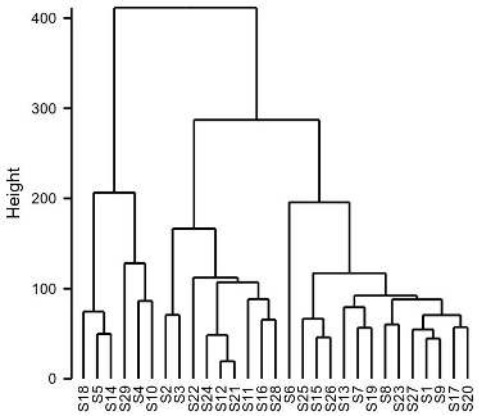
表达差异 - 聚类树状图



网址：https://www.xiantao.love

更新时间：2023.03.17

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：比较常用。距离算法包含有欧式距离等距离算法
  - 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

应用场景

聚类树状图主要以树状图的形式将进行聚类分析的分类（变量/分组）进行可视化，看分类之间的聚类情况

分析过程

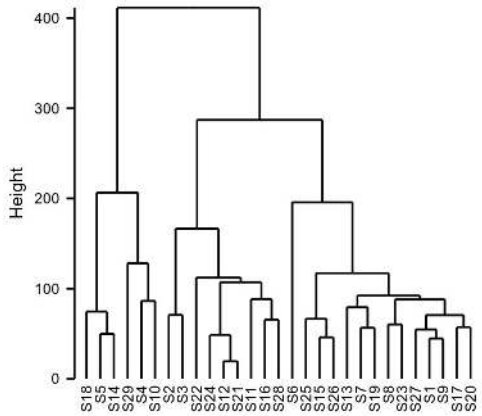
上传数据 ⟶ 数据处理(清洗) ⟶ 分析 ⟶ 可视化

- 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 列可以为分类类型也可以为数值类型数据
  - 数据第 2 列及以后都需要是数值类型数据

	A	B	C	D	E	F	G	H	I	J	K	L
1	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
2	0.00765304	0.015978322	0.00672803	0.00797679	0	0	0.023087251	0	0	0	0	0
3	7.778407912	14.32947477	11.98878983	10.22245185	6.934812539	2.818661038	3.710862683	9.291028609	13.92487537	5.835709865	12.47712847	4.946892192
4	0	0	0	0	0	0.620815021	0	0	0	0	0	0
5	5.013282687	7.086140197	13.2220069	24.89810779	11.76927704	22.57257131	6.864973374	9.136603542	24.50068057	5.104992259	10.14434879	13.64310724
6	85.80738601	36.25800351	40.58288461	46.47828169	57.21732036	21.68081293	36.14394143	46.91355377	53.40525194	46.30575277	48.55400313	63.93272425
7	0	0	0	0	0	0	0	0	0	0	0	0
8	0.6648926	0.494610785	0.938927105	0.684834611	0.204450779	0.501714813	1.589907047	1.389038643	0.262068262	0.168065912	0.504009848	0.278971274
9	636.1518011	202.0575591	244.2340909	978.934297	828.89732	627.5859501	609.79123	703.7965205	626.8261401	959.827265	366.3832583	418.908692
10	0.304751778	1.060454937	0.401875414	0.264703205	0.466709948	1.046769816	0.408603421	0.385717837	1.10175174	0.112449843	1.605588026	0.371479258
11	0.483464241	0.504697998	0.802832047	0	0	0.664246525	0.270090224	0.458932797	0.448164288	0	0.84904552	0.168377724
12	0	0	0	0	0	0	0	0.03892519	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0.401135166
14	138.6828532	62.76122836	97.38042908	109.6563671	53.7838189	239.7255678	79.43004548	81.67579243	149.6515304	100.9789216	184.2103069	125.2416676
15	0.594589869	0.716197229	0.048251335	0.276500769	0.966616155	0.216799453	0.487525157	0.312601737	1.175277084	0.25315005	0.276019576	0.238938256
16	19.3618486	0.312065174	0.788412073	79.3335556	30.05654546	146.2943489	0.346851039	12.48577472	36.73828726	0.407275768	1.552919802	3.940216705

