Predicting bird classifcation by bone lenghts

Jouke Profijt

October 8, 2018

Classification of differing ecological bird populations

Introduction

Materials & Methods

During this research we used data from an external source, used weka 3.8.0 for classification and R version 3.5.1 in combination with Rstudio for data analysis.

Data

Birds' Bones and Living Habits, Kaggle dataset

Bone measurements were measured from a skeleton collection of Natural History Museum of Los Angeles County, provided by Dr. D. Liu of beijing Museaum of Natural History

R and Rstudio

R was used for the exploration of the data, cleaning of the data and statistical analysis. In our EDA(Exploratory Data Analysis) we looked at the structure of the data and the distribution of hte data using basic graphs with the ggplot2 library.

Cleaning was done based on missing values which we all removed & on outliers which we also removed,

And lastly we compared different classification algorithms again using plots from ggplot2 to choose the correct algorithm.

weka 3.8.0

Weka (Waikato Environment for Knowledge Analysis) is a free to use datamining software written in Java. It is an java application that is capable of doing lots of things apliccable to datamining. We will be using it to determine what is the best classification algorithm for our usecase. As we will be creating a java application of our own that can be used to classify the different bird groups.

- First we use the explorer to make our datafiles usable in weka by removing duplicate id collums
- when the data is ready we can use the classify module in weka to test diffrent classification algorithm's that have reasonable accuracy's.
- at last when we found some interesting algorithm's we can use the experimenter to try diffrent settings to find optimal ones.

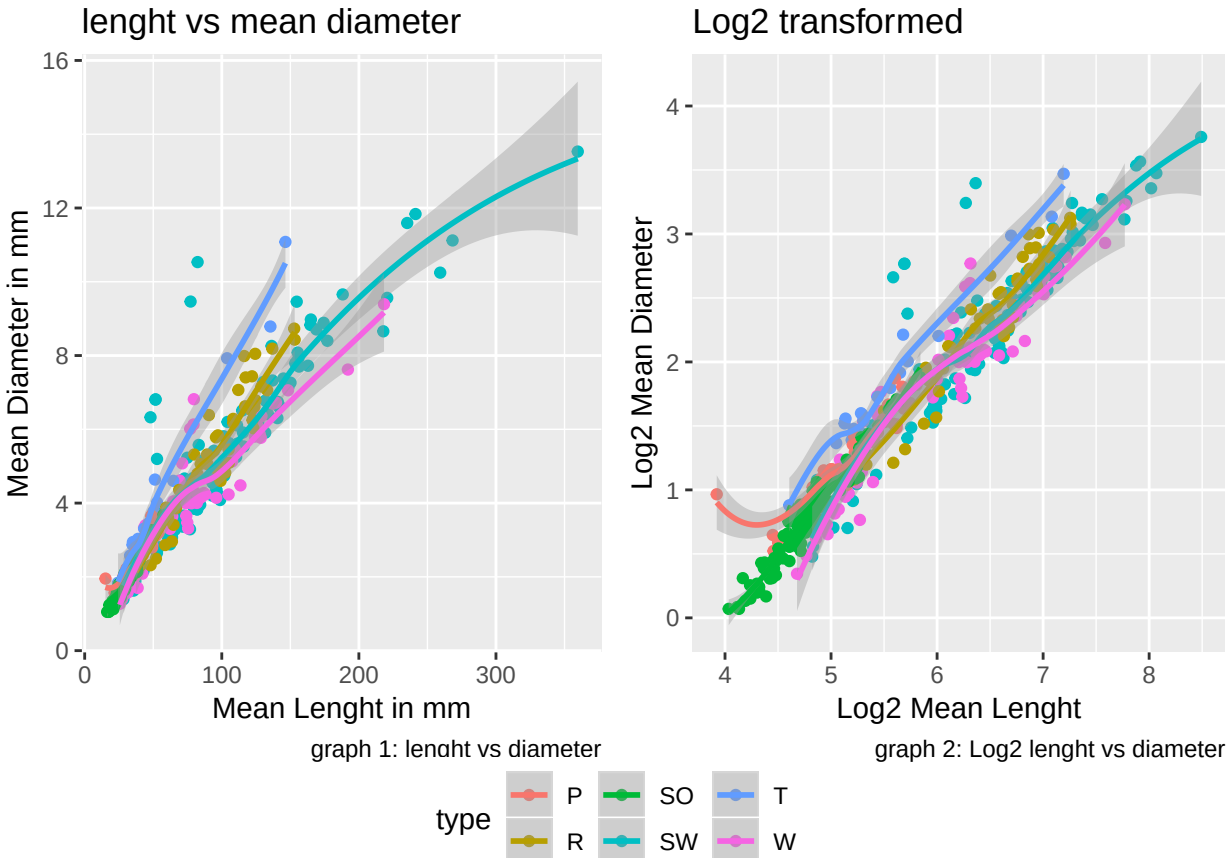
Every algorithm their initial output is a confusion matrix. This matrix is the measurement that we use to compare different algorithms. they are also inspected using a learning & ROC curve's.

