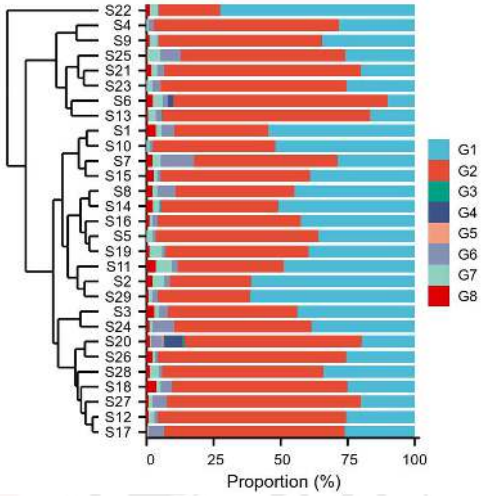
基础绘图 - 聚类叠加柱状图



网址：https://www.xiantao.love

更新时间：2023.06.14

目录

基本概念

- 聚类(Clustering)：是按照某个特定标准(如距离)把一个数据集分割成不同的类或簇，使得同一个簇内的数据对象的相似性尽可能大，同时不在同一个簇中的数据对象的差异性也尽可能地大堆叠（叠加）。
  - 层次聚类：常用且层次聚类算法对时间和空间需求很大
  - 丰度聚类：来源于生态学 vegan 包 vegdist 函数，包含多种生态常用距离算法
- 聚类树：一种展现有群组、层次关系的比例数据的一种分析工具
- 柱状图：用柱子的高度或者柱子的相对高度来表示数据的大小情况
- 叠加柱状图：
  - 叠加比例柱状图：用于查看不同分类中 分组的组成比例情况
  - 叠加数值柱状图：用于查看不同分类中 分组数值的差异。与叠加比例柱状图的差别在于：叠加比例柱状图每组都会计算每个分组的比例情况

应用场景

用聚类树与柱状图的结合来展示分类与分类之间、分类与分组之间的相关信息，不仅能反映分类之间的的相似性（聚类情况），也可以展示在分类中各分组的组成信息

分析过程

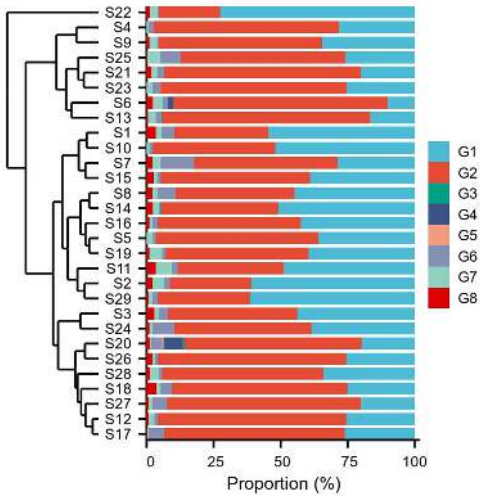
上传数据 ⟶ 数据处理(清洗) ⟶ 分析 ⟶ 可视化

- 数据格式：（具体数据格式要求可以看后面过程的“数据格式”部分）
  - 数据第 1 列为分类类型，对应分组信息（图注-颜色映射）
    - 至少需要提供 2 个不同的分类
  - 数据第 2 列及以后都需要是数值类型数据
    - 从第 2 列开始所有的变量只能是数值，不能含有非数值类型数据，或者混合数值与非数值类型数据
  - 不能含有无法识别的特殊字符或者是非字符

	A	B	C	D	E	F	G	H	I	J	K
1	group	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
2	G1	7.778407912	14.32947477	11.98878983	10.22245185	6.934812539	2.818661038	3.710862683	9.291028609	13.92487537	5.835709865
3	G2	5.013282687	7.086140197	13.2220069	24.89810779	11.76927704	22.57257131	6.864973374	9.136603542	24.50068057	5.104992259
4	G3	0.00765304	0.015978322	0.00672803	0.00797679	0	0	0.023087251	0	0	0
5	G4	0	0	0	0	0	0.620815021	0	0	0	0
6	G5	0	0	0	0	0	0	0	0	0	0
7	G6	0.6648926	0.494610785	0.938927105	0.684834611	0.204450779	0.501714813	1.589907047	1.389038643	0.262068262	0.168065912
8	G7	0.304751778	1.060454937	0.401875414	0.264703205	0.466709948	1.046769816	0.408603421	0.385717837	1.10175174	0.112449843
9	G8	0.483464241	0.504697998	0.802832047	0	0	0.664246525	0.270090224	0.458932797	0.448164288	0

