Predicting bird classifcation by bone lenghts

Jouke Profijt

October 8, 2018

Classification of differing ecological bird populations

Introduction

Are there bones in a bird that are more important for classification using machine learning? Or should we take all avalable information into account if we are creating a classificiaton algorithm using machine learning. And then what algorithm should we use?

In this research project we will be going through data of varying ecological bird groups. in this data the bone lengths and diameters are provided of an bird with a label to which ecological group they belong.

Using this data we will be preparing this data for classification, analysing results of different classification algorithms and see how to use the data for classification.

We will be looking into Random.Forest and IBk asl classification algorithms. Random.Forest is tree / decision based and IBk is a K-nearest neigbour classifier(lazy). As IBk is lazy we have to be cautious becouse there might form a bias towards groups with bigger data values.

After we have decided what classfication algorithm to use we can use this research to create a java wrapper application to make use of this classifier in the future if we want to study an unknown fossil ans see what the possible ecological goup is.

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

[2]

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

[1]R is a programming language used mostly in statistics and data analysis. developed by Ross Ihaka en Robert Gentleman.

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[3] (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.

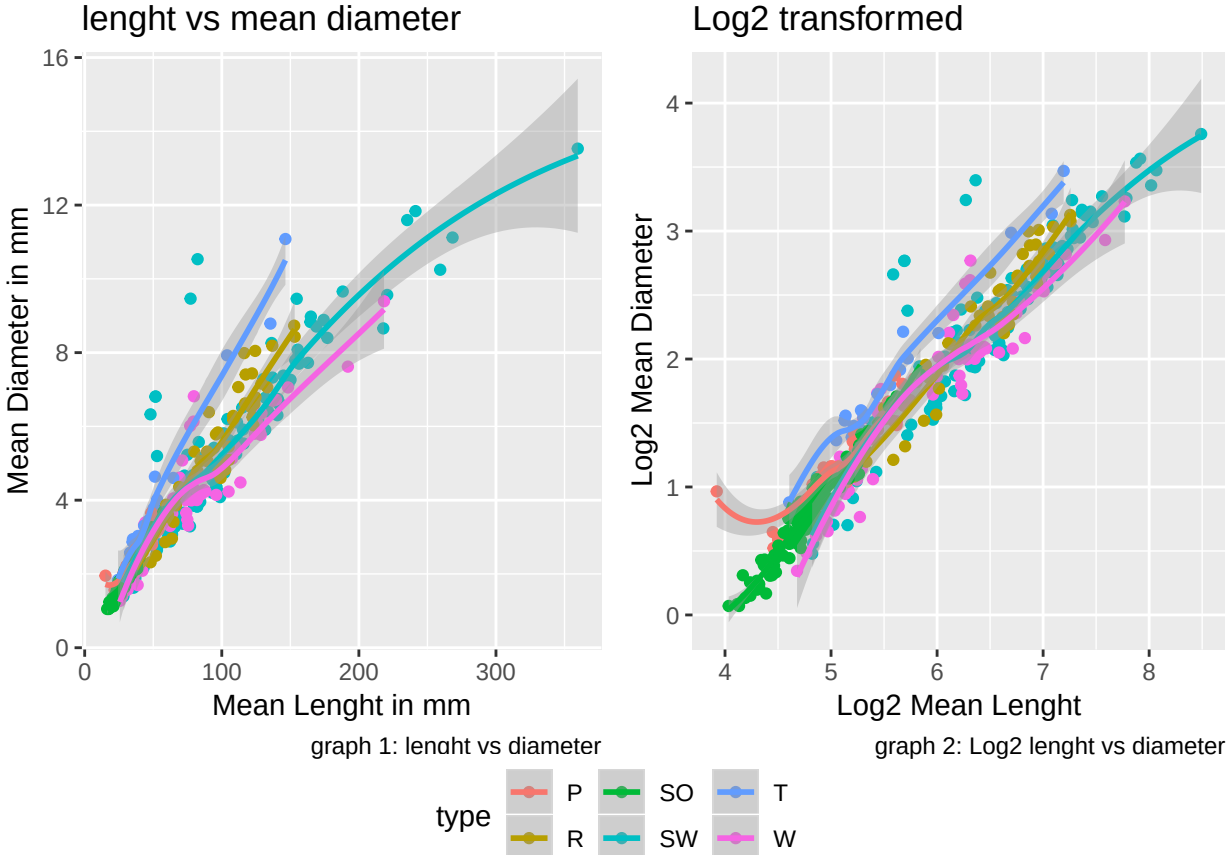
Exploratory Data Analysis

The data contains 420 bird samples where the bone lengths and diameters have been measured. The birds are separated in 6 diffrent groups:

- Swimming Birds, SW
- Wading Birds, W
- Terrestrial Birds, T
- Raptors, R
- Scansorial Birds, P
- Singing Birds, SO

Most samples have data for:

- Length and Diameter of the Humerus
- Length and Diameter of the Ulna
- Length and Diameter of the Femur
- Length and Diameter of the Tibiotarsus
- Length and Diameter of the Taesometatarsus



graph 1: lenght vs diameter

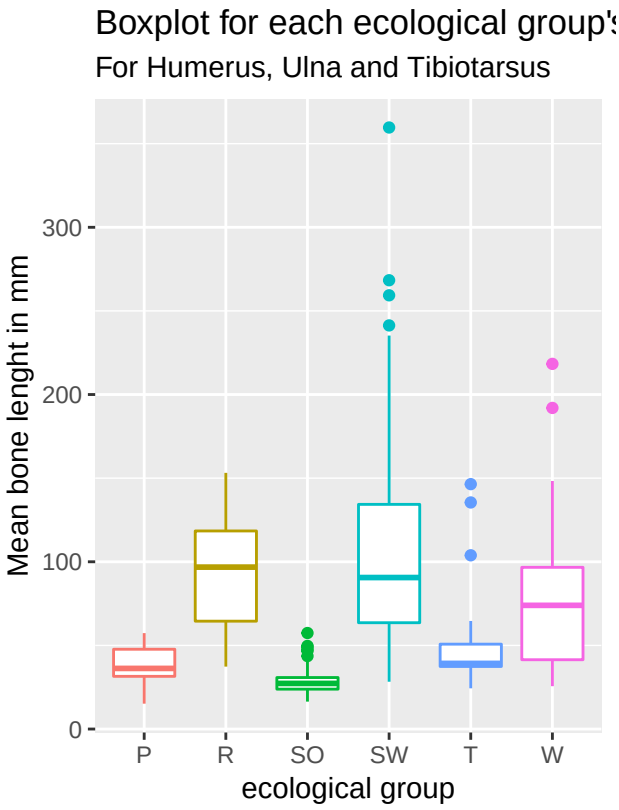
graph 2: Log2 lenght vs diameter

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).

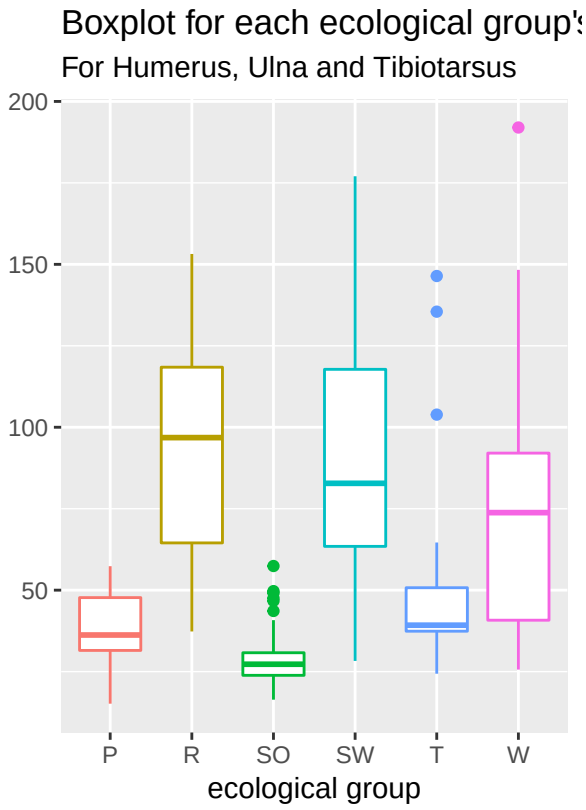
Graph 1 is the distribution without any changes. Here we see that the Swimming birds have quite large bones, which could be because of abnormally big bird samples. as you look closer at the swimming birds in graph 1 and 2 we see that there might be a lot of noise in these samples because there are a 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