- A Learning curve shows the false positive rate over how much data is provided to the learning algorithm
- A ROC curve shows the Sensitivity over the Specificity.

Resutls

To begin researching what classification algorithm is the best for our data we first need to explore the data. we want to know what the distribution of data is, if we need to remove datapoints because they are missing or misleading, so removing NA's & outliers. To find out what algorithms to use we can compare their performance by looking at accuracym ROC and their learning curve.

Exploratory Data Analysis



graph 1: lenght vs diameter

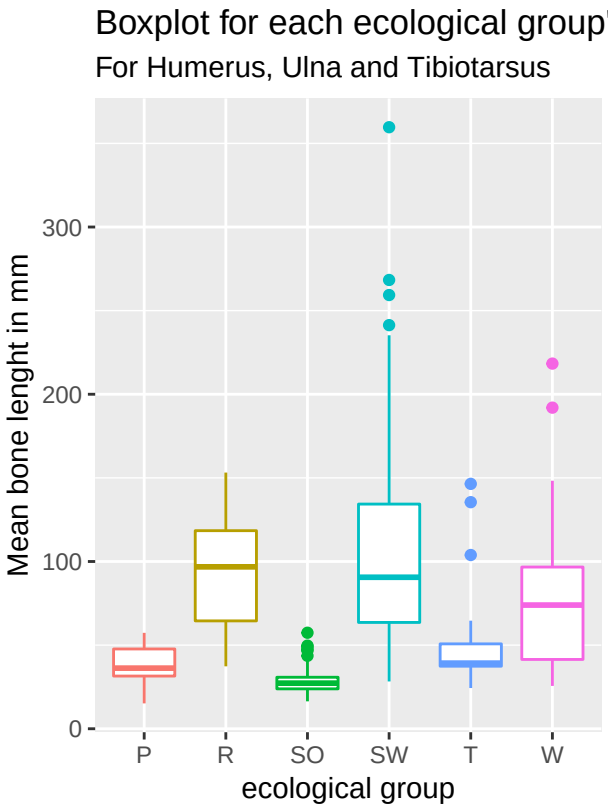
graph 2: Log2 lenght vs diameter

Graph 1 & 2 show the bone length & diameter distribution. As expected the lenght and diameter are strongly correlated (longer bones alse need to be thicker to keep their strenght).

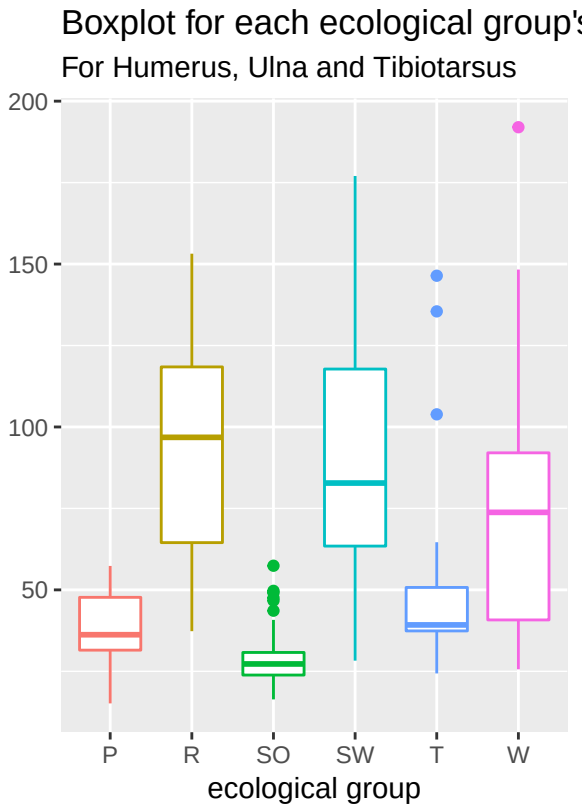
Graph 1 is the distribution without any changes. Here we see that the Swimming birds have quite large bones, which could be because of abnormally big bird samples. as you look closer at the swimming birds in graph 1 and 2 we see that there might be a lot of noise in these samples because there are a lot of points outside the main distribution. For the other groups we also see that the Scansorial Birds & the Sininging birds are the smaller groups.