- 数据处理：对上传数据各列数据进行相关处理
  - 如果第 1 列为数值类型，则从第 1 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
  - 如果第 1 列为分类类型，则从第 2 列开始所有的变量只能是数值
  - 不能含有无法识别的特殊字符或者是非字符
- 分析：对上传数据进行聚类分析
- 可视化：将分析得到的结果数据进行 ggplot2 包可视化

结果解读



- 聚类树状图横向坐标表示变量/分类，对应上传数据每一列(如果上传数据第 1 列为分类类型则为从第 2 列开始的每一列)
- 纵向坐标表示变量/分类所对应的相对距离
- 竖线表示从横向坐标最低端（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。
- 分类之间的连线（横线）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。

数据格式

	A	B	C	D	E	F	G	H	I	J	K	L
1	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
2	0.00765304	0.015978322	0.00672803	0.00797679	0	0	0.023087251	0	0	0	0	0
3	7.778407912	14.32947477	11.98878983	10.22245185	6.934812539	2.818661038	3.710862683	9.291028609	13.92487537	5.835709865	12.47712847	4.946892192
4	0	0	0	0	0	0.620815021	0	0	0	0	0	0
5	5.013282687	7.086140197	13.2220069	24.89810779	11.76927704	22.57257131	6.864973374	9.136603542	24.50068057	5.104992259	10.14434879	13.64310724
6	85.80738601	36.25800351	40.58288461	46.47828169	57.21732036	21.68081293	36.14394143	46.91355377	53.40525194	46.30575277	48.55400313	63.93272425
7	0	0	0	0	0	0	0	0	0	0	0	0
8	0.6648926	0.494610785	0.938927105	0.684834611	0.204450779	0.501714813	1.589907047	1.389038643	0.262068262	0.168065912	0.504009848	0.278971274
9	636.1518011	202.0575591	244.2340909	978.934297	828.89732	627.5859501	609.79123	703.7965205	626.8261401	959.827265	366.3832583	418.908692
10	0.304751778	1.060454937	0.401875414	0.264703205	0.466709948	1.046769816	0.408603421	0.385717837	1.10175174	0.112449843	1.605588026	0.371479258
11	0.483464241	0.504697998	0.802832047	0	0	0.664246525	0.270090224	0.458932797	0.448164288	0	0.84904552	0.168377724
12	0	0	0	0	0	0	0	0.03892519	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0.401135166
14	138.6828532	62.76122836	97.38042908	109.6563671	53.7838189	239.7255678	79.43004548	81.67579243	149.6515304	100.9789216	184.2103069	125.2416676
15	0.594589869	0.716197229	0.048251335	0.276500769	0.966616155	0.216799453	0.487525157	0.312601737	1.175277084	0.25315005	0.276019576	0.238938256
16	19.3618486	0.312065174	0.788412073	79.3335556	30.05654546	146.2943489	0.346851039	12.48577472	36.73828726	0.407275768	1.552919802	3.940216705

数据要求：

- 数据至少 3 列及以上，每列至少 4 个观测(4 行数据)，最多支持 100 列、60000 行数据
  - 数据第 1 列可以为分类类型也可以为数值类型数据，数据第 2 列及以后都需要是数值类型数据
    - 如果第 1 列为数值类型，则从第 1 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
    - 如果第 1 列为分类类型，则从第 2 列开始所有的变量只能是数值
  - 不能含有无法识别的特殊字符或者是非字符
- 每列数据为一个分类，每一列列名即为堆叠柱状图的横向坐标轴刻度名。
- 变量/分类不能重复（每一列数据的列名不能重复）

