数学学院案例分析报告

案例名称：木槿花谱系图

报告负责人：魏丹怡（202032164）

完成时间：2021.04.15

目　录

一、研究背景与意义

1、研究背景

木槿属是一种历史久远的草本、灌木或乔木。叶互生，掌状分裂或不分裂，具掌状叶脉，具托叶。花两性，5 数，花常单生于叶腋间；小苞片 5 或多数，分离或于基部合生；花萼钟状，很少为浅杯状或管状，5 齿裂，宿存；花瓣 59 各色，基部与雄蕊柱合生；雄蕊柱顶端平截或 5 齿裂，花药多数，生于柱顶；子房 5 室，每室具胚珠 3 至多数，花柱 5 裂，柱头头状。蒴果胞背开裂成 5 果爿；种子肾形，被毛或为腺状乳突。

对于古生物木槿的系统发育树的构建一直是困扰生物学家的难题。随着统计与机器学习等高端技术的发展，我们可能利用现代技术解决这一难题。由于直接对于古生物木槿进行聚类，并没有一个标准进行参考，故我们先利用现代生物木槿先通过统计方法进行分类。通过对于按照各种特征组合以及采用各种分类方法的分类结果与其真实的生物学分类进行对比，找到最适合进行分类的方法以及特征组合，再将其运用在古生物木槿的分类上，这样的分类结果更有理论保障。

2、研究意义

共同祖先学说表明地球上的一切生命形式都有一个共同的起源，无论动物、植物、真菌、原生生物，还是原核生物，它们都藉由一部共同的进化历史而有着或近或远的关联，所有的生物都来自共同的祖先。分子生物学发现了所有的生物都使用同一套遗传密码，生物化学揭示了所有生物在分子水平上有高度的一致性。共同祖先学说是构建系统发育树的理论基础(phylogenetic tree)。重建所有生物的进化历史并以一种树状结构即系统发育树的形式来表示生物类群之间的进化关系， 一直是系统发育学研究的核心问题，也是进化生物学研究的重要内容之一。建立可靠的系统发育关系不仅是生物分类和命名的基础，也是阐明类群起源和扩散、探讨性状演化以及揭示物种形成机制的前提。因此，对于木槿属植物进行系统发育树聚类有着极大的研究意义。

二、数据来源和基本情况

1、数据来源

来自于生物学家对于木槿花的观察得到的现生木槿花形态学数据集。

2、基本情况

该数据集来源木槿属的形态学数据，描述了共 41 个分类群的 38 种形态学数据，其中 26 种木槿属植物，14 个变种木槿属植物，1 个外类群蜀菊。每行代表一个分类单元，每列为一个形态学特征。其中特征取值包含 0,1,2,3,?,N。数据集的 20 个特征具体展式如下：

表 1 案例特征

类型	序号	性状	性状特征及编码
整体	1	植物生活习性	木本(0);草本(1)
叶片	2	叶缘形状	全缘或近全缘(0);具锯齿(1)
	3	叶片形态	椭圆形或长圆形(0);心形或卵形(1)
	4	叶片具裂片	是(0);否(1)

	5	叶裂片形状	无(0);钝圆形(1);三角形或长圆形(2)
	6	叶片质地	纸质(0);坚纸质或革质(1);厚革质(2)
	7	叶基部形态	楔形(0);钝至阔楔形或圆形(1);圆形、截形或心形(2)
托片	8	托叶习性	早落(0);宿存(1)
	9	托叶形状	叶状或佛焰苞状(0)；线形(1)
花萼	10	花萼宿存	是(0)；否(1)
	11	萼形状	钟形(0)；杯形或浅杯状(1)；管状或筒状(2)
	12	花萼膨大	是(0)否(1)
茎	14	茎直立	是(0);否(1)
	15	茎具刺	是(0);否(1)
花梗	13	小枝具毛	是(0);否(1)
	19	花梗具毛	无毛(0);被硬毛(1);短柔毛(2)
	20	花梗长度	长于叶柄(0);短于叶柄(1)
	21	花梗长度	１～３ｃｍ(0);４～１３ｃｍ(1)
	22	花梗具节	是(0);否(1)
花	16	花序类型	圆锥花序(0);花单生(1)
	17	花形态	花直立(0);花下垂(1)
	18	花柱枝被毛	是(0);否(1)
	28	雄蕊伸出花外	是(0);否(1)
	29	花瓣边缘分裂情况	不分裂或微具缺刻(0)；分裂或深裂成流苏状(1)
	30	花瓣颜色	黄色(0)；白色(1)；紫色、红色(2)；多色(3)
	31	花瓣层数	单瓣(0)；重瓣(1)
	32	花瓣长度	≤５ｃｍ(0)；＞５ｃｍ(1)
苞片	23	小苞片形状	线形或披针形(0);卵形(1);匙形(2)
	24	小苞片长度	１～２ｍｍ(0);６～１５ｍｍ(1);１５～３０ｍｍ(2)
	25	总苞合生	分离或仅基部合生(0)；１／３～１／２处合生(1)
	26	小苞片具附属物	是(0)；否(1)
	27	小苞片颜色	绿色(0)；红色(1)
果	33	果皮具毛	无毛(0);被柔毛(1);被硬毛(2);混合毛(3)
	34	蒴果具喙	是(0);否(1)
	35	蒴果具翅	是(0);否(1)
种子	36	种子形状	肾形(0)；球形(1)
	37	种子具腺状乳突	是(0)；否(1)
	38	种子被毛	无毛(0)；短柔毛或棉毛(1)；被长粗毛(2)

