Topological Data Analysis Applied in Identifying Leaf Morphology of Plants

Student : Yuqing Xing

Huazhong University of Science and Technology

Outline

Outline

Motivation and Significance

- **The beauty of plant forms**. There are about 391,000 species of vascular plants currently known to science worldwide. The roots, branches and leaves of plants have obvious geometric and topological features. For thousands of years, the morphological beauty of plants has attracted countless people.

Motivation and Significance

Motivation and Significance

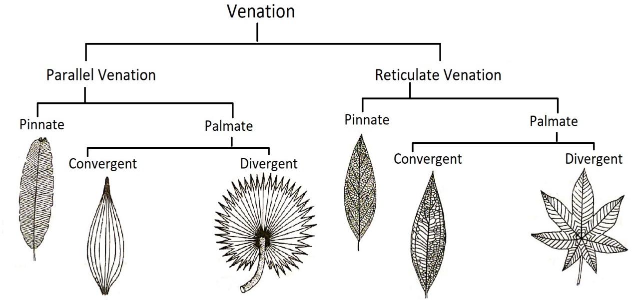
Intelligent classification of plants greatly reduces human labour. The study of plant leaves is a direct and effective method for plant classification. Leaf contour and vein can be used as the basis for plant classification.

To quantify the characteristics of plant leaves, a large number of researchers consider the geometric features of leaves, without additional studies on topological features of leaves. Leaf venation is the distribution form of leaf veins on leaves. **Leaf contour** characterizes the shape of the leaf. Both of them have fine topology and geometry structures, providing important bases for plant classification.

Reticular venation

Three main types of leaf venation:
(1) reticular venation (2) parallel venation (3) bifurcated venation

The figure below shows the venation structure of most leaves.



Background

Advanced mathematics models in algebraic topology, differential geometry and algebraic graph theory, have been proposed for topology data analysis (TDA) in recent years. Several invariants, such as topological invariant (Betti number), geometric invariant (curvature) and algebraic graph invariant (eigenvalue) are considered in TDA.

As an emerging mathematical method, it can be applied to many fields. Persistent homology is a fundamental method in TDA, which is a flourishing method to extract data features and classifying big data.

Background

In 2017, Li M, Frank MH, Coneva V, Mio W, Chitwood DH, Topp CN presented a morphometric technique based on topology, using a persistent homology framework to measure the morphology of leaves and study characteristics extracted from leaf contours and roots. For a given topological feature, persistent homology creates persistence barcodes or diagrams that can be used to compare the overall topological similarity between objects.

Combined with analysis methods (e.g., canonical variant analysis (CVA), principal component analysis(PCA)), persistent homology can be used to classify the morphology of plants. Topological features provide a framework to quantify leaves of plants comprehensively.

Background

At present, **machine learning** is a sub-field of artificial intelligence, which has been applied in various fields. The application of machine learning in plant species recognition can help botanists and the general public to quickly identify plant species. The steps of plant classification by machine learning include dataset acquisition, data preprocessing, feature extraction and classification.

In May 2021, Malarvizhi K, Sowmithra M, Gokula Priya D, Kabila B extracted more than 20 geometric features from leaf species and classified 32 leaf species with support vector machine(SVM), K-Nearest Neighbor(KNN) and random forest(RF) classification methods. However,**we haven't found research concerning extracting topological features of leaves to classify leaves by machine learning so far**.

Outline

Research problems

The problems we focus on:
(1) Using topological data analysis(TDA) to quantify topological characteristics of plant leaf veins and leaf contours.
(2) Extracting engough topological and geometric features of leaves to quantify the morphology of plant leaves.
(3) Classifying plant leaves by machine learning methods.

More research:
(4) Analyzing the correlation and interaction between leaf contours and veins.
(5) Searching for more topological invariants for topological data analysis and ameliorating persistent homology theory.

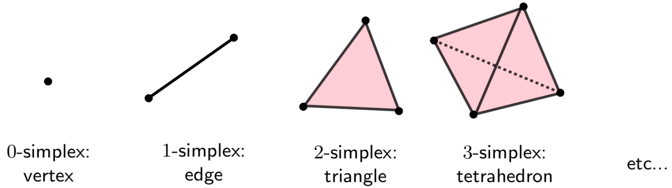
Outline

Homology

Given a set $P = \{p_0, ..., p_n\}$ of affinely independent points in Euclidean space R^{n+1}, the **k-simplex** σ spanned by P is the set of linear combinations

$$\sum_{i=1}^{n} \lambda_i p_i, \qquad with \sum_{i=1}^{n} \lambda_i = 1,\ \lambda_i \geq 0$$

Elements of P are called vertices of σ. Given two simplices A and B, A is called a face of B if each vertice of A is a vertice of B.