参数说明

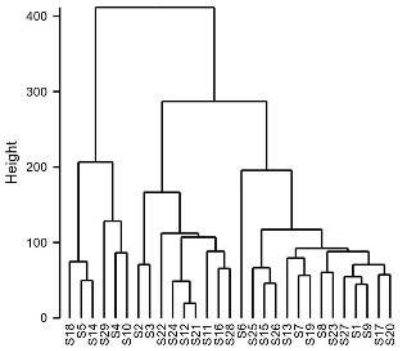
（说明：标注了颜色的为常用参数。）

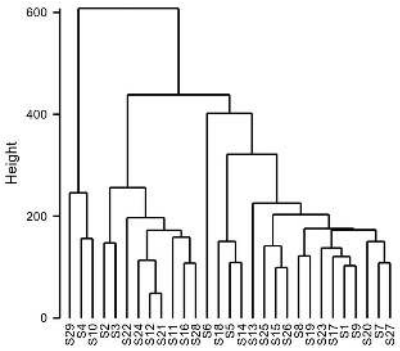
数据处理

- 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

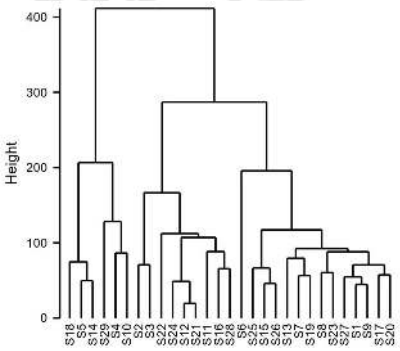
聚类

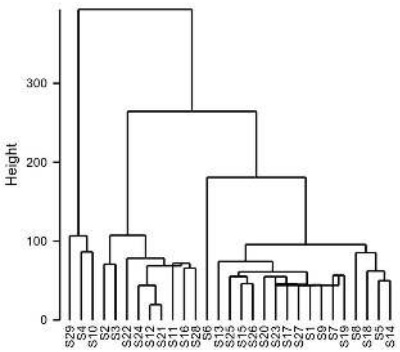
- 类型：可以选择聚类的类型，默认选择常用于一般数据特征的层次聚类-欧氏距离，计算距离的方法默认欧氏距离，其他常用的方法有：曼哈顿距离、堪培拉距离等，如下：