三、数据预处理

1、数据展示

将整个数据集展示如图 2 所示，从表中可以看出，数据集存在很多没有观察到的特征即用“?”表示，还有这个数据中草木槿有个别性状并没有在该花中体现，所以记为“N”。

由于在进行下面步骤对数据进行聚类时，不能对于拥有这些缺失值或者异常值的点进

行聚类，因此我需先进行数据预处理，填补缺失值进而方便于我对于数据的进一步探索。

表 2 数据展示

编号	分类群	性状
1	大叶木槿	0 0 1 0 0 0 2 1 0 0 0 1 1 0 1 1 0 0 1 0 0 1 0 2 0 1 0 1 0 0 0 1 2 1 1 1 1 1
2	黄槿	0 0 1 0 0 1 2 1 0 0 0 1 0 0 1 1 0 1 0 0 0 1 0 1 1 1 0 1 0 0 0 0 1 0 1 1 0 0
3	高红槿	0 0 1 0 0 1 2 1 0 1 0 1 0 0 1 1 0 0 0 0 0 1 0 2 1 1 0 0 0 1 0 1 ? ? ? ? 0 ?
4	樟叶槿	0 0 0 0 0 1 1 1 1 0 0 1 0 0 1 0 0 0 1 0 1 1 0 1 1 1 0 1 0 0 0 1 0 1 1 1 1 1
5	滇南芙蓉	0 1 1 0 0 0 2 1 0 0 0 1 1 1 1 1 0 0 0 0 1 1 2 1 0 1 0 1 0 2 0 0 1 1 1 1 1 2
6	旱地木槿	0 1 1 0 0 2 2 0 1 0 1 1 1 0 1 0 0 1 0 0 1 0 1 1 0 1 0 1 0 2 0 0 1 1 1 1 1 1
7	光柱旱地木槿	0 1 1 0 0 0 2 0 1 0 1 1 1 0 1 0 0 0 0 0 1 0 1 1 0 1 0 1 0 2 0 0 1 1 1 1 1 1
8	海滨木槿	0 0 1 0 0 2 2 1 0 0 0 1 1 0 1 0 0 1 0 0 0 1 0 1 1 1 0 1 0 0 0 ? 3 1 1 1 0 0
9	吊灯扶桑	0 1 0 0 0 0 0 1 1 1 0 2 1 0 0 1 0 1 0 1 1 1 0 0 0 0 1 0 0 1 1 0 1 0 1 1 1 1 ?
10	朱槿	0 1 0 0 0 0 1 0 1 0 0 1 1 0 1 0 1 0 0 1 1 0 0 1 0 1 0 1 0 3 0 1 0 0 1 1 1 ?
11	重瓣朱槿	0 1 0 0 0 1 2 0 1 0 0 1 0 0 1 0 0 0 0 1 1 0 0 1 0 1 0 1 0 3 1 1 0 0 1 1 1 ?
12	庐山芙蓉	0 1 1 1 2 0 2 1 1 0 0 1 1 0 1 0 0 0 0 0 0 1 2 2 0 1 0 1 1 2 0 1 3 1 1 1 1 2
13	长柄庐山芙蓉	0 1 1 1 2 0 2 1 1 0 0 1 1 0 1 0 0 0 0 1 0 0 2 2 0 1 0 1 1 2 0 1 3 1 1 1 1 2
14	美丽芙蓉	0 1 1 1 2 0 2 1 1 0 1 1 1 0 1 0 0 0 0 1 1 0 2 2 0 1 0 1 0 3 0 1 2 1 1 1 1 1
15	全叶美丽芙蓉	0 1 1 0 1 0 2 1 1 0 1 1 1 0 1 0 0 0 0 1 1 0 2 2 0 1 0 1 0 3 0 0 2 1 1 1 1 1
16	台湾芙蓉	0 1 1 1 2 0 2 1 1 0 0 1 1 0 1 0 0 0 0 1 0 1 0 1 0 1 0 1 0 1 0 0 ? ? ? ? ? ?
17	木芙蓉	0 1 1 1 2 0 2 1 1 0 0 1 1 0 1 0 0 0 0 1 0 0 0 1 0 1 0 1 0 1 0 0 3 1 1 1 1 2
18	洋槿	0 0 0 0 0 0 1 1 1 0 0 1 1 0 1 0 0 ? 0 0 ? 0 2 1 ? 1 0 ? 0 0 0 1 0 1 1 1 1 1
19	贵州芙蓉	0 1 1 1 2 0 2 1 1 0 0 1 1 0 1 0 0 0 0 0 0 1 0 2 0 1 0 1 0 3 0 1 ? ? ? ? ? ?
