基于拓扑数据分析的植物叶片形态量化

邢雨晴

华中科技大学

2022 年 1 月 26 日

目录

研究背景

竹色溪下绿，荷花镜里香。从古至今，植物的形态美吸引着无数的人。全世界约有 391000 种维管束植物，植物的分类对保护生物圈至关重要。植物叶形稳定，为对植物进行分类，对植物叶子的研究是一种直接而有效的方法。

叶子轮廓形态各异，有扇形、圆形、掌形、心形、针形等；叶脉是叶片上分布的粗细不同的维管束，种子植物的叶脉脉序主要分为网状脉序、平行状脉序、叉状脉序等类型。叶子轮廓和叶脉都具有良好的拓扑和几何结构，它们的拓扑特征为植物叶片的全面量化提供了一个研究框架。

叶子轮廓

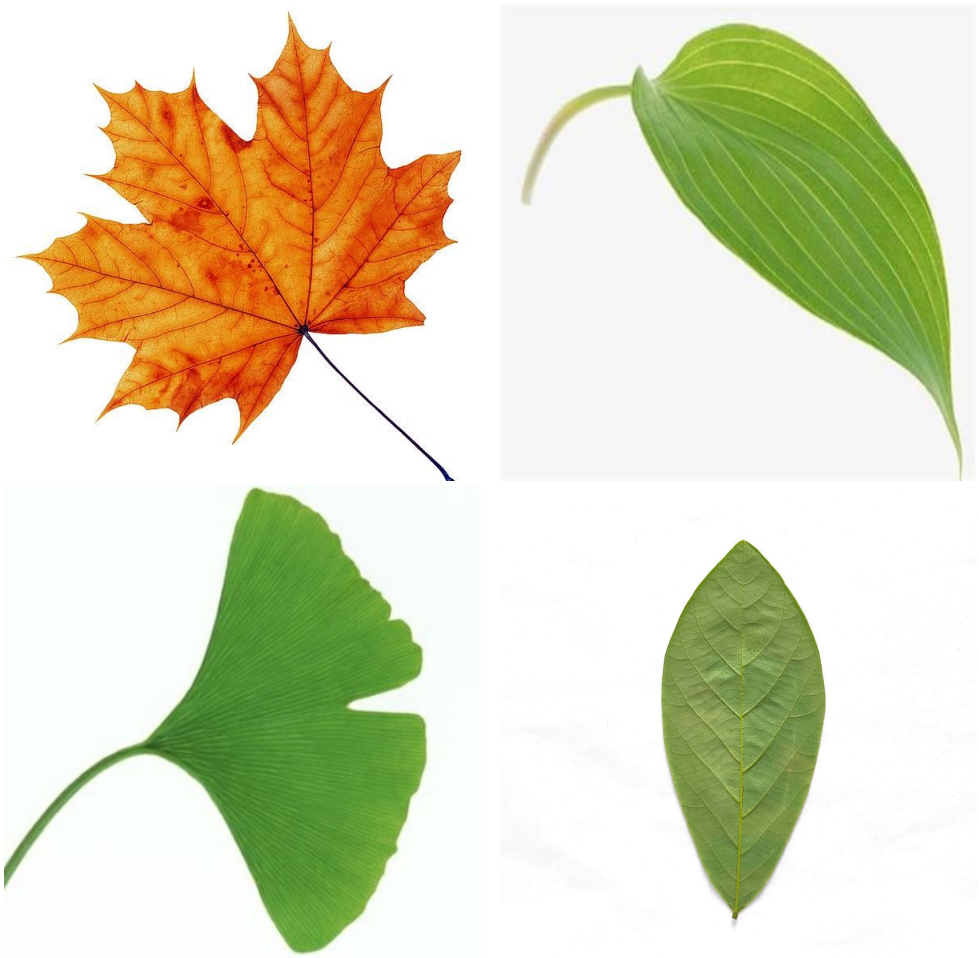


图 1　四种不同的叶片

叶脉脉序

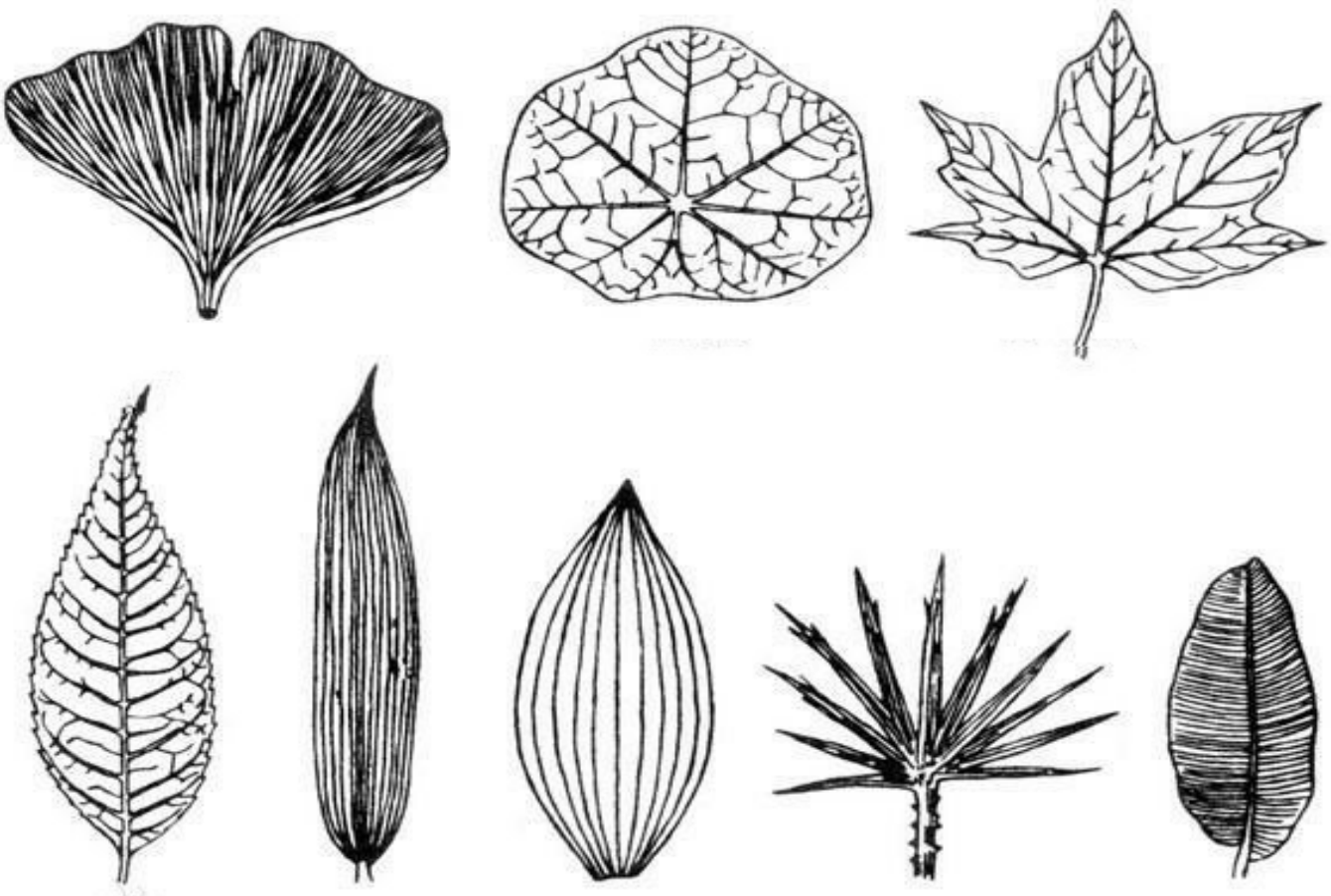


图 2　叶脉脉序的几种模式

国内外研究现状

- Frosini 在 1990 年提出了尺度函数及其相关理论，这等价于 0 维持续同调; 1999 年，Robins 研究了样本空间的同调，描述了由包含诱导的同调群的像; 2002 年，H.Edelsbrunner，D.Letscher 和 Afra Zomorodian 提出了计算持续同调的算法。
