功能聚类 - 【一致性聚类】分析

一致性聚类
分析入口

网址：https://www.xiantao.love

更新时间：2023.04.24

目录

仙桃学术

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。
- 计算距离方法：
  - pearson，即 1-皮尔森相关系数，当数据满足正态分布，且数据之间的差距不大时可选该距离算法。
  - spearman，即 1-斯皮尔曼，该相关性算法时非参数检验的秩统计方法，使用范围较广。
  - euclidean，欧式距离，最常用的距离算法，在 N 维空间中两个点之间的真实距离，或者向量的自然长度（即该点到原点的距离）。
  - maximum，计算最大距离，计算数据集中单位之间或两个不同数据集中观测值之间的最大距离。
  - manhattan，曼哈顿距离，用于几何空间度量，表示两个点在标准坐标系上的绝对轴距距离总和。
  - Canberra，坎贝拉距离，用来衡量两个向量空间的居间，被用作比较排名列表和计算机安全中的入侵检测的测量。
  - binary，二进制距离，常用于定义了二进制向量之间的各种相似度或距离度量。
  - minkowski，闵可夫斯基距离，是欧氏空间中的广义距离函数，其参数 p 值的不同代表着对空间不同的度量。
- 聚类算法：
  - hclust，层次聚类，该算法是基于簇之间的相似度在不同层次上分析数据，从而形成树形的聚类结构。

- hclust 层次聚类方法：
  - average，两个类/群所有节点的平均距离作为两类的距离。
  - complete，两个类/群中最长的节点作为两类的距离。
  - single，两个类/群中最短的节点作为两类的距离。
  - centroid，两个类/群中的中心点的距离作为两类的距离。
  - ward.D，保证类合并时离均差平方和增量最下。
  - ward.D2，在 ward.D 基础上进行了平方。

仙桃学术

应用场景

一致性聚类是一种无监督聚类方法，可以将数据集中的样本区分成不同的亚型，随后对亚型进行比较分析。

主要结果

这里为任务式模块，提交任务后需要到历史记录中刷新并等待任务完成，（分析时间大概在 几分钟 左右，如果任务执行时间过长，刷新后任然在执行阶段，建议删除后重新提交。）

仙桃学术

数据格式

	A	B	C	D	E	F	G
1	Genes	s1	s2	s3	s4	s5	s6
2	gene1	-1.50991	-3.2774	-2.46946	1.483049	2.459767	0.234124
3	gene2	-3.73909	1.40959	2.689937	1.599547	-3.60629	-3.70625
4	gene3	-1.39978	-2.94408	-0.30223	2.743997	1.011225	3.558182
5	gene4	-1.50669	-3.62763	-4.33151	-1.91955	1.139198	0.095649
6	gene5	0.675272	-1.67351	1.240537	2.171307	-0.95152	-1.36593
7	gene6	-1.50667	-0.35609	1.839982	-1.87729	0.156029	-1.75589
8	gene7	2.823646	-1	-0.71288	2.890608	-0.02103	3.73674
9	gene8	-0.05454	-0.6439	-1.68796	-1.74898	1.993951	0.886564
10	gene9	-3.99559	0.849275	1.245652	0.095505	-3.93367	-4.39669
11	gene10	-2.11254	-0.44921	-2.81493	-3.62388	-0.44068	-1.74903
12	gene11	1.831823	-1.9357	-2.1209	1.363133	-2.33771	-2.16575
13	gene12	1.387253	-1.22456	-0.9716	2.012572	0.496663	1.712959

数据要求：

- 除第一列分子名称外，数据至少有 10 列（样本）以上，每列至少 3 个观测（如果不满足这个条件，则不会进行分析），每一列均需要是数值类型。
- 每列数据为一个样本，计算距离及聚类分析即以样本进行分析。每一行为分子（特征/基因），若验证数据时返回报错，需要在上传数据内进行相应的调整，然后再上传数据。
- 注意：一致性聚类分析是通过重复多次抽样聚类的方式，来计算聚类结果一致的概率，从而验证聚类的合理性，因此，样本越多效果越好。
- 最多支持 500 个样本（500 列），6000 行。

