Plant Species Identification using Persistent Homology

Hemming Ma
hemmingm@kth.se

KTH Royal Institute of Technology
Purdue university

April 2023

Introduction

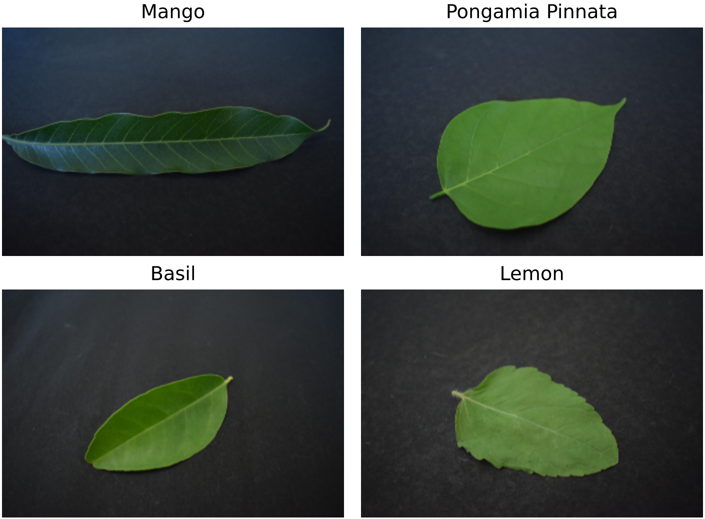

Figure: Source: [Cho+19].

Introduction

Question

Can we use the venation of a plant's leaves to identify its species?

Introduction

Question

Can we use the venation of a plant's leaves to identify its species?

My project

Extracting venation from leaves of different species, quantifying them using persistent homology, and using this to train a classification algorithm.

Outline

Dataset

- Raw images from *A Database of Leaf Images: Practice towards Plant Conservation with Plant Pathology* [Cho+19]
- Healthy mango, pongamia pinnata, basil, and lemon leaves
- 40 of each species

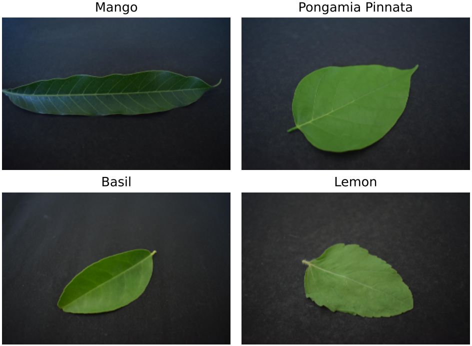

Figure: Sample leaves from [Cho+19].

Pre-processing

- Raw images are large at $\sim 1 - 2$ MB each
- Pre-process by downsampling and removing excess background
- Done using OpenCV, a Python computer vision library

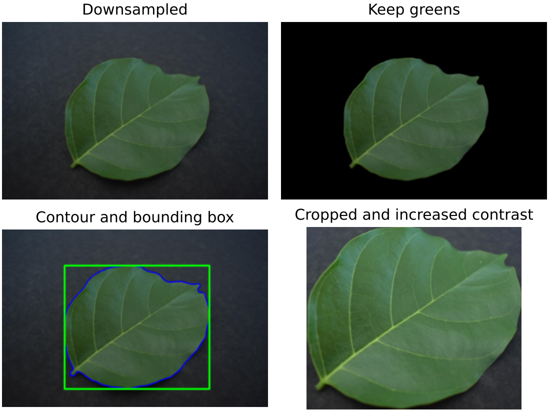

Figure: Pre-processing of a Pongamia Pinnata leaf.

Venation extraction using ilastik

- Venation extraction of the pre-processed images using ilastik, an interactive image classification and segmentation tool.
- Accuracy of ilastik model determined by eye.
- Remove redundant points by skeletonizing.

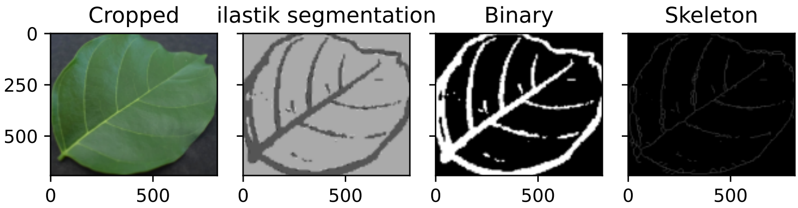

Figure: Extracted venation from a Pongamia Pinnata leaf.