- 2004 年，Afra Zomorodian,Gunnar Carlsson 和 Leonidas J.Guilbas 等人定义了拓扑描述符“条码”，用于提取过滤过程中的拓扑不变量；2005 年，Afra Zomorodian 和 Gunnar Carlsson 又公布了一种计算任意主理想域上任意维数的持久同调群的算法，推广了之前的持续同调算法；2009 年 Gunnar Carlsson 系统阐述了拓扑数据分析相关理论及其应用方向。
- 2016 年，吴杰, Stephane Bressan, 李京艳, 任世全等提出了超图的嵌入同调及超图序列的持续嵌入同调，之后给出了超图同调的算法，并于 2020 年证明了超图持续同调的稳定性。
- 关于拓扑数据分析的理论、算法及应用，都在蓬勃发展。

国内外研究现状

就植物分类来讲，之前的研究者在分析植物形态特征时，往往考虑几何方法。若用持续同调方法提取植物的拓扑和几何特征，并用机器学习算法对已提取特征的植物进行分类，这种基于持续同调的方法将为植物分类提供新思路。

近些年，持续同调被一些学者应用于量化植物的特征。2017 年,Li M, Frank MH, Coneva V, Mio W, Chitwood DH, Topp CN 应用持久同调方法量化番茄亲本 cv M82 及其渗入系品种中叶片、茎、根的形态，并利用主成分分析、典型相关分析等方法分析不同番茄品种形状的区别。2018 年，Li M 等人分析了来自世界各地的 141 个植物科和 75 个地区的 182707 片叶子的拓扑特征, 应用持续同调 (PH)，以组合形态特征描述传统的形状描述符，结合线性判别分析 (LDA) 量化植物叶片。