Homology

A **simplicial complex** K is a collection of simplices satisfying:
(i) each face of a simplex of K is a simplex of K,
(ii) the intersection of any two simplices of K is either empty or a common face of both.

Let $P = \{p_1, ..., p_n\}$ be a finite set. An **abstract simplicial complex** K with vertex set P is a collection of subsets of P satisfying:
(i) each single point set of P belongs to K,
(ii) if $\tau \in K$ and $\sigma \subseteq \tau$, then $\sigma \in K$.
Elements of K are called simplices.

Remark: The abstract simple complex is a generalization of simple complex.

Homology

Let X be a simplicial complex of dimension n and X^k be the set of all k-simplices in X. Define a **k-chain on X** as a linear combination $\sum_{\sigma_i \in X^k} a_i \sigma_i$ of k-simplices in X, where $a_i \in \mathbb{Z}/2\mathbb{Z}$ and σ_i is a k-simplex represented by $[v_0, v_1, ..., v_k]$. **A k-th chain group** $C_k(X)$ is the set of all k-chains with multiplication induced by the addition of coefficients .

The boundary operator $\partial_k : C_k(X) \to C_{k-1}(X)$ is a linear map

$$\partial_k \left[v_0, v_1, ..., v_k\right] = \sum_{i=0}^{k} (-1)^i \left[v_0, ..., \hat{v_i}, ..., v_k\right]$$

A k-chain c is a **k-boundary** if c is in the image of the boundary map ∂_{k+1}. A k-chain c is a **k-cycle** if $\partial_k(c) = 0$.

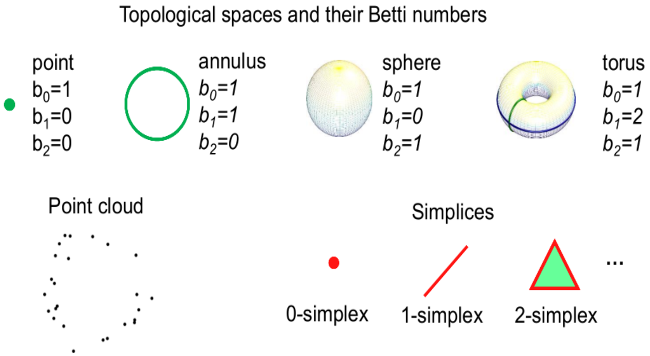
Homology

Remark: $\partial_p \circ \partial_{p+1} = 0$ follows from the definition of boundary operator. Thus $Im(\partial_{k+1})$ is a subgroup of $ker(\partial_k)$.

The k-th homology group is the quotient group $H_k(X) = ker(\partial_k)/Im(\partial_{k+1})$.

The k-th Betti number is the rank of the k-th homology group.

A topology invariant: Betti number



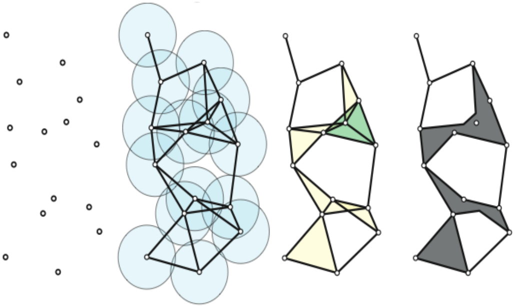
Introduction of persistent homology

Let S be a finite set of points in Euclidean space, d be a positive real number, **the Rips complex** $R_\rho(S)$ is the abstract simplicial complex whose k-simplices are determined by subsets of k $+$ 1 points in S with diameter at most ρ, where the diameter of a simplex is the maximal distance between all two points in it.

A filtration is a finite increasing sequence of sets $K = \{K_i \mid K_i \subset K_j \quad i < j,\ i, j \in R\}$. Given a dataset, we can get complex generated by constructing a filtration.

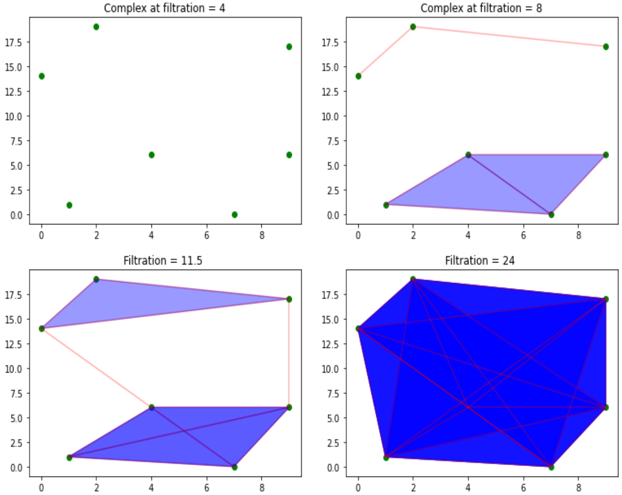
Constructing Rips filtration

The following figures show the filtration process of a dataset at a filtration value ρ. Let r be the radius of the yellow sphere, here $\rho = 2r$.