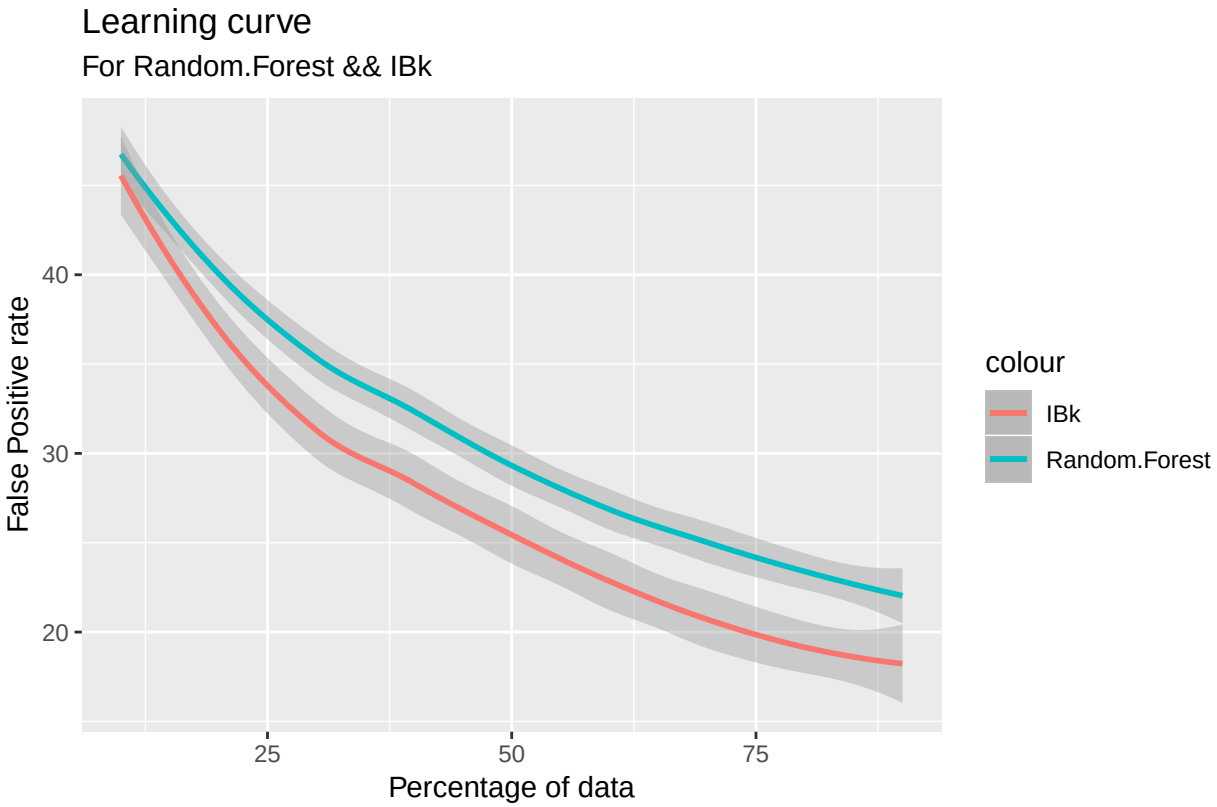
	Correct	Incorrect
Instances	342	62
Percentage	84.6535 %	15.3465 %