国内外研究现状

将机器学习应用于植物物种识别，可以帮助植物学家和大众快速识别植物物种。用机器学习进行植物分类的步骤包括预处理、分割、特征提取和分类。2021 年 5 月，Malarvizhi K, Sowmithra M, Gokula Priya D, Kabila B 用机器学习方法从叶片种提取了 20 多个几何特征，并用支持向量机 (SVM)、k-近邻 (KNN) 和随机森林 (RF) 等算法对 32 种叶片进行了分类。然而，目前在有关机器学习分类叶片的论文中，提取的植物叶片特征是几何特征，尚未涉及拓扑特征。

目录

研究问题

目前进行的工作:

(1) 用拓扑数据分析 (TDA) 分析植物叶轮廓、叶脉拓扑特征;
(2) 提取叶片的拓扑和几何特征，量化植物叶片的形态;
(3) 用随机森林 (BF)、K-近邻算法 (KNN) 等机器学习方法对植物叶片进行分类。

更多研究:

(4) 用离散曲率刻画植物轮廓几何特征；将持续加权同调应用于植物叶片的分类。

研究步骤

用持续同调提取叶轮廓特征的步骤：

1. 获取原始叶子图像，图像预处理
2. 从图像中提取叶轮廓数据集
3. 构造过滤复形
4. 计算持续同调
5. 获得叶轮廓的拓扑特征统计量

用机器学习分类植物的步骤：

1. 获取叶子形态数据集
2. 数据集预处理
3. 综合提取叶子的 10 组几何和拓扑特征统计量
4. 用机器学习算法分类植物，比较不同算法的有效性

目录

持续同调原理

设 P 为一有限点集，r 是一正整数。定义 $Vietoris-Rips$ **复形**为抽象单纯复形 $R(P,r)=\{\sigma\subset P \mid diameter\,\sigma\leq 2r\}$; 定义 **Céch 复形**为抽象单纯复形 $C(P,r)=\left\{\sigma\subset P \mid \bigcap_{x\in\sigma} B_x(r)\neq\emptyset\right\}$。

定义**过滤集** K 为一递增集合序列 $K=\{K_i \mid K_i\subset K_j,\, i<j,\, i,j\in R\}$ 。给定数据集 P ，记 $d_1,d_2,...,d_m$ 为逐渐增大的过滤值，可由 P 构造出 $Vietoris-Rips$ **过滤复形** $\{K(d_1),K(d_2),...,K(d_m)\}$。

Rips 过滤

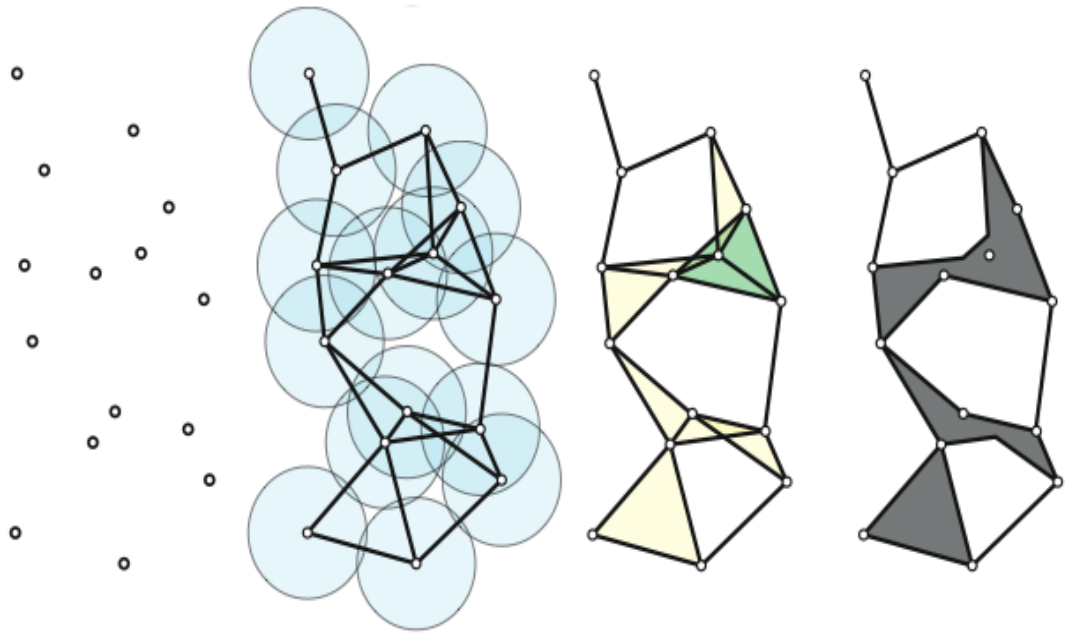


图 3　数据集在过滤值为 r 时生成的 Vietoris-Rips 复形 $R(P, r)$ 及对应的拓扑空间

Nerve 定理

定义

设 $U = \bigcup_{i \in \wedge} U_i$ 是一拓扑空间的开覆盖。定义 U 的 **Nerve** 为一抽象单纯复形 $N(U)$，满足 (1) U 是 $N(U)$ 的顶点集，
(2) $\sigma = [U_{i_0}, ..., U_{i_k}] \in N(U)$ iff $\bigcap_{j=0}^{k} U_{i_j} \neq \varnothing$。