Constructing Rips filtration

The following figures show four different Rips complex of seven points as the filtration value increases.

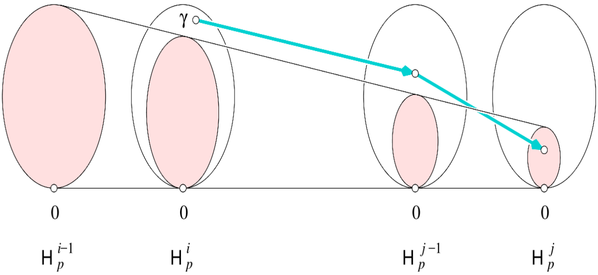


Persistent homology group

The inclusion $f^{i,j} : K_i \to K_j$ can induce a homomorphism $f_p^{i,j} : H_p(K_i) \to H_p(K_j)$. **The p-th persistent homology group** $H_p^{i,j}$ is defined as the image of $f_p^{i,j}$. The p-th persistent betti number is defined as the rank of $H_p^{i,j}$.

Let γ be a class in $H_p(K_i)$, γ is born at K_i if $\gamma \notin H_p^{i-1,i}$; furthermore, γ dies entering at $K_j(j > i)$ if $f_p^{i,j-1}(\gamma) \notin H_p^{i-1,j-1}$ and $f_p^{i,j}(\gamma) \in H_p^{i-1,j}$. If homology class γ is born at K_i and dies entering K_j , define the **persistence** of γ as $j - i$.

Persistent homology group



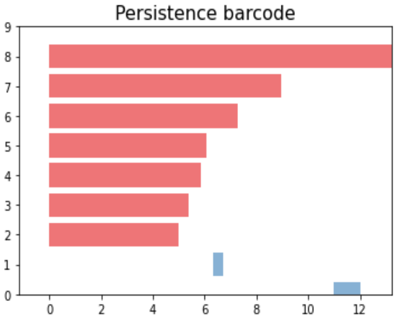
The above graph illustrates the meaning of "persistence" in persistent homology group. Class γ is born at time i and dies at time j, so the persistence of γ is $j - i$.

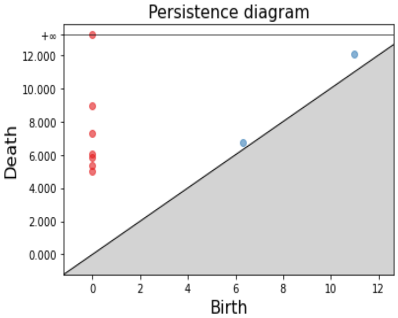
Persistence barcode/Persistence diagram

Persistent homology can be represented in two equivalent forms: either as a **persistence barcode** or as a **persistence diagram**. We use matrix reduction algorithm to get the persistence barcode or diagram.

A persistence barcode is a collection of bars whose left endpoints represent birth and right endpoints represent death respectively. Each interval in the persistence barcode can be represented in the persistence diagram by a point in the plane equivalently, with its **birth** on the horizontal axis and with its **death** on the vertical axis. Points of a persistence diagram all lie above the diagonal $y = x$.

Persistence barcode/Persistence diagram





Remark: We care about long bars or points far away from the diagonal y=x. The number of n-dimension bars keep alive until the complete complex is reached is equal to the n-th betti number in the filtration.

Applications

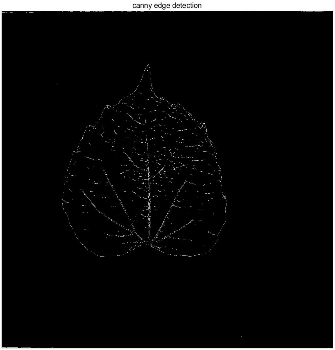
Steps of the PH method

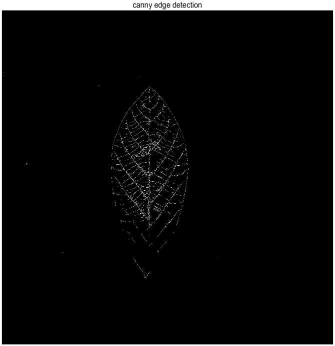
1. Getting two-dimensional leave images
2. Extracting datasets from leave images
3. Computing persistent homology(e.g., persistent barcode, bottleneck distance, Betti number, Euler characteristic curve)
4. Getting topological features of leaf veins and contours