Persistent homology

- Persistent homology is tool from TDA that counts the number of connect components and “holes” in your data
- We use giotto-tda, a Python TDA library to compute it
- Need to convert extracted venation into point cloud data

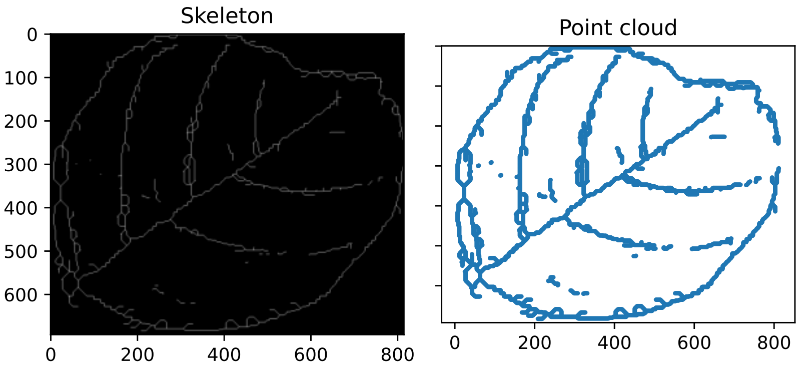

Figure: Venation represented by 7391 data points.

Persistent homology

- Point cloud representation has $\sim 10^4$ points, this is too much!
- Solution: consider a subset consisting of $k \in [0, 1]$ of all points chosen at random

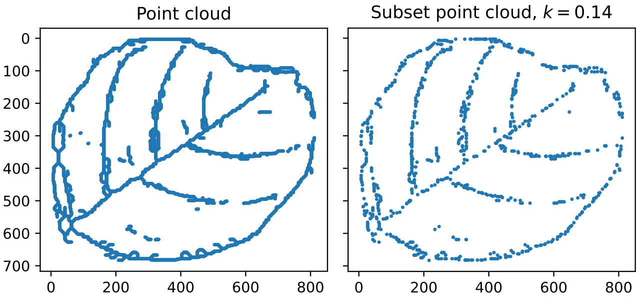

Figure: Subset of total point cloud consistinng of $k = 0.14$ of all points, 1000 points total.

Persistent homology

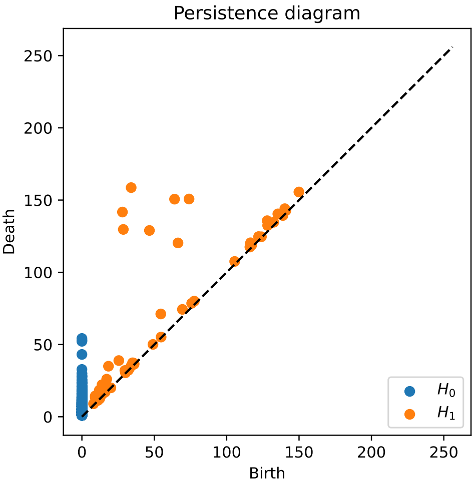

Figure: Persistence diagram of subset point cloud.

Persistent entropy

We use persistent entropy turn the persistence diagram into a $2d$-feature vector.

Definition

If $D = \{(b_i, d_i)\}_{i \in I}$ is a persistence diagram, then its *persistent entropy* is

$$S(D) = \sum_{i \in I} p_i \log\left(\frac{1}{p_i}\right)$$

where

$$p_i = \frac{d_i - b_i}{L_D} \quad \text{and} \quad L_D = \sum_{i \in I}(d_i - b_i).$$

[AGR17] has shown persistent entropy to be a stable way of comparing persistence diagrams.