定理

设 $U = \bigcup_{i \in \wedge} U_i$ 是欧式空间的子空间 X 的一个开覆盖，且 U 中任意两开集无交或交是可缩的，则 X 和 $N(U)$ 同伦等价。

根据 *Nerve* 定理，图 3 的第 1 个图表示的数据集在过滤值为 r 时生成的 Céch 复形 $C(P, r)$ 和第 4 个图表示的拓扑空间有着相同的同调。由于 $C(P, r) \subseteq R(P, r) \subseteq C(P, \sqrt{2}r)$，因此可用点集生成的 Vietoris-Rips 复形的同调刻画对应拓扑空间的同调。

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

持续同调

设 $f^{i,j}: K_i \to K_j$ 为一个包含映射，它可诱导同调群的群同态 $f_p^{i,j}: H_p(K_i) \to H_p(K_j)$。**p 阶持续同调群** $H_p^{i,j}$ 定义为 $f_p^{i,j}$ 的像。**p 阶持续同调群的维数**定义为 $H_p^{i,j}$ 的秩。若 γ 是 $H_p(K_i)$ 中的一个同调类，称 γ **出生**于 K_i 若 $\gamma \notin H_p^{i-1,i}$；称 γ **死亡**于 $K_j (j > i)$ 若 $f_p^{i,j-1}(\gamma) \notin H_p^{i-1,j-1}$ 且 $f_p^{i,j}(\gamma) \in H_p^{i-1,j}$。若同调类 γ 出生于 K_i，死亡于 K_j，则称 γ 的**持续时间**为 $j-i$。

持续同调可由持续二维码 (Persistent barcode) 或者持续图 (Persistent diagram) 直观描述。持续二维码中的条形码显示了每个同调代表元的出生时间和死亡时间。每一个条形码均可在持续图中由横坐标表示出生时间、纵坐标表示死亡时间的一点来表示。