20	木槿	0 1 0 0 0 1 0 0 1 0 0 1 1 0 1 0 0 1 0 0 0 1 0 1 0 1 0 1 0 1 0 0 0 1 1 1 1 2
21	长苞木槿	0 1 0 0 0 1 0 0 1 0 0 1 1 0 1 0 0 1 0 0 0 1 0 2 0 1 0 1 0 1 0 0 0 1 1 0 1 2
22	短苞木槿	0 1 0 0 0 1 0 0 1 0 0 1 1 0 1 0 0 1 0 0 0 1 0 0 0 1 0 1 0 1 0 0 0 1 1 0 1
23	百花重瓣木槿	0 1 0 0 0 1 0 0 1 0 0 1 1 0 1 0 0 1 0 0 0 1 0 1 0 1 0 1 0 2 1 0 0 1 1 0 1 2
24	粉紫重瓣木槿	0 0 0 0 0 1 0 0 1 0 0 1 1 0 1 0 0 1 0 0 0 1 0 1 0 1 0 1 0 1 1 0 0 1 1 0 1 2
25	雅致木槿	0 1 0 0 0 1 0 0 1 0 0 1 1 0 1 0 0 1 0 0 0 1 0 1 0 1 0 1 0 1 1 0 0 1 1 0 1 2
26	大花木槿	0 1 0 0 0 1 0 0 1 0 0 1 1 0 1 0 0 1 0 0 0 1 0 1 0 1 0 1 0 1 0 0 0 1 1 0 1 2
27	牡丹木槿	0 1 0 0 0 1 0 0 1 0 0 1 1 0 1 0 0 1 0 0 0 1 0 1 0 1 0 1 0 1 1 0 0 1 1 0 1 2
28	紫花重瓣木槿	0 1 0 1 2 1 0 0 1 0 0 1 1 0 1 0 0 1 0 0 0 1 0 1 0 1 0 1 0 1 1 0 0 1 1 0 1 2
29	百花单瓣木槿	0 1 0 1 2 1 0 0 1 0 0 1 1 0 1 0 0 1 0 0 0 1 0 1 0 1 0 1 0 2 0 0 0 1 1 0 1 2
30	华木槿	0 1 1 1 2 1 1 0 1 0 0 1 1 0 1 1 0 1 0 0 0 1 0 2 0 1 0 1 0 1 0 1 ? ? ? ? ? ?
31	光籽木槿	0 1 1 1 1 1 1 0 1 0 0 1 1 0 1 0 0 0 0 0 0 1 0 1 1 1 0 1 0 3 0 0 3 0 1 0 1 0
32	红秋葵	1 1 0 1 2 0 0 ? 1 0 0 1 0 0 1 0 0 1 0 1 0 1 0 2 0 1 0 1 0 1 0 1 0 0 1 1 1 2
33	芙蓉葵	1 1 0 0 0 0 1 1 1 0 0 1 0 0 1 0 0 0 0 1 0 0 0 0 0 1 0 1 0 3 0 1 0 1 1 0 1 0
34	云南芙蓉	1 1 1 0 0 ? 1 1 1 0 1 1 1 0 1 0 0 ? 0 0 0 1 0 1 0 1 0 1 0 0 ? 0 2 1 0 0 0 0
35	刺芙蓉	1 1 1 1 2 ? 0 0 0 0 1 1 1 0 0 0 0 ? 0 0 0 1 0 1 0 0 0 ? 0 0 ? 0 2 0 1 0 1 2
36	辐射刺芙蓉	1 1 1 1 2 ? 1 0 1 0 0 1 1 0 0 0 0 1 0 0 0 0 0 2 0 0 0 1 0 1 0 1 2 0 1 2 1 1
37	野西瓜苗	1 1 1 ? 2 ? ? 0 1 0 0 0 1 1 1 0 0 1 0 0 0 1 0 1 0 1 0 1 0 0 0 0 2 1 1 0 0 0
38	玫瑰茄	1 1 1 ? 1 ? 1 0 1 0 1 1 1 0 1 0 0 ? ? 0 0 0 0 1 0 0 1 ? 0 0 0 1 2 0 1 0 1 0
39	大麻槿	1 1 1 ? 2 1 1 0 1 0 0 1 0 0 0 0 0 1 1 0 0 1 0 1 0 1 0 1 0 0 0 1 2 0 1 0 1 0
40	草木槿	1 1 1 ? 2 0 1 0 1 0 1 1 1 0 1 1 0 1 0 1 0 0 — — — — — 1 0 0 0 0 2 0 1 ? 1 1
41	外类群蜀葵	1 1 1 1 2 0 1 0 0 0 1 1 1 0 1 1 0 0 1 0 0 1 2 0 ? 1 0 1 0 3 0 0 1 1 1 0 1 0