- 数据处理：对上传数据各列数据进行相关处理
  - 对上传数据第 1 列分类类型数据，第 2 列及以后各列数值类型数据进行相应处理
- 分析：对处理后的数据进行聚类分析
- 可视化：将分析得到的结果数据进行 ggplot2 包可视化

结果解读



左侧为聚类树状图（从右往左看）：

- 聚类树状图纵坐标表示样本/变量/分子（对应上传数据除了第 1 列外的每一列）；横向的线表示分类所对应的相对距离
- 横线表示从纵向坐标最低端（每个分类）开始将最近的两个分类聚为一类，然后将其看作一个整体计算与其它分类之间的距离，继续聚类，直至所有的分类都被聚为一类
- 样本之间的连线（竖线）表示其对应的分类都被聚为一类，有多少条连线就表示经过多少次聚类

右侧为堆叠柱状图（从左往右看）：

- 堆叠柱状图纵坐标表示样本/变量/分子（对应上传数据除了第 1 列外的每一列）；横向坐标表示各分组的百分比（默认）或频数值；图中直接展示了每个分组在分类中的所占比例或具体数值
- 可以直观比较不同分类中不同分组的占比情况
- 一种颜色表示一个分组

数据格式

	A	B	C	D	E	F	G	H	I	J	K
1	group	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
2	G1	7.778407912	14.32947477	11.98878983	10.22245185	6.934812539	2.818661038	3.710862683	9.291028609	13.92487537	5.835709865
3	G2	5.013282687	7.086140197	13.2220069	24.89810779	11.76927704	22.57257131	6.864973374	9.136603542	24.50068057	5.104992259
4	G3	0.00765304	0.015978322	0.00672803	0.00797679	0	0	0.023087251	0	0	0
5	G4	0	0	0	0	0	0.620815021	0	0	0	0
6	G5	0	0	0	0	0	0	0	0	0	0
7	G6	0.6648926	0.494610785	0.938927105	0.684834611	0.204450779	0.501714813	1.589907047	1.389038643	0.262068262	0.168065912
8	G7	0.304751778	1.060454937	0.401875414	0.264703205	0.466709948	1.046769816	0.408603421	0.385717837	1.10175174	0.112449843
9	G8	0.483464241	0.504697998	0.802832047	0	0	0.664246525	0.270090224	0.458932797	0.448164288	0

数据要求：

- 数据至少 2 列以上，每列至少 2 个观测，最多支持 50 列和 30 行数据
- 第 1 列为分类类型，表示分组
  - 至少需要提供 2 个不同的分类
  - 不能含有重复的分组名
  - 第 1 列作为分组，其排列的顺序与上传数据中的顺序一致，如果需要调整，需要在上传数据之前修改后再上传数据进行处理分析
- 除第 1 列外，从第 2 列开始每列数据代表一个分类(样本)，列名即为堆叠柱状图的横向坐标轴刻度名，分类不能重复（每一列数据的列名不能重复）
  - 每 1 行表示一个观测，不能含有小于 0 的数
  - 不能含有非数值类型数据，或者混合数值与非数值类型数据
- 不能含有无法识别的特殊字符或者是非字符