例子：不同形状数据集的持续同调比较

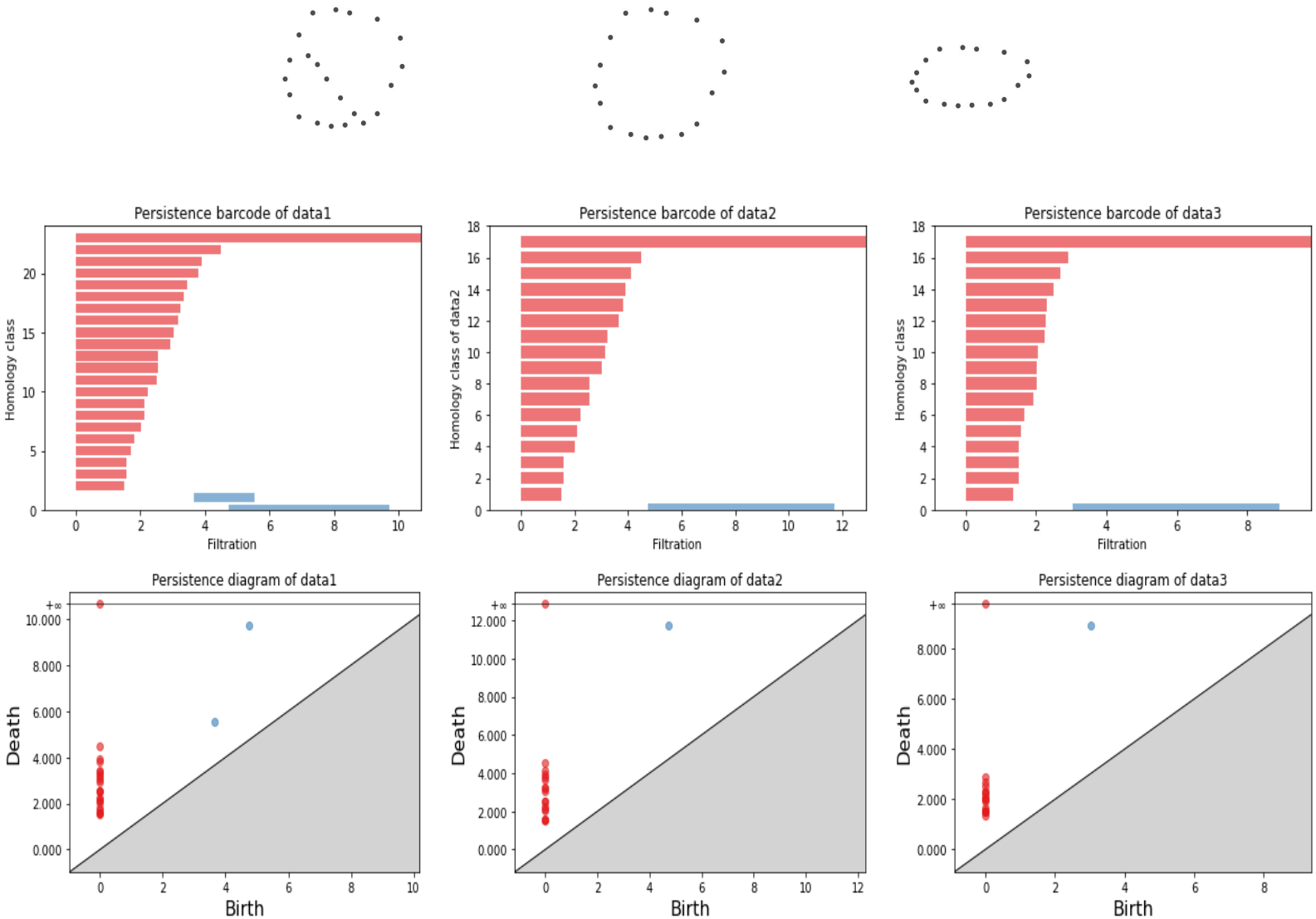


图 4　三种不同数据集的持续同调条形码和持续图的比较

持续同调应用：叶子分类

实验数据取自 Li M, An H, Angelovici R 等人的论文 “Topological Data Analysis as a Morphometric Method: Using Persistent Homology to Demarcate a Leaf Morphospace” (2018)。从叶片轮廓数据库中选取了 11 种植物的成熟健康的叶片进行实验，每种叶片随机选 15 个轮廓数据样本，每个样本含 800 个数据点。这 11 种植物分别为：土豆，六出花，芸薹，彩叶草，棉花，常春藤，三角枫，西番莲，辣椒，马铃薯，拟南芥。

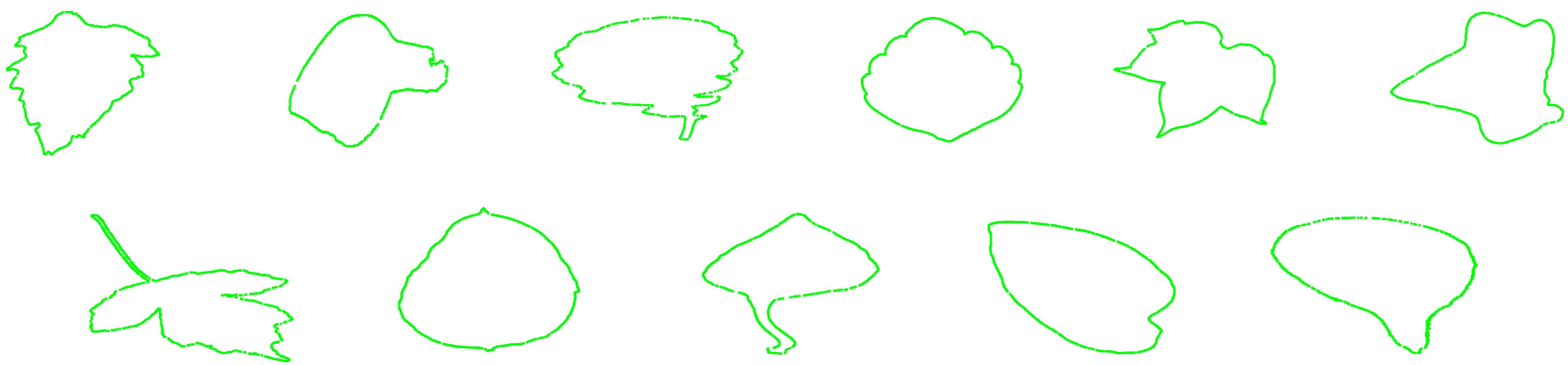


图 5　11 种植物样本的轮廓形状

提取叶子轮廓拓扑特征

对于来自 11 种植物的叶片轮廓数据集，持续同调分析可描述轮廓的形状。我们根据 165 个样本数据，得到每个样本的持久条形码/持续图，以及各维条码的条数、出生时间、死亡时间。

从每个叶片的持续同调图中提取 4 个拓扑特征统计量：**次长 0 维条码死亡时间，第 3 长 0 维条码死亡时间，最长 1 维条码出生时间，最长 1 维条码死亡时间**。

