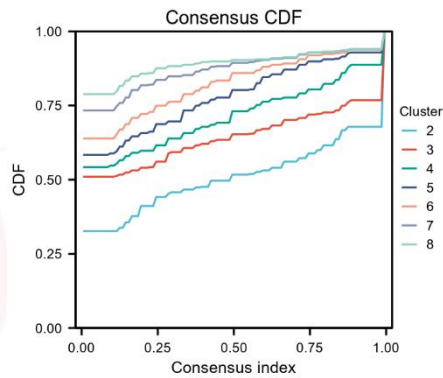


功能聚类 — 【一致性聚类】 CDF



网址: <https://www.xiantao love>



更新时间: 2023.04.24

目录

基本概念	3
应用场景	4
主要结果	5
云端数据	6
参数说明	7
分析参数	7
线	7
标题	8
图注(Legend)	9
坐标轴	9
风格	11
图片	11
结果说明	12
主要结果	12
方法学	13
如何引用	14
常见问题	15

基本概念

- 一致性聚类分析：使用不同的距离方法计算样本间的距离，对距离矩阵以重复抽样的方式来进行聚类分析，将多次聚类结果换算成概率，获得一致性聚类矩阵，从而验证聚类的合理性。
- **CDF 累积分布图**：利用累积分布图展示每个K值的一致性得分分布情况。

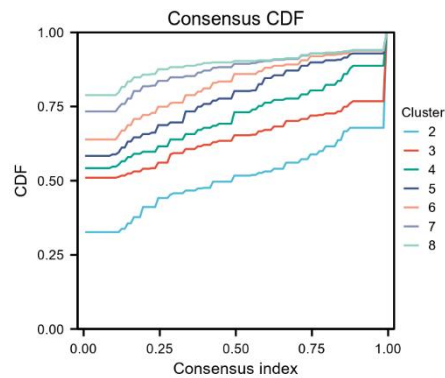


应用场景

通过一致性聚类分析可以获得不同 K 值下的一致性得分分布，使用 CDF 累积分布图可以直观评估和选择最合适的聚类数量 (K)。



主要结果



一致性聚类 CDF:

- 图中每一条线代表对应 K（分类簇数目）值下一致性得分的累积分布，横坐标是一致性得分频数，纵坐标是累积频率。
- 当 CDF 曲线下面积相对变化不再大幅度增长时，此时可以考虑对应 K 值进行分类。

云端数据

云端数据

	记录名称	来源模块	时间	补充说明
<input checked="" type="checkbox"/>		一致性聚类分析 @1.0	2023-04-24 10:01:32	数据记录可以在历史记录中找到

这里的云端数据与历史记录汇总 一致性聚类 分析模块的数据记录是保持一致的，可以在历史记录中找到相应的数据记录。

根据需要可视化的项目 选择好对应的云端数据记录。默认使用最近生成的分析记录。



参数说明

(说明：标注了颜色的为常用参数。)

分析参数



分析参数

K值 8

- K 值：一致性聚类分析的分类簇数目，默认为 8，选项是 2-8 之间。一般使用默认即可，展示所有 K 之下的分布情况。

线



线

颜色

线条类型 实线

线条粗细 0.75pt

不透明度 1

- 颜色：分布曲线对应的颜色，有多少个分类簇会提取多少个颜色，最多支持修改 7 个颜色。受配色方案全局性修改。

- 线条类型：可以选择实线或者虚线。
- 线条粗细：线的粗细。
- 不透明度：线条的不透明度，默认为 1，代表完全不透明，此参数请输入 0-1 之间的值。

标题

标题

大标题

大标题内容

x轴标题

x轴标题内容

y轴标题

y轴标题内容

- 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)