Table 3: Ibk performance

	Correct	Incorrect
Instances	371	33
Percentage	91.8317 %	8.1683 %

Graph 5: ROC

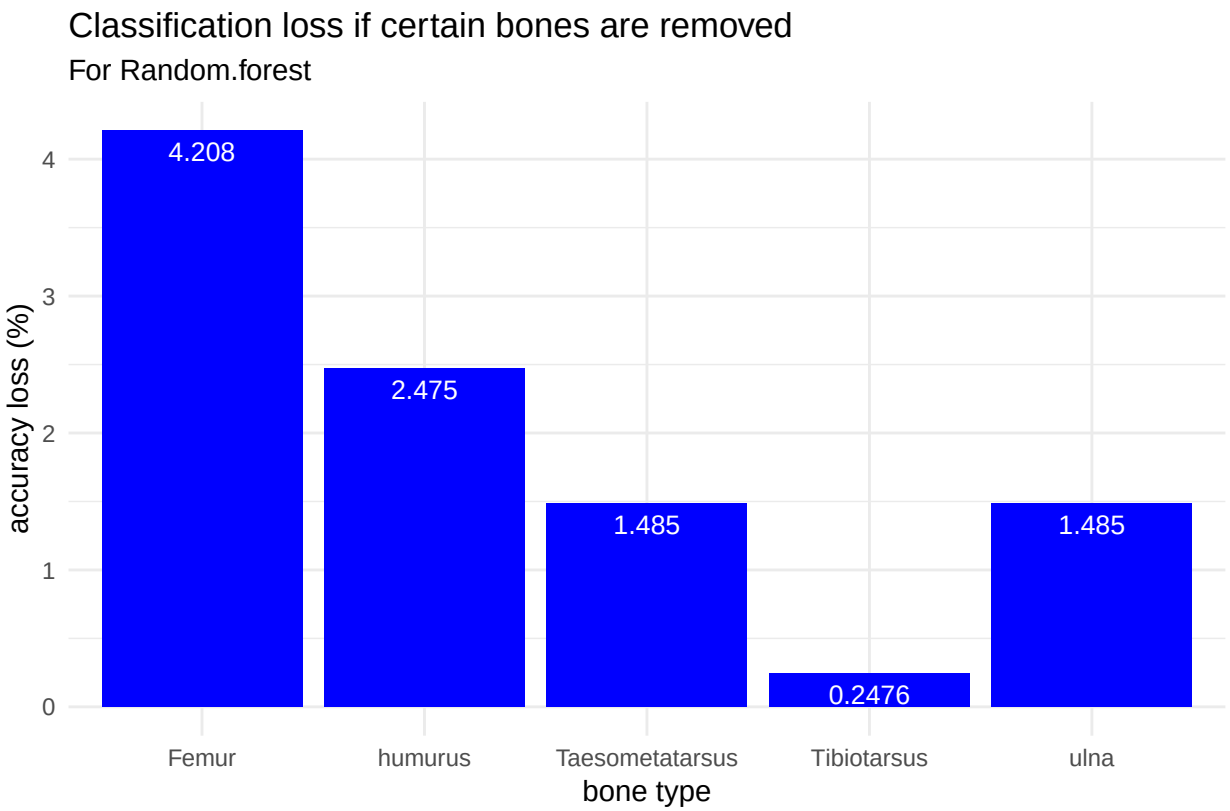
In graph 5 we can see the ROC comparison between IBk and Random.Forest. As the two algorithm's come closer to the upper left we see that Ibk cuts off quit a bit earlier than Random.Forest. this could give an indication that Random.forest might have a higher overall accuracy. But in the IBk algorithm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

Are there bones in a bird that are more important for classification using machine learning? Or should we take all avalable information into account if we are creating a classificiaton algorithm using machine learning. And then what algorithm should we use?

To answer part one of the question we can take a look at graph 7 here all different kinds of bones were listed and removed from classification one by one, and the results show that the Tibiotarsus collums in the data are the least important in classificaton. What is suprising is that the femeur which was the smallest from all inspected bones, has the highest classification accuracy loss.

So if bones are missing in for example the fossil you're investigating the least important one is the Tibiotarsus but it is always preferred to have the most data availalbe.

For the algorithm to use we chose Random.Forest. Altough IBk gives better results the data is skewed and not all groups have the same amount of datapoints. which gave us reason to believe that the data was overfitted to the trainings data.

If we were to redo this classification experiment we would strongly recommend normalizing the data, because as said in our conclusion the data is skewed towards the Sininging and Swimming Birds. use a method that equalizes the amount of datapoints per ecological group so that there is no more data for one type than the oters.

Further we only removed outliers based on the Humerus. it might be worth looking at all available bones for results that differ too much. Also debate on removing datapoints with NA values as it might still be possible to classifiy if only one bone is missing as demonstrated in graph 7. If there is too much missing then yes, remove that instance but only one might not be that important.

Minor Proposal

In the minor Application design (which i chose) we will be focussing on creating userfriendly software. For this project i created a java application/wrapper that can classifiy bird bone data to a specific ecological group. If we want archeologists or biologists to use our application we need to make it userfriendly, As they might have no understanding of the commandline.

We could make an GUI application that lets you enter bone lenghts & diameters that you found in the field, Or select a file you have created. Selecting a file could be a problem because atleast in the current stat can only accept one type of file (arff) so we would need to create some sort of convertor or implement other filetypes.

A web application could also be possible this would almost be the same as the GUI but here we could implement a database system where found data could be saved into an database for others to use or create new datasets.

References

1. Dalgaard, Peter - 2002 - Introductory Statistics with R - ISBN 0387954759.
2. user zhangjuefei, kaggle - 2016 - Birds' Bones and Living Habits - dataset
3. Ian H. Witten, Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes en Sally Jo Cunningham - 1999 - pdf
4. Jouke Profijt - 2018 - Research project - (link)github-research
5. Jouke Profijt - 2018 - java application - (link)github-javaApplicaton