提取叶子轮廓几何特征

同一植物的叶片可能大小不等，无法用轮廓面积、周长等参数准确识别叶片，这要求提取的几何特征具有平移、旋转、伸缩不变性。本工作选取**纵横轴比、偏心率、矩形度、等效圆半径、周长凹凸比、形状参数**刻画叶片轮廓的几何特征。设 x, y 分别表示叶片凸包的横轴和纵轴长度，z_1, z_2 分别表示叶片短轴和长轴长度，L,S 分别表示叶片的周长和面积，Lc,Sc 分别表示叶片凸包的周长和面积。六个特征的具体计算公式及描述如下：

纵横轴比 (P1) $= y/x$，描述叶片最小外接矩形与正方形的接近程度；

偏心率 (P2) $= z_1/z_2$，用于描述叶片的细长性；

矩形度 (P3) $= S/Sc$，测定叶片在最小外接矩形中的填充程度；

等效圆半径 (P4) $= \sqrt[2]{S/\pi}$，用于描述叶片和圆的相似程度；

周长凹凸比 (P5) $= L/Lc$，刻画叶片边缘的锯齿度和规则性；

形状参数 (P6) $= 4\pi S/L^2$，反映叶片面积的紧凑性。

11 种叶片的拓扑特征比较

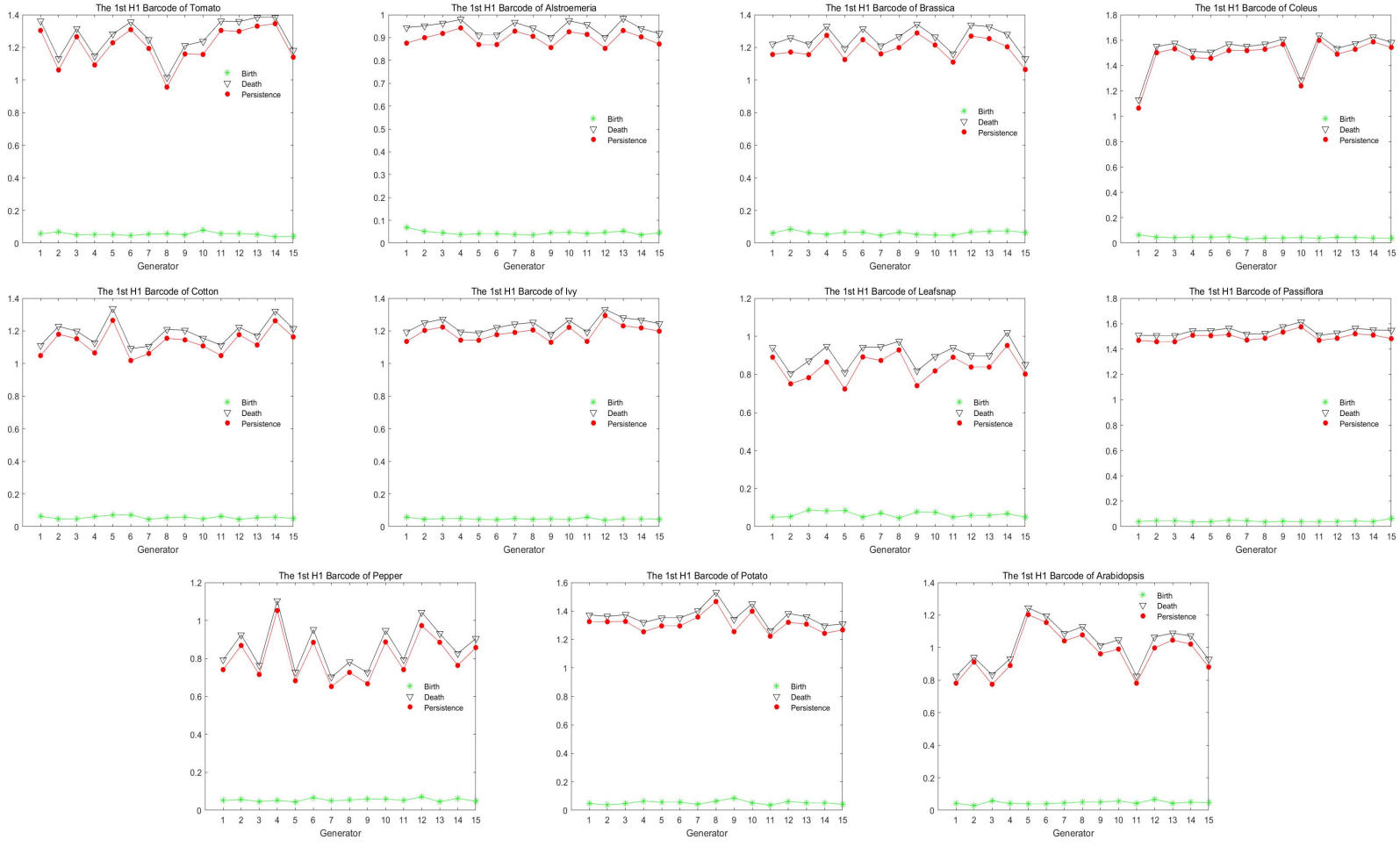