2、缺失值处理

缺失值处理有多种方式，常见的方式有：删除、插补、特殊值填充、平均值（众数、中位数）填充、就近补齐以及 K 最近距离邻法等方法。

针对于本数据集，由于我本身做的是一个对于木槿花的分类问题，故不能运用 K 最近距离邻法，又因为数据的上下两个特征并没有很大的联系，所以就近补齐法也不能采用。同时，由于木槿花数据本来的特征就不够多，所以删除并不是一种很好的方法，并且我对于木槿花的相关生物知识并没有充分的了解，所以特殊值填充法也无法发挥作用。因此我打算利用每个特征的统计量进行插补填充。又因为每个特征并非数值型变量，故平均数填充法不靠谱，我准备采用众数进行填充，因为如果更多的花具有这个特征，极大概率可以认为确实特征的木槿花也具有相同的特征。

针对于没有相应特征的草木槿而言，打算采用两种方法进行对比。第一是同样利用众数填充，其二是删除该花。将对比两种方法对于所有花统一的分类情况，再决定利用哪一种最后对于木槿花进行聚类。

四、系统发育树的构建

通过层次聚类的方法进行系统发育树的构建，同时将聚类结果是否为木本的分类准确性作为评价所构建的系统发育树优良性的依据，对聚类结果进行评价。

1、形式化和分类器数学推导

（1）聚类分析谱系图

二叉树是一种常见的数据结构，二叉排序树的建立方法就是将待插入的数据项与树根的结点值做比较，若前者小于后者，则进入左子树，否则进入右子树;在子树中又与子树根比较，如此进行下去，到达终结点后，插入该数据项，由此建立起排序的二叉树。该类树左子树所有结点都小于等于根，右子树所有结点都大于等于根，采用中序遍历(左、根、右的顺序)后，得出由小到大排列的一组结点值。

聚类分析谱系图具有二叉树的结构(但未画出终结点以外的结点值和根)，因为聚类分析谱系图是根据联结表中类别之间的相互关系按照一定规则绘制的，所以通过分析联结表应该能够建立起谱系图的二叉树.在聚类分析中，由距离系数矩阵形成联结表的过程是：

[1]在距离系数矩阵中选择最小值，将最小值和与其对应的两类的类序号一起写入联结表；

[2]按照“保留小号，划掉大号”的原则，重新计算这两类中较小的类序号所在行、列的距离系数值，划掉这两类中较大类序号所在的行和列，形成新的距离系数矩阵；

[3]重复进行 1,2 步骤，直到矩阵中剩下最后两类，一起写入联结表。

（2）层次聚类

层次聚类通过对数据集在不同层次进行划分，从而形成树形的聚类结构。数据集的划分可采用“自底向上”的聚合（agglomerative）策略，也可采用“自顶向下”的分拆（divisive）策略。“自底而上”的算法开始时把每一个原始数据看作一个单一的聚类簇，然后不断聚合小的聚类簇成为大的聚类。“自顶向下”的算法开始把所有数据看作一个聚类，通过不断分割大的聚类直到每一个单一的数据都被划分。