Terrestrial Birds, Raptors and Wading Birds seem to be in the middle of the rest with the Terrestrial Birds having notisable thicker bones that the others.

Data Cleaning



graph 3: mean bone lenghts



graph 4: Cleaned data

For our cleaning of data we removed all rows with NA's which cost us 7 datapoints. After that we calculated the 1st and 3rd quartile outliers from the humurus which can be seen in graph 3 & 4, and removed 9 rows.

- 1st quartile(Q1): 25.17 mm
- 3rd quartile(Q3): 90.31 mm

$Smallthreshold = Q1 - 1.5 * (Q3 - Q1)$

$Largethreshold = Q3 + 1.5 * (Q3 - Q1)$

Table 1: Removed datapoints

NA's	Outliers
7	9

After all cleaning we left with 404 datapoints to use for Weka Analyses and classification.

Weka (classification)

In weka the goal was to find a classification algorithm that has the highest possible accuracy because classifying a birds herritage it does not matter that much if there is a False positive, the goal was to keep false negatives low and keep true positives high.

Table 2: Random.Forest performance

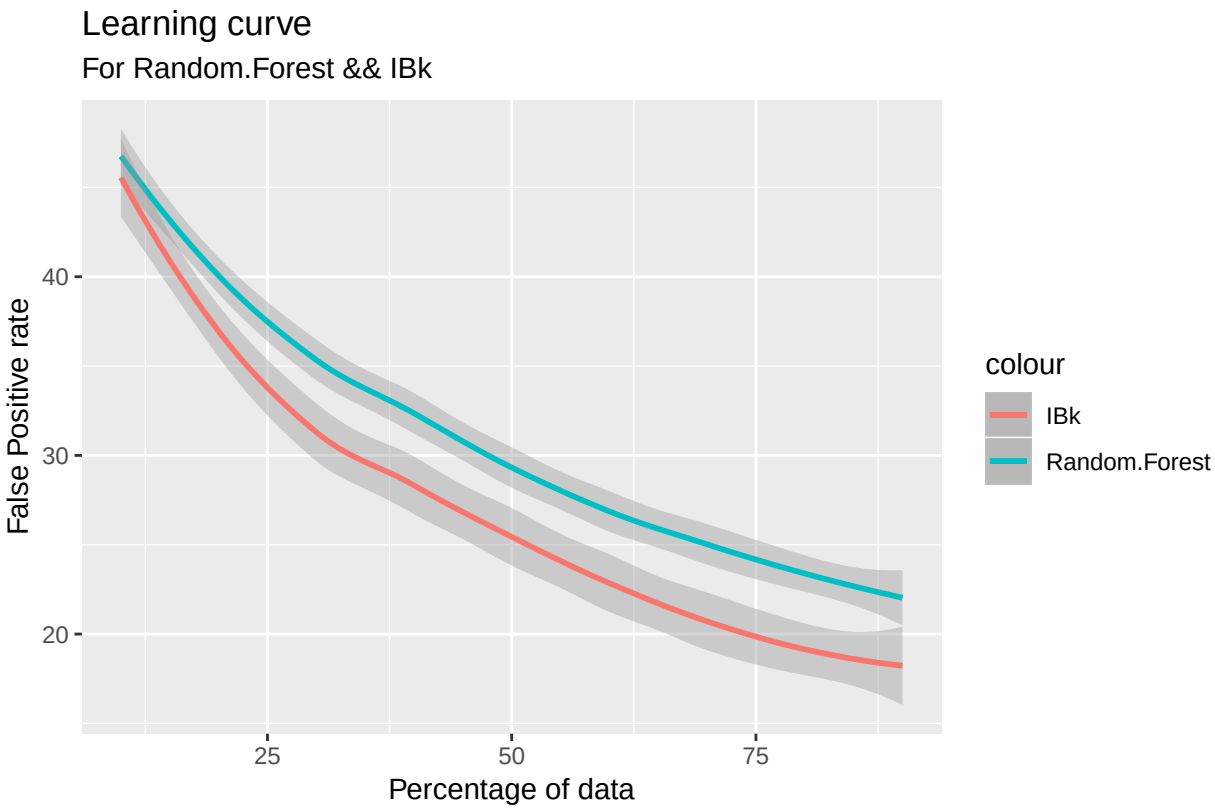
	Correct	Incorrect
Instances	342	62
Percentage	84.6535 %	15.3465 %