图 6　比较分析取自 11 种叶片的 165 个样本的最长 1 维条码出生时间、死亡时间、持续时间，不同植物叶片有一定的差异性。

11 种叶片的拓扑特征比较

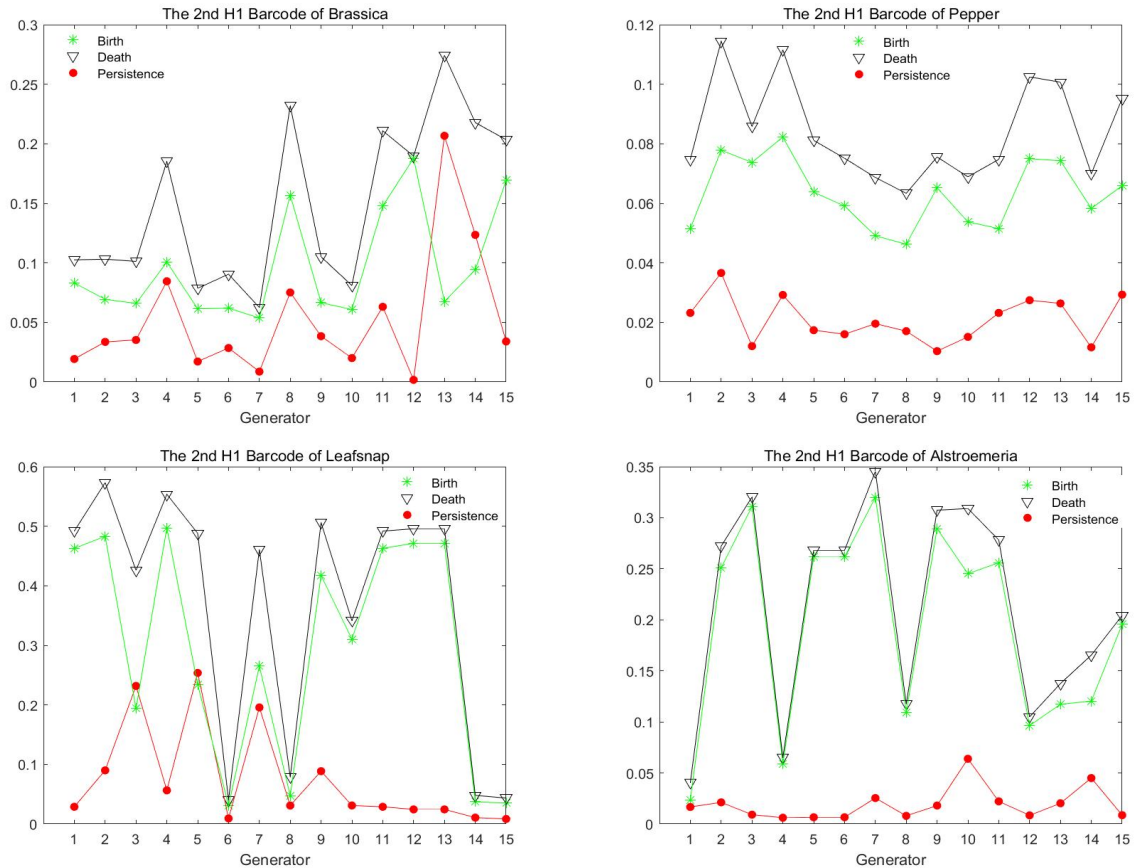


图 7　比较分析芸薹, 三角枫, 辣椒, 六出花的 1 维条码出生时间、死亡时间、持续时间，出生、死亡时间表现不稳定，波动幅度总体较大，不能较好地反映出叶片的几何特征。因此在选取拓扑特征统计量时舍去次长 1 维条码出生时间、死亡时间这两个拓扑统计量。

机器学习分类

根据提取的 4 个轮廓拓扑统计量、6 个轮廓几何统计量、10 个轮廓综合统计量，用 K-近邻算法 (KNN)、随机森林 (RF) 分别对叶片进行分类，选取的训练集和测试集的比例为 4:1。
KNN 算法的拓扑特征、几何特征、综合特征分类准确率分别为 **54.55%、72.73%、70.00%**; RF 算法的拓扑特征、几何特征、综合特征分类准确率分别为 **48.48%、70.00%、78.80%**。

问题：根据两种机器学习算法的分类效果，提取的 4 个轮廓拓扑统计量未能很好地对叶片进行分类，说明提取的拓扑统计量未能完全反映数据集的信息。因此下一步需要提取更合适的拓扑统计量来完善分类。