Dependence on k

- Choose k so that it affects results minimally
- choose k as large as possible that still has a small computation time, take $k = 0.2$

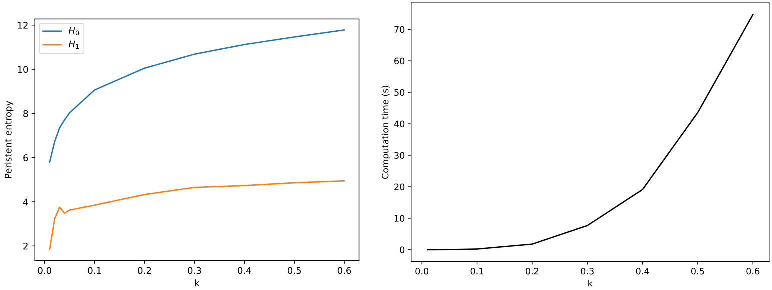

Figure: Persistent entropy values for H_0, H_1 and computation time as a function of subset proportion k.

k-NN

- We use a k-nearest neighbours (k-NN) classification algorithm
- k-NN is a *non-parametric supervised learning* classification model
- Simple, explainable, easy to visualise

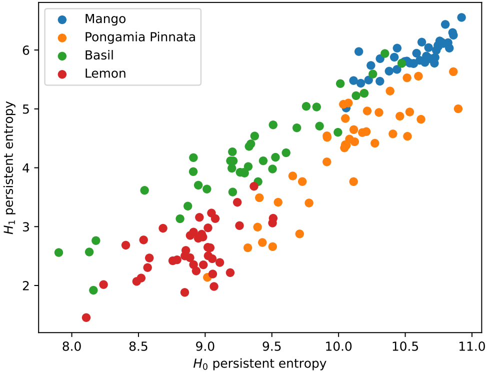

Figure: Feature vectors.

Accuracy and choosing k

We use accuracy = (correct predictions)/(total predictions) as a measure of goodness.

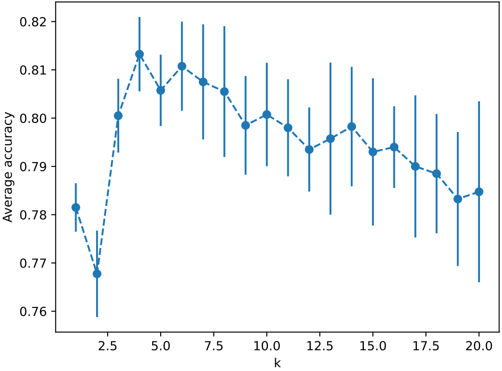

Figure: Average repeated 5-fold cross validation, 25 repetitions. Maximum accuracy of 0.813 at $k = 4$.

Result

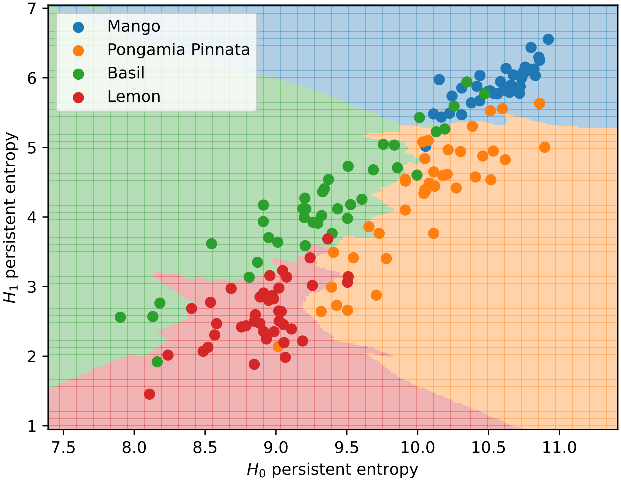

Figure: 4-NN decision boundary with average accuracy of 0.813.

Remarks and future work

- Use more data, increase types of leaves
- Do persistent homology computation on a more powerful computer
- Quantify accuracy of ilastik model
- k-NN is $\mathcal{O}$(number of points) which can be slow, use different model such as Random forest
- Classifier limited to leaves taken on black background
- Take more features into account

Acknowledgement

I want to thank Professor censorAlexandria Volkening for the idea to study the persistent homology of leaf venation!

References

[AGR17] Nieves Atienza, Roco Gonzalez-Diaz, and Matteo Rucco. "Persistent Entropy for Separating Topological Features from Noise in Vietoris-Rips Complexes". In: *CoRR* abs/1701.07857 (2017). arXiv: 1701.07857. URL: http://arxiv.org/abs/1701.07857.

[Cho+19] Siddharth Singh Chouhan et al. *A Database of Leaf Images: Practice towards Plant Conservation with Plant Pathology*. Mendeley Data. 2019.

The end

Thank you!

Questions?

`hemmingm@kth.se`