Table 3: Ibk performance

	Correct	Incorrect
Instances	371	33
Percentage	91.8317 %	8.1683 %

Graph 5: ROC

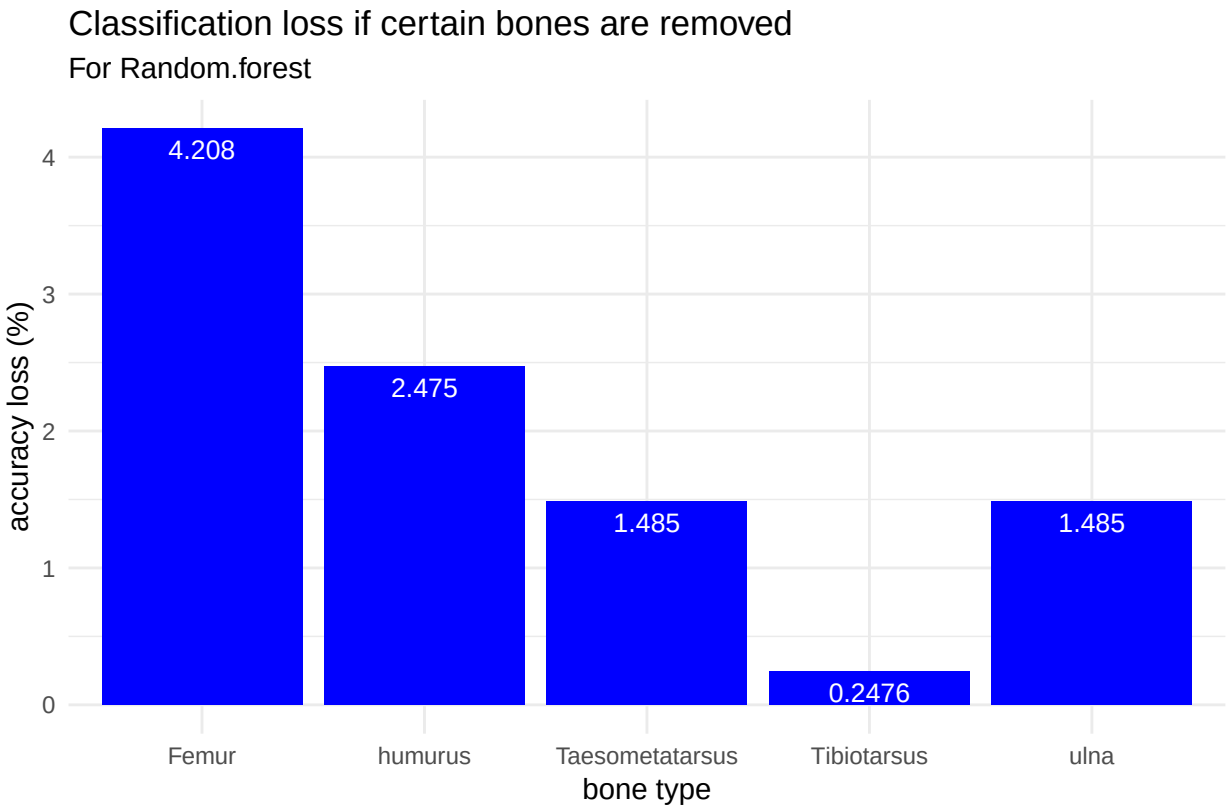
In graph 5 we can see the ROC comparison between IBk and Random.Forest. As the two algorithm's come closer to the upper left we see that Ibk cuts off quit a bit earlier than Random.Forest. this could give an indication that Random.forest might have a higher overall accuracy. But in the IBk algorithm we got fewer datapoint to work with.



graph 6: Learning curve inspected algorithm's

Graph 6 shows the learning curve for both algorithms. As we reduce the amount of data fed to the learning algorithms we can see that both lose about the same amount of accuracy. But as shown in table 2 & 3 the overall accuracy of ibk is about 8% better and that is shown here again.

Ibk is also a small amount better in classification when less data is provided.



graph 7: accuracy loss per bone

in graph 7 we can see what the importance is from certain bones if we are talking about classification. the differences we can see are because of their importance for the functionallity of the bird groups.

Conclusion & discussion

What is the most important bone for each ecological group their function? we can conclude that (At least for classification) the Femur is the most important as when we remove this from our classification algorithems the accuracy loss is great. we can see this in graph 5 where the loss per bone is displayed. while researching the subject we made the assumption that the Femeur and Taesometatarsus were the least important for classification. Becouse of this the first classification algorithems were done only using the longer bones but further in reverted this desicion. The classification algorithem that was chosen however wasn't changed. In the future we should first collect some data about the importance for some bones before making such rational desisions.

We do wan to think about removing some bones as when we have less data to be put in the more unknown fossils with missing bones could possibly be classified

Minor Proposal

References