凝聚型层次聚类的策略是先将每个对象作为一个簇，然后合并这些原子簇为越来越大的

簇，直到所有对象都在一个簇中，或者某个终结条件被满足。最小距离的凝聚层次聚类算法流程：

[1] 将每个对象看作一类，计算两两之间的最小距离；

[2] 将距离最小的两个类合并成一个新类；

[3] 重新计算新类与所有类之间的距离；

[4] 重复[2]、[3]，直到所有类最后合并成一类。

（3）欧几里得距离

对于层次聚类的两类之间的距离，我们采用欧几里得距离。

n 维空间的公式：

$$d(x,y):=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2}$$

2、评估方法

（1）混淆矩阵

混淆矩阵向我们展式了查准率(准确率)与查全率(召回率)：

查准率(P) = $\frac{TP}{TP+FP}$ ，即在被判别为正类别的样本中，确实为正类别的比例是多少；

查全率(R) = $\frac{TP}{TP+FN}$ ，即在所有正类别样本中，被正确判别为正类别的比例是多少。

（2）评价指标

将聚类结果与已知植物生活习性（草本/木本）进行对比并绘制混淆矩阵，计算准确率，以此评价聚类结果是否良好。选择准确率最高的一组特征作为判断木槿花类型的指标。

五、结果分析

我们使用木槿花的全部特征进行聚类分析，然后将木槿花特征分为叶片、托片、花萼、茎、花梗、苞片、花、果、种子，共 9 组，分别进行聚类分析。

1、全部特征

由于草木槿的特征中存在不适用数据，因此我们使用了两种方法来处理缺失值问题：

（1）视为缺失值，按照众数填补；

（2）删除不适用数据，即不考虑草木槿数据。

聚类结果如下：

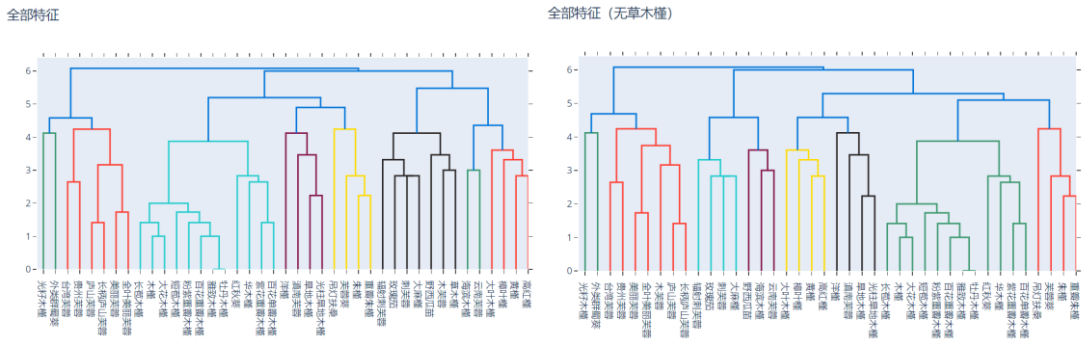

图 1　全部特征聚类结果

绘制混淆矩阵如下：

0	23	7
1	9	1
	0	1

此模型将本应分类为木本植物（标签为 0）的 7 个样本误分为草本植物（标签为 1），同时将本应分类为草本植物（标签为 1）的 9 个样本误分为木本植物（标签为 0）。按全部特征聚类得到的准确率为 60%。

2、叶片

在叶片相关特征中，不存在不适用数据，因此按照众数填补缺失值，聚类结果如下：

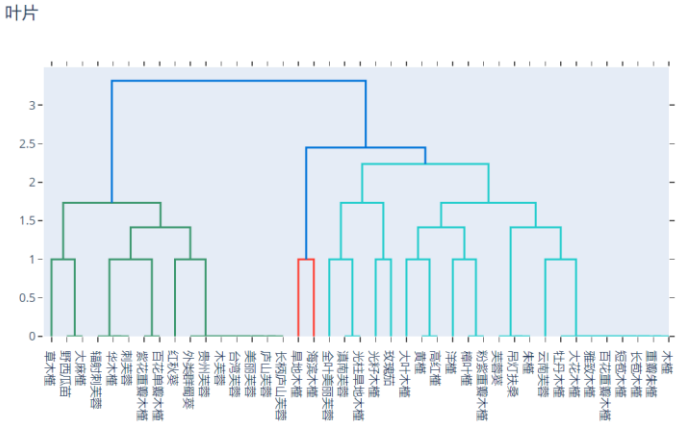

图 2　叶片特征聚类结果

绘制混淆矩阵如下：

0	18	12
1	3	7
	0	1

此模型将本应分类为木本植物（标签为 0）的 12 个样本误分为草本植物（标签为 1），同时将本应分类为草本植物（标签为 1）的 3 个样本误分为木本植物（标签为 0）。按全部

