paper, of important variables is generated from data set that is balanced and filtered by `n>=20`. Except for `n`, all of the variables are geological factors (Table 1), which supports IBD. Although I have applied filtering to reduce the noise introduced by the number of individuals used to infer genetic distance, it's still an important variable as it ranked second from the list, which indicates the intrinsic noise in the methodology. As the genetic sequences were not originally generated for the propose of this study, `n` is out of control in this database-mining study. Further attention should be paid to controlling confounding factors related to experiment design.

4 Code availability

R markdown is used to prepare data, run random forest and knit the report, and Jupyter notebook is used for neural network. For codes of all sections:

- Random forest: `Rscripts/random_forest.Rmd` and its pdf output `Rscripts/random_forest.pdf`
- Data preparation for neural network: `Rscripts/data_preparation_NN.Rmd` and its pdf output `Rscripts/data_preparation_NN.pdf`

- Neural network:
- Report: `report/report.Rmd` and pdf report `report/report.pdf`

5 Appendix

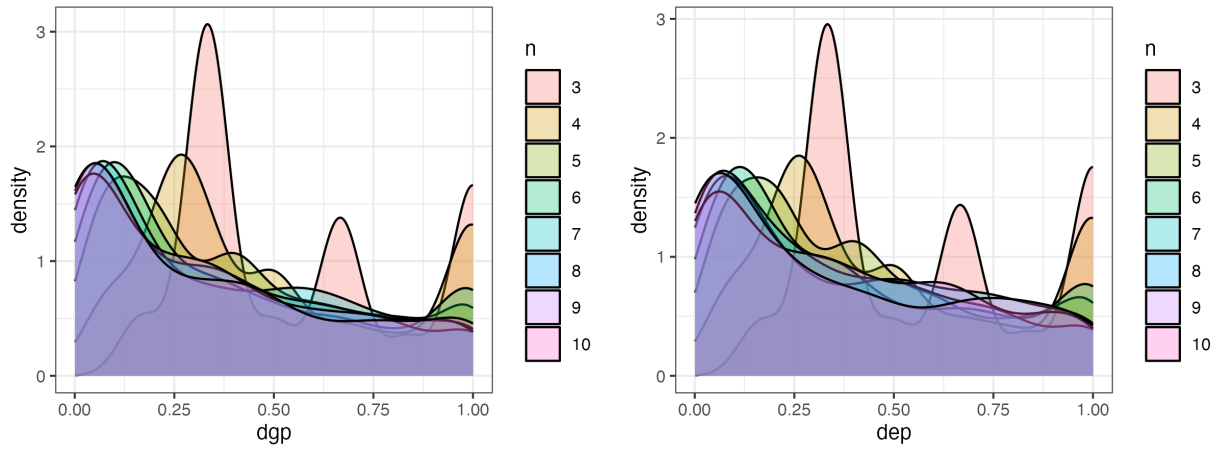

Figure 2: Histogram of p-values of IBD and IBE by different n