KNN几何特征分类准确度：
0.7272727272727273

KNN拓扑特征分类准确度：
0.5454545454545454

KNN10个综合特征分类准确度：
0.696969696969697

BF几何特征分类准确度：
0.696969696969697

BF拓扑特征分类准确度：
0.48484848484848486

BF综合特征分类准确度：
0.7878787878787878

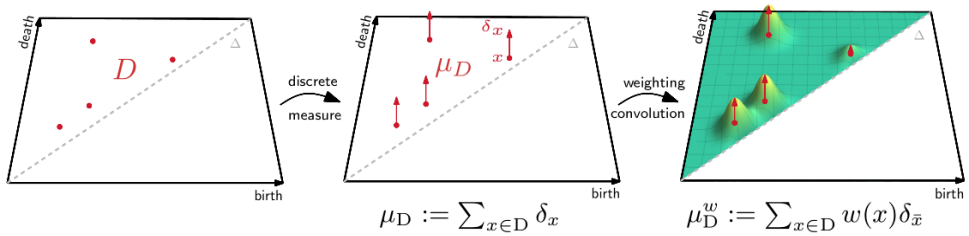
未解决的问题

(1) 叶脉是叶片的重要组成部分，若能获取叶脉数据库并提取出拓扑和几何特征，则叶片分类效果更好。目前植物形态分类采用较多的植物结构有叶子、花、果实等，在提取植物叶片特征时较常选取轮廓和纹理特征，而叶脉特征的提取需要精密的仪器设备，难度较大；

(2) 若难以获得大样本叶脉数据集，本工作需将研究重点放在 TDA 理论扩展上，将叶片轮廓分类作为一个理论实践。计算叶片轮廓离散曲率，尝试应用持续加权同调分类叶片。

(3) 目前关于构造抽象单纯复形计算持续同调，处理持续同调图时选取距离均为欧式距离。是否可以选取其他的度量 (如双曲度量) 描述持续同调图之间的距离？

展望: 寻找合适的拓扑特征



持续图中，远离对角线的点是持续时间较长的点，反映出明显拓扑特征，相比于靠近对角线的点更为重要。如何突出远离对角线的点？可以用高斯加权核函数来刻画。

给定二维持续图 D，定义 D 上的加权测度 $\mu^{\omega}_{D} = \sum_{x\in D} \omega(x)\delta(x)$，其中 x 是 D 中的点，$\delta(x)$ 是点 x 的 Dirac 测度，$\omega(x)$ 是点 x 的权重，表达式为 $\omega(x) = arctan(\lambda d(x,\Delta)^t)$，$\lambda, t \geq 0$，反映 x 到对角线的欧式距离。

D 上的高斯核 $K(x,\Delta) = e^{(-\frac{||x-\Delta||^2}{2\sigma^2})}$，其中 Δ 表示对角线的过点 x 的垂线的垂点，σ 是参数，$||\cdot||$ 表示 Euclidean 范数。

高斯加权核

持续加权高斯核函数 $K_{PWG}(x) = \frac{1}{\sqrt{2\pi}\sigma} w(x) e^{(-\frac{|x-\Delta|^2}{2\sigma^2})}$

高斯加权核函数将远离对角线的点赋予较大值，靠近对角线的点赋予较小值，较好地反映数据集拓扑特征，避免了贴近坐标轴的点带来的拓扑特征干扰。

对任一 k 维持续图，求其中每个点对应的高斯加权核；这等价于求持续条形码中的每个条码对应的加权高斯核，因此高斯加权核可作为持续拓扑特征应用于分类。如何高效提取加权持续图的高斯核特征？难点在于，由 800 个数据集生成的 1 维持续图中将出现 800 个点，分别计算这 800 个点的加权高斯核较为繁琐，应适当选取远离对角线的点求加权高斯核。

[2] Persistence weighted Gaussian kernel for topological data analysis, Kisano, Hiraoka, Fukumizu, ICML, 2016.

展望：TDA 判别虫眼病害叶

健康的植物叶子没有虫眼，叶片内部无洞，边缘规则；若植物叶子被病虫侵害出现虫眼，叶片内部将出现洞，边缘缺失；采用叶片面积、纵横轴比、偏心率、矩形度等几何特征，难以判断有微小虫洞的叶片和健康无洞叶片；但这种形态变化可通过叶片数据集的 1 维同调分辨。后续，拓扑数据分析可用于叶片虫害检测、植物疾病防治的研究。

总结

- 我们用持续同调方法将植物的拓扑特征可视化，提取植物的 4 个拓扑特征和 6 个几何特征，并用随机森林、K-近邻算法对植物进行分类。但因提取的拓扑特征不完善，使得拓扑特征的分类效率不高。
- 本工作需要继续提取拓扑不变量，计划用高斯加权核函数描述植物的形态，将植物分类框架应用到植物叶片病害检测的研究中。