特征聚类得到的准确率为 62.5%。

3、托片

在托片相关特征中，不存在不适用数据，因此按照众数填补缺失值，聚类结果如下：

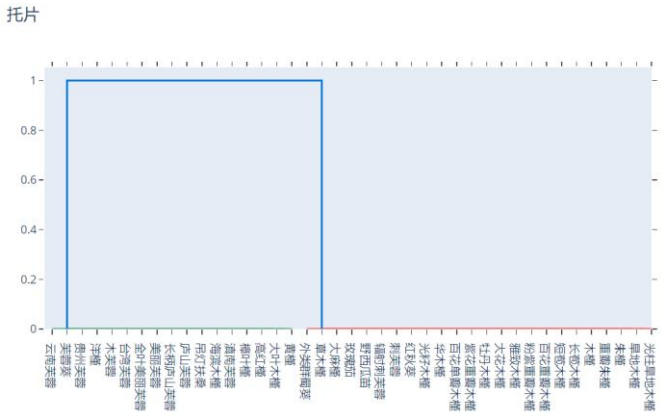

图 3　托片特征聚类结果

绘制混淆矩阵如下：

0	14	16
1	2	8
	0	1

此模型将本应分类为木本植物（标签为 0）的 16 个样本误分为草本植物（标签为 1），同时将本应分类为草本植物（标签为 1）的 2 个样本误分为木本植物（标签为 0）。按全部特征聚类得到的准确率为 55%。

4、花萼

在花萼相关特征中，不存在不适用数据，因此按照众数填补缺失值，聚类结果如下：

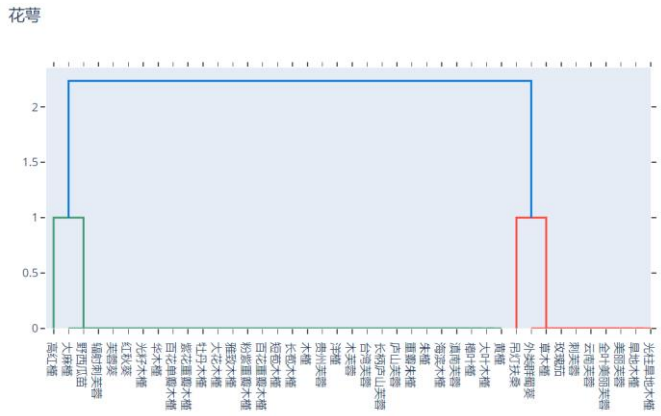

图 4　花萼特征聚类结果

绘制混淆矩阵如下：

0	25	5
1	5	5
	0	1

此模型将本应分类为木本植物（标签为 0）的 5 个样本误分为草本植物（标签为 1），同时将本应分类为草本植物（标签为 1）的 5 个样本误分为木本植物（标签为 0）。按全部特征聚类得到的准确率为 75%。

5、茎

在茎相关特征中，不存在不适用数据，因此按照众数填补缺失值，聚类结果如下：

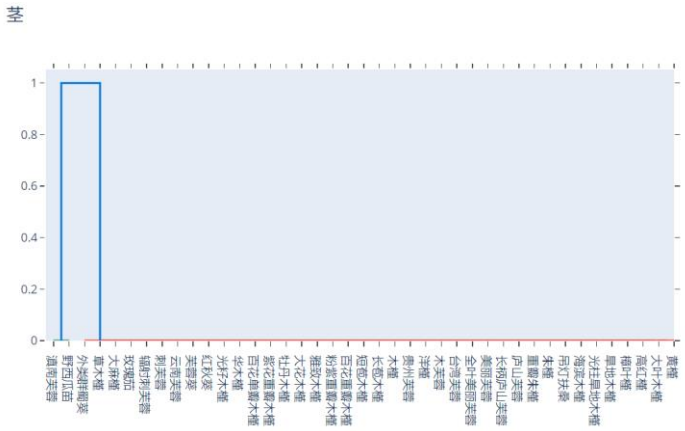

图 5　茎特征聚类结果

绘制混淆矩阵如下：

0	29	1
1	9	1
	0	1

此模型将本应分类为木本植物（标签为 0）的 1 个样本误分为草本植物（标签为 1），同时将本应分类为草本植物（标签为 1）的 9 个样本误分为木本植物（标签为 0）。按全部特征聚类得到的准确率为 75%。

6、花梗

在花梗相关特征中，不存在不适用数据，因此按照众数填补缺失值，聚类结果如下：

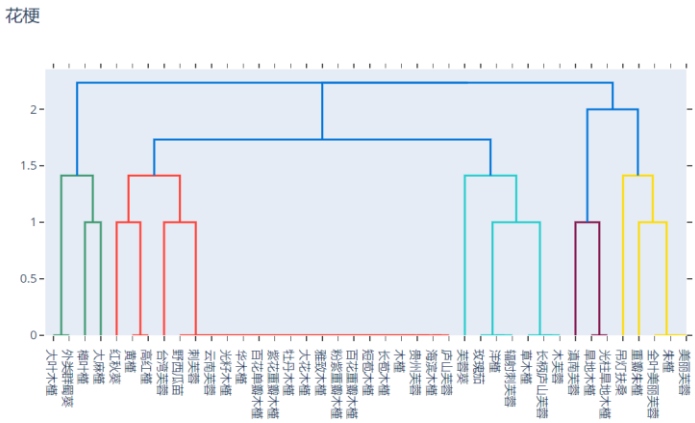

图 6　花梗特征聚类结果

绘制混淆矩阵如下：

0	21	9
1	7	3
	0	1

此模型将本应分类为木本植物（标签为 0）的 9 个样本误分为草本植物（标签为 1），同时将本应分类为草本植物（标签为 1）的 7 个样本误分为木本植物（标签为 0）。按全部特征聚类得到的准确率为 60%。