参数说明

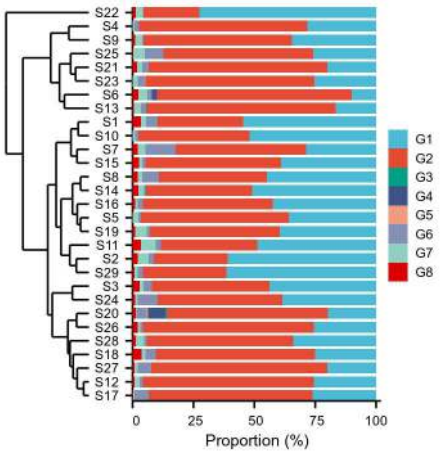
（说明：标注了颜色的为常用参数。）

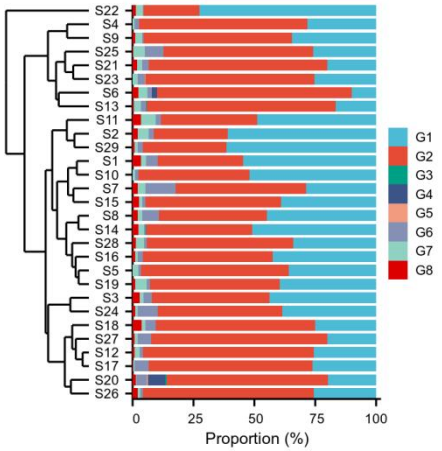
数据处理

- 归一化：可以选择是否对上传数据进行归一化处理，默认不归一化，还可以选择对行归一化、对列归一化

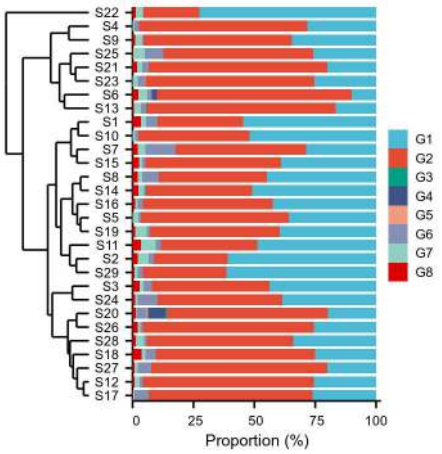
聚类

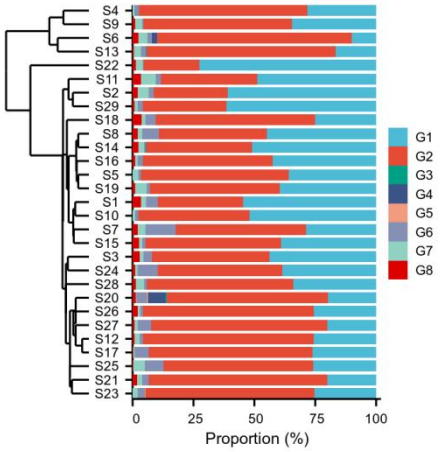
- 类型：可以选择聚类的类型，默认选择常用于一般数据特征的层次聚类-欧氏距离，计算距离的方法默认欧氏距离，其他常用的方法有：曼哈顿距离、堪培拉距离等，如下：



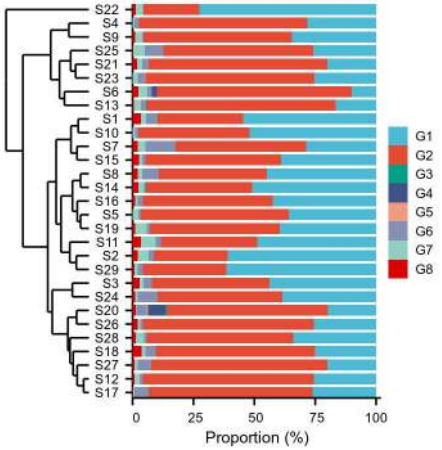


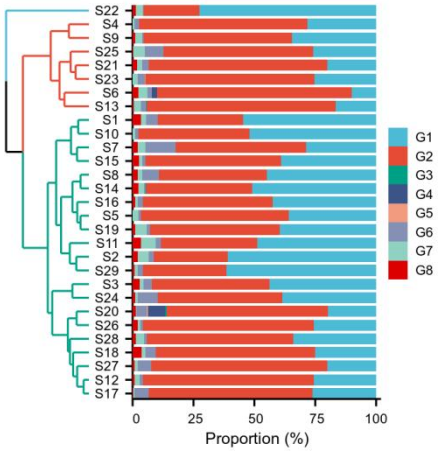
- 方法：可以选择聚类的方法，默认选择类平均法，也可选择常用中间距离法、最长距离法、最短距离法等，如下：





- 切割类型：可以选择对数据进行切割（分组），默认不切分，可选择不切分、切分 2 类、切分 3 类等，如下：





柱

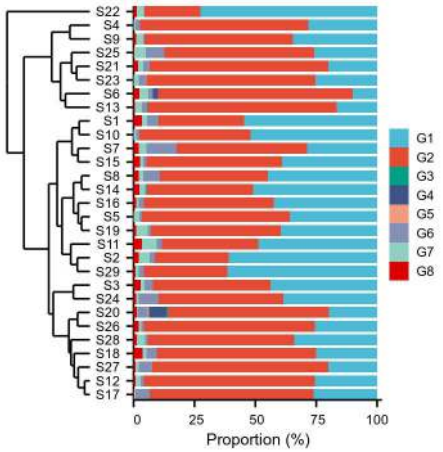
- 堆叠类型：可以选择绘制柱状图的样式，默认为百分比堆叠样式（堆叠比例柱状图），还可以选择频数(数值)堆叠样式（堆叠数值柱状图）
- 填充色：可以修改绘制柱状图的填充颜色
- 描边色：可以修改绘制柱状图的描边颜色
- 描边粗细：可以选择并修改柱状图外框的粗细
- 宽度：可以选择柱状图的每一根柱子的宽度
- 不透明度：可以修改柱状图每一根柱子的不透明度