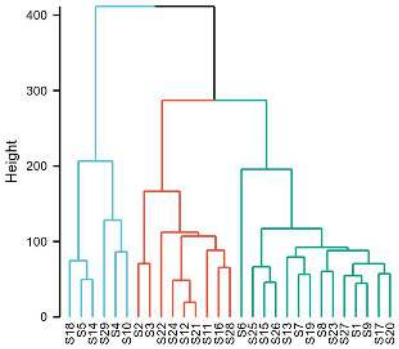


- 方法：可以选择聚类的方法，默认选择类平均法，也可选择常用中间距离法、最长距离法、最短距离法等





- 切分：可以选择是否对数据进行聚类分群的操作，默认为不进行切分，还可以选择切分以及切分成几类，并提供聚类分群的相关补充结果，如下：



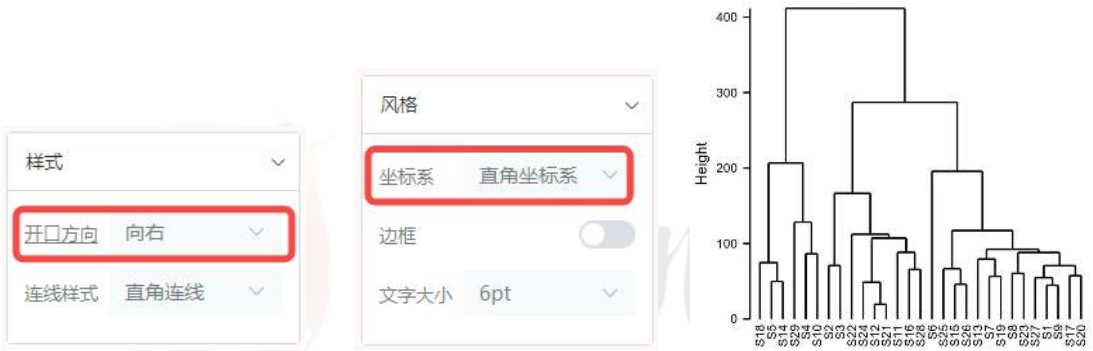
聚类分群

提供切分后的分群的情况

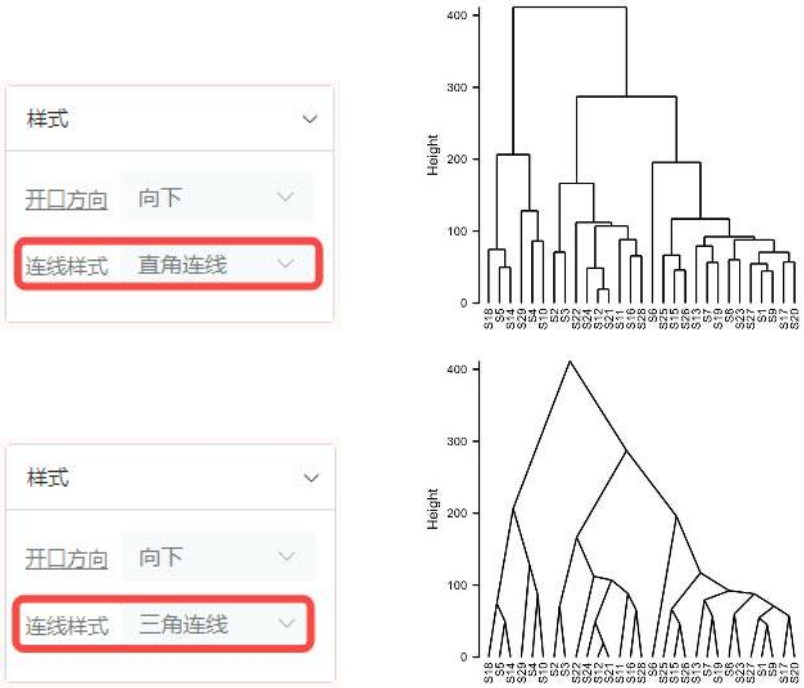
聚类群	包含个数
1	6
2	9
3	14

样式

- 开口方向：可以选择并修改聚类树的开口方向(只能在选择坐标系为直角坐标系时才有用)，如下：

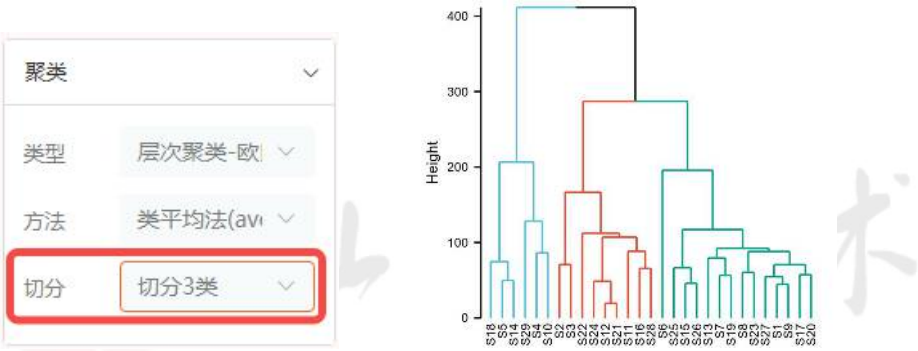


- 连线样式：可以选择并修改聚类树各线条的连线样式，如下：



线

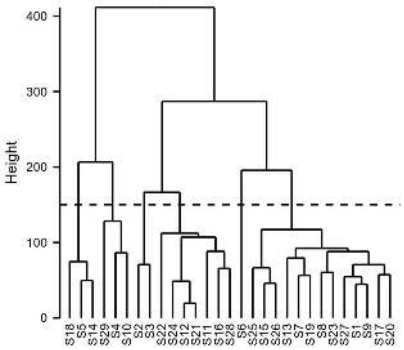
- 颜色：当选择对数据进行聚类分群(切分)操作时，可以修改各个分群的颜色



- 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

分割线

- 位置:可以输入需要进行参考线绘制的值(数值,并且不能超过 y 轴坐标范围),如下:



- 颜色：可以修改参考线的颜色
- 线条类型：可以修改参考线的线条类型
- 线条粗细：可以修改参考线的线条粗细，默认为 0.75pt

标题文本

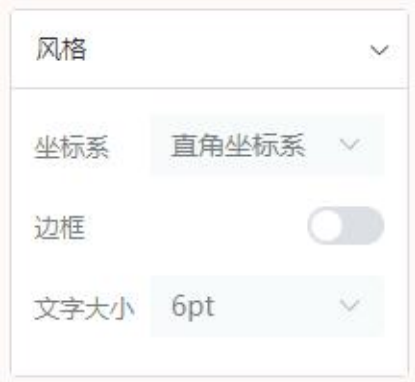
- 大标题：大标题文本
- x 轴标题：x 轴标题文本
- y 轴标题：y 轴标题文本

