Machine learning approaches to the evolution and ecology of leaf shape

Peter Zeng

Supervisor:
Will Pearse[1] Gemma Bramley[2] Alexander Zuntini[2]

Supervisor Affiliation:
[1]Department of Life Sciences(Silwood Park)
[2]Kew Gardens

Supervisor Contact Email:
Will Pearse: will.pearse@imperial.ac.uk
Gemma Bramley: g.bramley@kew.org
Alexander Zuntini: a.zuntini@kew.org

KeyWords: Machine Learning, Phylogeny, Leaf Shape, Quantification, Modelling, Classification

Introduction

Since people discover the potential of machine learning, different kinds of machine learning methods have been applied to scientific fields, which make it feasible to develop new algorithms or helps scientists save their time analyzing the data. Of these, the computer vision, an important part of machine learning, were widely used in the biological study [Jähne and Haußecker, 2000]. The automatic recognition of image based on features like shapes and patterns make it possible for people to use machine learning method to do the classification of images [Nanni et al., 2017]. In this case, the image of plant's leaf can be ideal for the machine learning method to study [Fu et al., 2004], as the leaf surface area, perimeter length and leaf compactness can be differ among different species. When people are studying the phylogeny of the plant, it will be effective to utilize machine learning methods to help people perform classification accurately and efficiently.

In the hope of helping develop new methods for studying the phylogeny of plant species, I will conduct a bioinformatics project based on machine learning methods. I will use the Angiosperm leaf's images, provided by the scientists in Kew Gardens, to quantify the morphology of the species. In this project, I will (1) update and extend an existing pipeline called "stalkless", which is aimed to record the individual leaf features [Pearse et al., 2018] (2) develop a machine learning pipeline, train it with the features of leaf, and perform classification (3) model the evolution of the estimated leaf shape using the phylogeny.

Methods and Materials

1. Images of leaf

The angiosperm leaf images are collected by the scientists in Kew Gardens, which is related to the genera used in the BigTree, a part of the PAFTOL project held by the Kew Gardens [Baker et al., 2021]. The PAFTOL project aims to equip people with the tools to accelerate efforts to document, identify and classify plants and fungi, explore their useful properties, understand their origins and evolution, and predict how species will respond to future environmental change.

2. Update and extend the pipeline

I will update and extend an existing pipeline called "stalkless", a Python and R pipeline to record individual leaf surface area, perimeter length and leaf compactness, which is aimed to extract the morphology features of leaf images, and used for training the machine learning pipeline. The pipeline was used to quantify the leaf shape from herbarium specimens, and now I will try to apply it to the leaf images provided by scientists in Kew Gardens.

3. Develop the machine learning pipeline

With the features provided by the stalkless pipeline, I will try to develop and train a machine learning model to perform the classification of these plant species. After training, the model should be accurate in classifying the plant species based on the features of their leaf shape.

4. Model the evolution

I will be modelling the evolution of the estimated shape using the phylogeny. The phylogeny is provided from the Kew Garden

Anticipated Outcomes

By the end of the project, I expect to (1) update and extend the ”stalkless” pipeline, make it compatible with the plant images provided by Kew Gardens (2) develop a machine learning pipeline that can be used to classify the leaf image based on the morphology features of the leaf (3) model the evolution of the estimated leaf shape using the phylogeny.

Timeline

The scheduled timeline for the project is shown as Fig. 1:

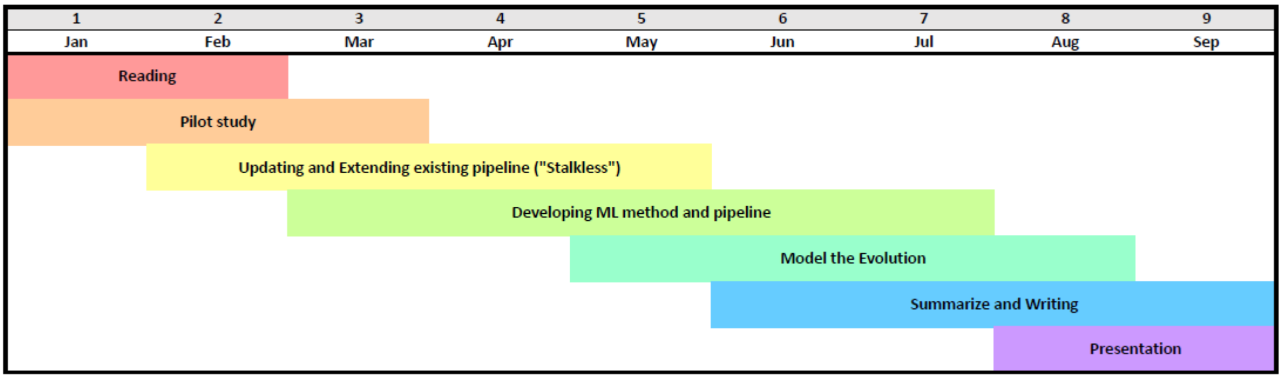

Figure 1: Gantt chart of the project

Budget

The project will need high performance computing for approximately 800 hours.

References

[Baker et al., 2021] Baker, W. J., Bailey, P., Barber, V., Barker, A., Bellot, S., Bishop, D., Botigué, L. R., Brewer, G., Carruthers, T., Clarkson, J. J., et al. (2021). A comprehensive phylogenomic platform for exploring the angiosperm tree of life. *bioRxiv*.

[Fu et al., 2004] Fu, H., Chi, Z., Feng, D., and Song, J. (2004). Machine learning techniques for ontology-based leaf classification. In *ICARCV 2004 8th Control, Automation, Robotics and Vision Conference, 2004.*, volume 1, pages 681–686. IEEE.

[Jähne and Haußecker, 2000] Jähne, B. and Haußecker, H. (2000). Computer vision and applications.

[Nanni et al., 2017] Nanni, L., Ghidoni, S., and Brahnam, S. (2017). Handcrafted vs. non-handcrafted features for computer vision classification. *Pattern Recognition*, 71:158–172.

[Pearse et al., 2018] Pearse, W. D., Cavender-Bares, J., Hobbie, S. E., Avolio, M. L., Bettez, N., Roy Chowdhury, R., Darling, L. E., Groffman, P. M., Grove, J. M., Hall, S. J., et al. (2018). Homogenization of plant diversity, composition, and structure in north american urban yards. *Ecosphere*, 9(2):e02105.

I have seen and approved the proposal and the budget.

Supervisor Signature:

Date: Dec 14 2021