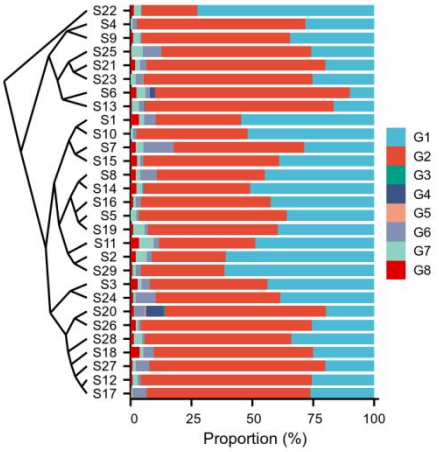
连线

- 展示：可以选择是否对柱状图之间进行连线操作，
  - 选择展示：连线的宽度与其连接的柱状图的宽度一致
- 不透明度：首先选择展示，则可以修改柱状图间连线的不透明度

样式

- 连线样式：可以选择并修改聚类树各线条的连线样式，如下：





线

- 颜色：当选择对数据进行聚类分群(切分)操作时，可以修改各个分群的颜色
- 线条类型：可以选择聚类树各分类表示的线条（竖线与横线）用实线或者虚线绘制
- 线条粗细：可以选择聚类树各分类表示的线条（竖线与横线）粗细
- 不透明度：可以修改聚类树各分类表示的线条（竖线与横线）的不透明度，1 表示完全不透明，0 表示完全透明

标题文本

- 大标题：大标题文本
- x 轴标题：x 轴标题文本
- y 轴标题：y 轴标题文本

补充：在要换行的中间插入\n。如果需要上标，可以用两个英文输入法下的大括号括住，比如 {{2}}；如果需要下标，可以用两个英文输入法下的中括号括住，比如 [[2]]

图注（Legend）

- 展示：可以选择是否展示图注操作
- 图注标题：首先选择展示，则可以修改需要上传的图注标题信息
- 图注位置：首先选择展示，则可以选择展示图注的位置

坐标轴

- x 轴标注旋转：可以选择分类表示的横向坐标轴（x 轴）标注旋转的角度
- y 轴范围+刻度：可以控制 y 轴范围和刻度，可只提供 2 个值来控制范围。形如 0.1，0.2，0.3（最小值和最大值不能超过可视化数据范围 20%，如果调整过大可能会无作用）

风格

- xy 颠倒：可以选择是否进行 xy 颠倒的操作
- 文字大小：控制整体文字大小，默认为 6pt

图片

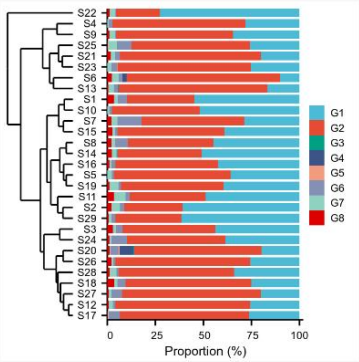
- 宽度：图片横向长度，单位为 cm
- 高度：图片纵向长度，单位为 cm
- 字体：可以选择图片中文字的字体

结果说明

主要结果

聚类叠加柱状图

聚类叠加柱状图: 用叠加柱状图来展示结果以及数据的组成情况，用聚类的方法展示样本间的聚类效果



聚类叠加柱状图.pdf　聚类叠加柱状图.tiff　聚类叠加柱状图.pptx

叠加柱状图部分：

· 横坐标（默认）表示各样本在各分组中的值/所占比值

· 纵坐标表示样本/变量/分子（对应上传数据除了第1列外的每一列）

聚类树部分：

方法学

统计分析和可视化均在R 4.2.1版本中进行

涉及的R包：ggplot2包（用于可视化）

处理过程：

(1) 使用dist函数计算各分类之间的距离

(2) 使用hclust函数构建分类之间的聚类模型

(3) 使用ggplot2包对聚类模型进行可视化

(4) 将清洗后的数据进行统计分析，得到各分类中各分组的占比

(5) 使用ggplot2包对占比数据（或原始数据）进行可视化

如何引用

生信工具分析和可视化用的是 R 语言，可以直接写自己用 R 来进行分析和可视化即可，可以无需引用仙桃，如果想要引用仙桃，可以在致谢部分（Acknowledge）致谢仙桃学术（www.xiantao.love）。

方法学部分可以参考对应说明文本中的内容以及一些文献中的描述。

常见问题

1. 聚类方法的选择?

答: 一般常用层次聚类, 除个别数据集 (如: 菌群) 使用丰度聚类, 不提供 kmeans 等聚类方法（速度慢，内存消耗较大）

2. 第一列为数值类型的数据可不可以?

答：不可以。第一列作为分组信息，需要是分类类型的数据