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]]

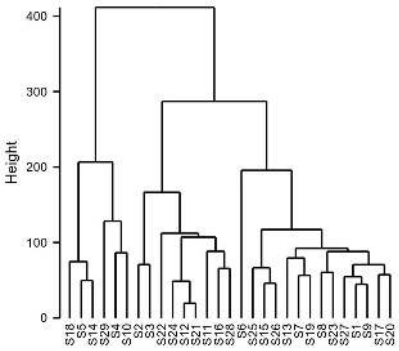
坐标轴

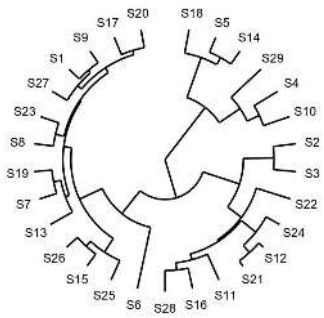
- x 轴标注旋转：可以选择分类表示的横向坐标轴（x 轴）标注旋转的角度

风格



- 坐标系：通过坐标系的类型可以选择并修改聚类树的样式，可以选择绘制直角坐标系类型的聚类树，也可以绘制环形坐标系类型的聚类树，如下：





- 边框：可以选择是否进行添加边框的操作
- 文字大小：控制整体文字大小，默认为 7pt

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

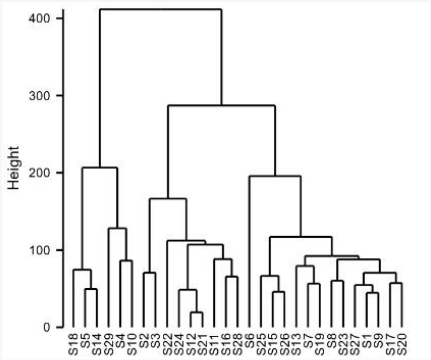
主要结果

聚类树状图

聚类树状图: 是一种展现有群组、层次关系的比例数据的一种分析工具

作用: 以树状图的形式将进行聚类分析的分类（变量/分组）进行可视化，看分类之间的聚类情况

聚类: 此次聚类类型：层次聚类-欧氏距离(euclidean)；方法：类平均法(average)；切割类型：不切分

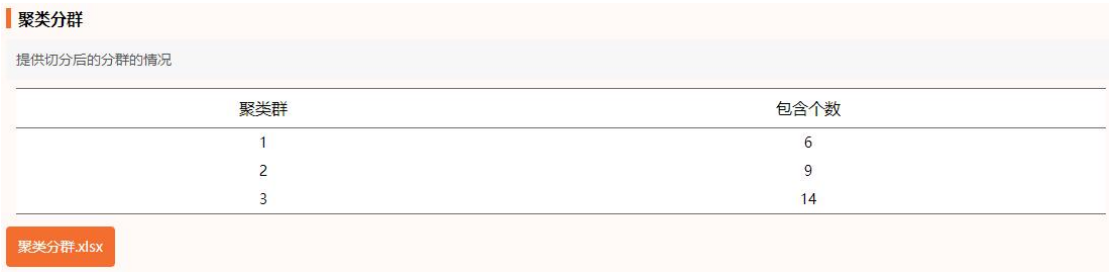


聚类树状图.pdf　聚类树状图.tiff　聚类树状图.pptx

（1）横向坐标表示分类

（2）纵向坐标表示分类所对应的相对距离

补充结果



聚类分群

提供切分后的分群的情况

聚类群	包含个数
1	6
2	9
3	14

聚类分群.xlsx

当选择对数据进行聚类分群(切分)操作时，提供切分后的分群情况：

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

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，这个模块目前不提供 kmeans 等聚类方法（因为速度慢并且内存消耗较大）。

2. 聚类过程提供分组聚类?

答：此模块不提供分组聚类内容。