图注配置面板，包含以下选项：

- 是否展示：开关按钮，当前处于开启状态。
- 图注标题：图注标题内容
- 图注位置：默认

- 是否展示：是否展示图注
- 图注标题：可以添加图注标题
- 图注位置：可以选 默认、右、上、右下。

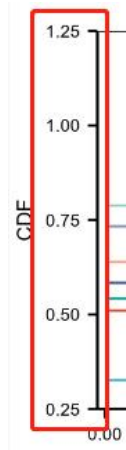
坐标轴



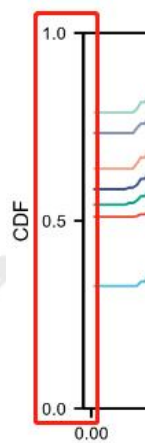
坐标轴配置面板，包含以下选项：

- x轴范围：逗号隔开
- y轴范围：逗号隔开

- x 轴范围：可以控制 x 轴范围和刻度，提供 2 个值来控制范围。形如 -1, 1 (最小值和最大值不能超过 3, 调整过大可能会无作用)。
- y 轴范围：可以控制 y 轴范围和刻度，提供 2 个值来控制范围。形如 -1, 1 (最小值和最大值不能超过 3, 调整过大可能会无作用)。
- 如果只是想要修改范围，可以只输入两个范围值，比如 0.25,1.25



- 如果同时想要修改范围+刻度，可以输入比如：0,0,0.5,1,1。注意，此时最大和最小值会被当做范围值，不会作为刻度，如果需要刻度，需要类似于 1 那样同时写两次。



风格



- 边框：是否添加主图外框
- 网格：是否添加网格
- 文字大小：控制整体文字大小，默认为 6pt

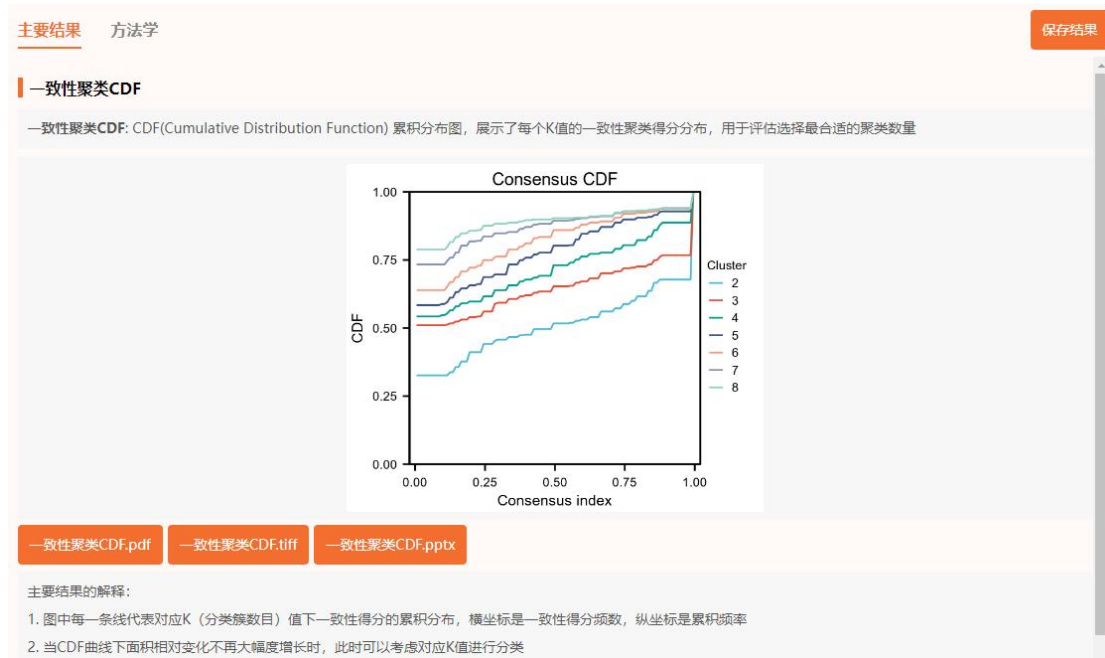
图片



- 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



主要结果格式为图片格式, 提供 PDF、TIFF、PPTX 格式下载。

方法学

所有分析和可视化均在 R 4.2.1 中进行

涉及的 R 包: ggplot2 包 (用于可视化)

处理过程:

- (1) 计算样本的距离矩阵, 通过重复多次抽样和聚类分析计算各样本在不同分类簇数目 K 下的一致性聚类分数
- (2) 计算不同 K 值下一致性聚类得分的概率密度积分
- (3) 使用 ggplot2 包绘制分布曲线进行可视化



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. 如何选择合适的 K 值?

答:

当一致性聚类 CDF 下降坡度更小, 且曲线下面积相对变化不再大幅度增长时, 对应的 K 值较为合适。