Steps of plants classification by Machine Learning

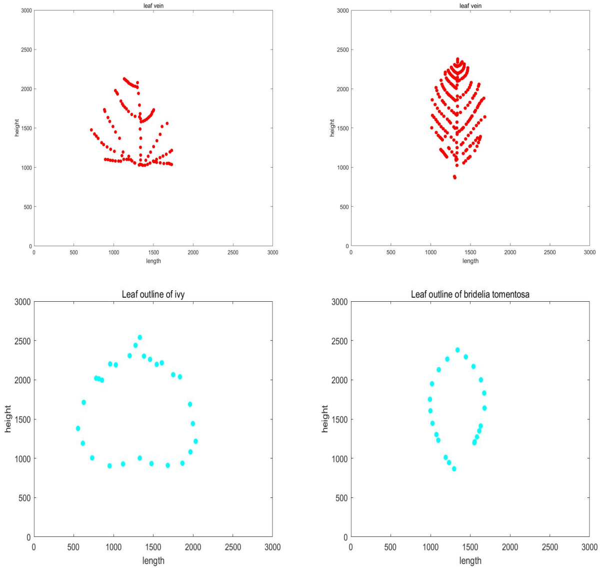
1. Collecting leaf images of plants
2. Image pre-processing
3. Feature extraction of leaves
4. Classification of plant species

Image extraction of leaf venations and contours





Getting Point clouds of venations and contours

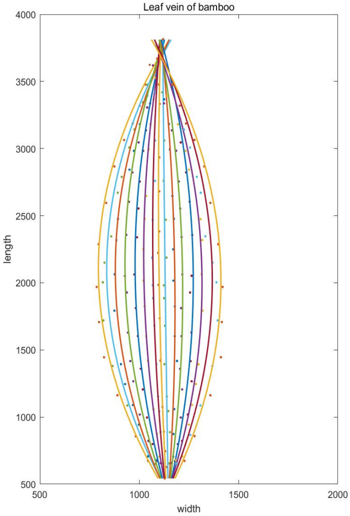


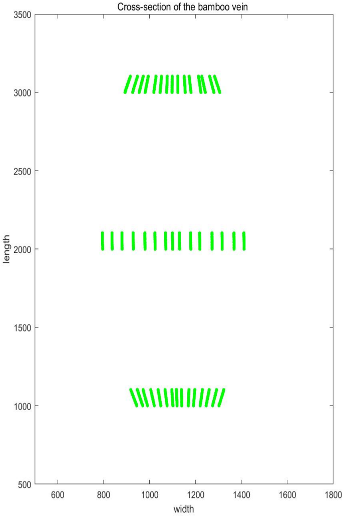
PH analysis of a bamboo leaf vein

We have conducted experiments on the simulation vein data of bamboo leaves. Specifically, we establish a height function, screen out the point set of leaf vein longitudinal coordinates between 1002-1102, 2001-2102 and 3002-3102, and obtain three cross-sections of bamboo leaf vein pointset, then make persistent homology analysis.

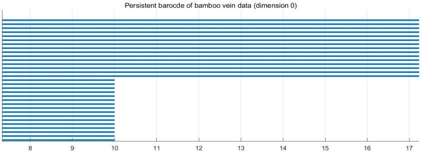
Graphs are gotten by using Python library **GUDHI**. The left figure in the next slide is a simulated photo of bamboo leaf veins, and the right is a dataset of three cross-sections of the veins.