7、苞片

由于草木槿的特征中存在不适用数据，因此我们使用了两种方法来处理缺失值问题：

（1）视为缺失值，按照众数填补；

（2）删除不适用数据，即不考虑草木槿数据。

聚类结果如下：

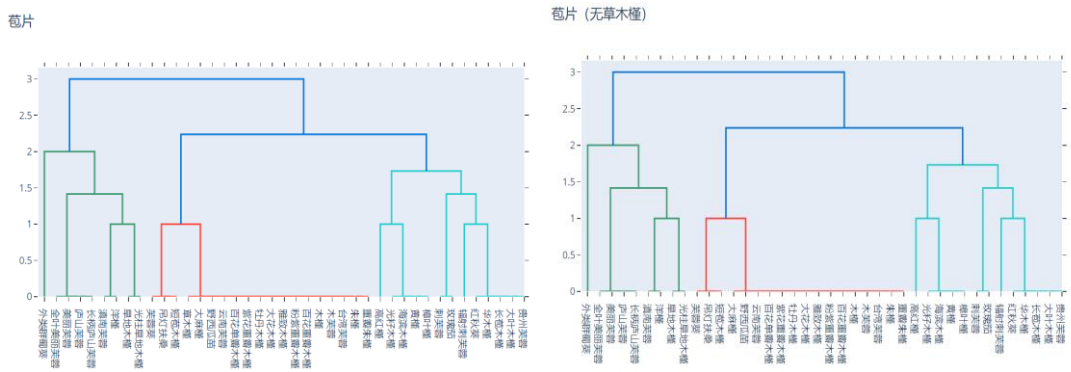

图 7　苞片特征聚类结果

绘制混淆矩阵如下：

0	22	8
1	9	1
	0	1

此模型将本应分类为木本植物（标签为 0）的 8 个样本误分为草本植物（标签为 1），同时将本应分类为草本植物（标签为 1）的 9 个样本误分为木本植物（标签为 0）。按全部特征聚类得到的准确率为 57.5%。

8、花

在花相关特征中，不存在不适用数据，因此按照众数填补缺失值，聚类结果如下：

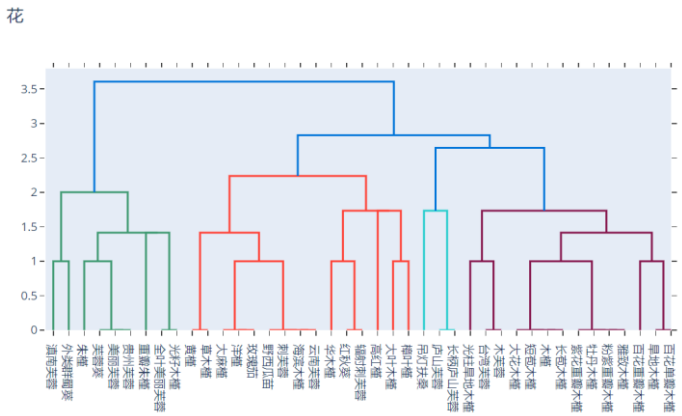

图 8　花特征聚类结果

绘制混淆矩阵如下：

0	23	7
1	8	2
	0	1

此模型将本应分类为木本植物（标签为 0）的 7 个样本误分为草本植物（标签为 1），同时将本应分类为草本植物（标签为 1）的 8 个样本误分为木本植物（标签为 0）。按全部特征聚类得到的准确率为 62.5%。