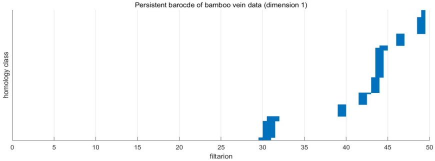
PH analysis of a bamboo leaf vein





Enlarged PH barcode with 15 long bars





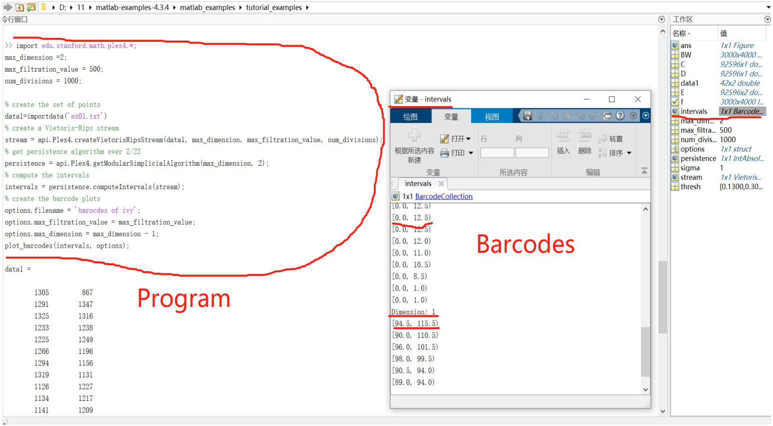
The above barcodes are resulted from the pointset of leaf vein longitudinal coordinates between 1002-1102. It show that the number of H_0 bars for the cross-section sharply reduces to 15 when filtration value is 10, corresponding to 15 topological connected components of a cross section. However, the characteristic of H_1 barcode is not apparent.

Extracting features

Geometric features:
curvature of the main vein, aspect ratio, rectangularity, roundness, shape parameter, etc.

Topological features:
• Compute the first/second longest H_0/H_1 BT/DT/PL; the mean of BTs/DTs/PLs of all H_0/H_1 barcodes; numbers of H_1 barcodes of each dataset.
• Rotate the persistence diagram 45° clockwise, divide the resulting rotation persistence diagram into $n \cdot n$ grids (or pixels), and then construct a persistence function on each grid. The number of points in each grid in the rotation persistence diagram also can be measured.
• For persistent barcodes, construct a function associated with each barcode as a topological feature.

Extracting features



For extracted leaf features, we are working on using several machine learning algorithms(SVM,CNN,RF) to classify leaves of plants and evaluate the effectiveness of algorithms.

Outline

Summary

In our work, the plant morphology is quantified by **persistent homology**, and the plant leaves are classified by machine learning algorithm. Previous researchers often used geometric methods to analyze the morphological characteristics of plants. **The innovation of our work** is extracting and quantifying the morphology of leaf veins by topological data analysis and use them as topological features for classifying.

Our work haven't innovated the theory of persistent homology yet. In the future work, we try to make a deep study on persistent homology theory. We plan to **construct more topological invariants** to describe plant morphology and **give a plant classification framework** based on topological data analysis.

References

[1] Willis KJ: State of the World's Plants. 2016.

[2] Li M, Frank MH, Coneva V, Mio W, Chitwood DH, Topp CN. The Persistent Homology Mathematical Framework Provides Enhanced Genotype-to-Phenotype Associations for Plant Morphology, Plant Physiology Aug 2018, 177 (4) 1382-1395.

[3] Gunter Rote, Vegter G . Computational Topology: An Introduction[J]. Springer Berlin Heidelberg, 2006.

[4] Chambers E W , Silva V D , Erickson J , et al. Rips Complexes of Planar Point Sets[J]. Discrete and Computational Geometry, 2007, 44(1):75-90.

[5] Ghrist R . Barcodes: The persistent topology of data[J]. Bulletin of the American Mathematical Society, 2008, 45(1):61-75.

References

[5] Edelsbrunner H , Harer J . Computational Topology: An Introduction[J]. American Mathematical Society, 2009.

[6] Adams H, Moy M. Topology Applied to Machine Learning: From Global to Local[J]. 2021.

[7] Azlah, Muhammad Chua, Lee Suan Rahmad, Fakhrul Abdullah, Farah Wan Alwi, Sharifah Rafidah. (2019). Review on Techniques for Plant Leaf Classification and Recognition. Computers. 8. 77. 10.3390/computers8040077.

[8] M Li, LL Klein, KE Duncan, N Jiang, CN Topp. Characterizing 3D inflorescence architecture in grapevine using X-ray imaging and advanced morphometrics: implications for understanding cluster density, Journal of Experimental Botany, Volume 70, Issue 21, 1 November 2019, Pages 6261–6276.

References

[9] Sack L, Scoffoni C. Leaf venation: structure, function, development, evolution, ecology and applications in the past, present and future. The New Phytologist. 2013. Jun; 198(4):983-1000. DOI: 10.1111/nph.12253. PMID: 23600478.

[10] Mao L, Duncan K, Topp C N, et al. Persistent homology and the branching topologies of plants[J]. American Journal of Botany, 2017, 104(3).

[11] Chi S P , Xia K , Si X L . Persistent-Homology-based Machine Learning and its Applications – A Survey[J]. SSRN Electronic Journal, 2018.

[12] Malarvizhi K , Sowmithra M , Gokula Priya D , Kabila B, Machine Learning for Plant Species Classification using Leaf Vein Morphometric, International Journal of Engineering Research and Technology (IJERT) Volume 10, Issue 04 (April 2021).