9、果

在果相关特征中，不存在不适用数据，因此按照众数填补缺失值，聚类结果如下：

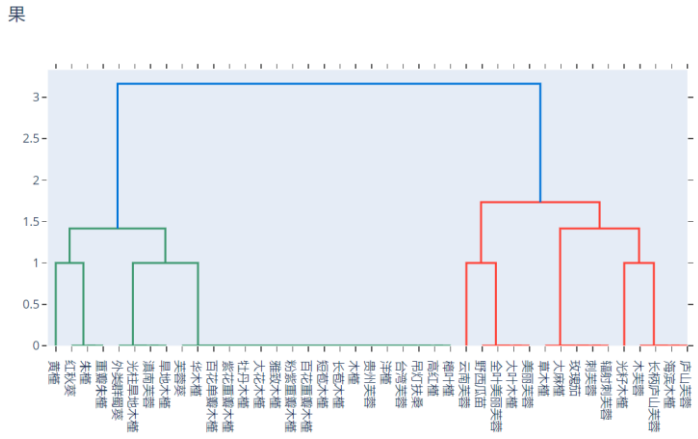

图 9　果特征聚类结果

绘制混淆矩阵如下：

0	22	8
1	3	7
	0	1

此模型将本应分类为木本植物（标签为 0）的 8 个样本误分为草本植物（标签为 1），同时将本应分类为草本植物（标签为 1）的 3 个样本误分为木本植物（标签为 0）。按全部特征聚类得到的准确率为 72.5%。

10、种子

在种子相关特征中，不存在不适用数据，因此按照众数填补缺失值，聚类结果如下：

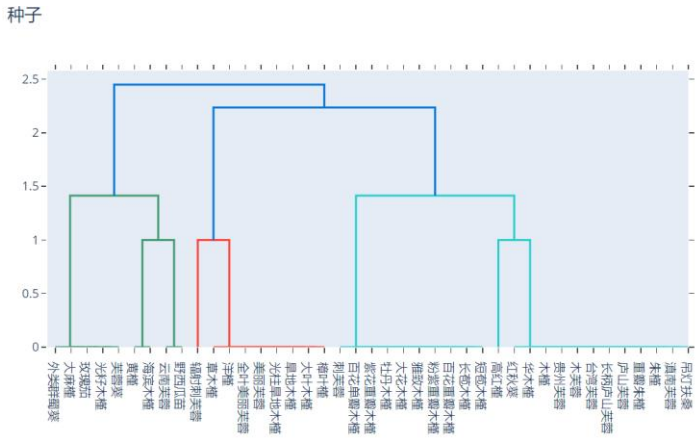

图 10　种子特征聚类结果

绘制混淆矩阵如下：

0	27	3
1	4	6
	0	1

此模型将本应分类为木本植物（标签为 0）的 3 个样本误分为草本植物（标签为 1），同时将本应分类为草本植物（标签为 1）的 4 个样本误分为木本植物（标签为 0）。按全部特征聚类得到的准确率为 82.5%。

六、结论

通过对比选取不同特征时的聚类结果准确率，我发现使用种子的相关特征进行聚类时，得到的结果最好，茎和花萼的聚类结果较好，其次为果的聚类结果也较为良好。这意味着不同类型木槿花的种子很好区分，在种植时较好分类；当木槿花开花后，通过对比其茎和花萼的特征，也可以较为容易地判断其类型；最后，当木槿花结果时，也可以观察其果实判断木槿花类型。

按照准确率最高地聚类结果绘制生物进化谱系树，结果如下：

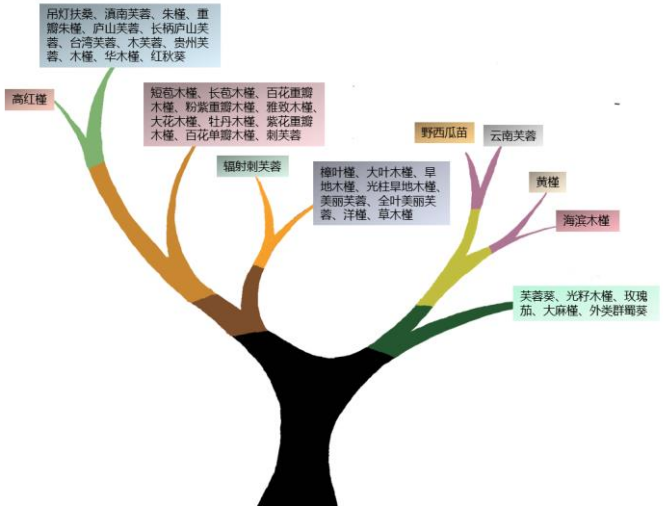

图 11　生物进化谱系树

七、实践总结

本次实践利用统计方法对于无标签木槿花数据集进行系统发育树的构建。在分析过程中，我进行了数据预处理，学习了处理缺失值的相关方法，并利用 python 对于缺失数据进行处理，并对处理后的数据进行聚类分析。在实现系统发育树的构建的过程中，我加深了对无监督学习聚类分析的理论知识的理解，同时在不断的尝试与纠错中，增加了许多编程经验。

与此同时也存在着不足，在对聚类后的系统发育树进行评价时，由于缺乏相关的生物学基础，导致只能运用数据集中所给的的信息对于聚类结果进行评价，不够严谨。这启示如若将来要从事某方面数据分析，应先系统学习该领域知识，从而使我对于数据分析后的结果更具有可用性。