这里为任务式模块，提交任务后需要到历史记录中刷新并等待任务完成，（分析时间大概在 几分钟 左右，如果任务执行时间过长，刷新后任然在执行阶段，建议删除后重新提交。）

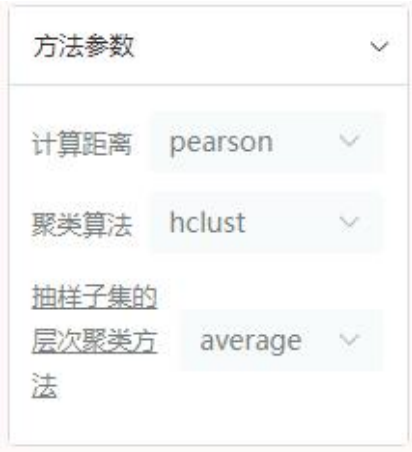
参数说明

（说明：标注了颜色的为常用参数。）

数据处理

- 归一化：默认为 无，即不做数据处理，当选择 "对行(变量)归一化" 时，将对数据矩阵中的行作 Scale 处理，对特征进行归一化可以有效减少特征之间数量级过大的问题。

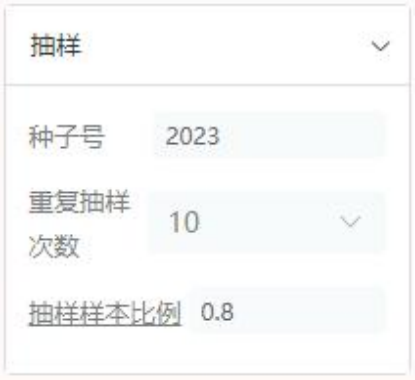
方法参数



- 计算距离：计算距离方法默认为 pearson，计算距离方法的选择可以参考“基本概念”中计算距离方法的说明。

- 聚类算法：默认为 hclust，选择不同的聚类方对样本间的抽样距离矩阵聚类，聚类算法的选择可以参考“基本概念”中聚类算法的说明。
- 抽样子集的层次聚类方法：当聚类算法选择层次聚类 hclust 时，此参数为层次聚类的方法，默认使用 average。可以参考“基本概念”中层次聚类方法的说明。

抽样



- 种子号：设置种子数可以保证重抽样结果可重复，默认为 2023，此参数请输入非零整数。
- 重复抽样次数：对样本间距离进行聚类分析时，重复抽样的次数，默认为 10，可以选 10、50、100、500、1000。
- 抽样样本比例：对样本间距离进行聚类分析时，通过重复多次抽取一定比例的样本距离数据进行聚类，默认为 0.8，此参数请输入 0-1 之间的值，1 代表比例为 100%，每次抽取所有样本。（总样本数目 * 抽样样本比例）不能小于 K 值 8。

结果说明

这里为任务式模块，提交任务后需要到历史记录中刷新并等待任务完成，（分析时间大概在 20s 左右，如果任务执行时间过长，刷新后任然在执行阶段，建议删除后重新提交。）。

注意：一致性聚类 的可视化需要到对应的 可视化模块中进行。如果删除了数据记录或者无该分析记录，将无法进行可视化。

仙桃学术

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：ConsensusClusterPlus 包

处理过程：

⑴ 将清洗后的数据计算样本的距离矩阵，通过重复多次抽样和聚类分析计算不同 K 值下各样本的一致性分数矩阵。

仙桃学术

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。

仙桃学术

常见问题

1. 最少样本数?

答:

一致性聚类分析通过重复抽样的方式进行聚类结果验证，本模块默认计算 K 值为 2-8 的所有一致性得分。样本数目与抽样比例、K 值均有关联，根据默认参数遵循（抽样）样本数大于或等于 K，因此原则上至少需要 10 个样本及以上。

一致性聚类分析是通过重复多次抽样聚类的方式，来计算聚类结果一致的概率，从而验证聚类的合理性，因此，样本越多效果越好。

2. 如何进行可视化的操作?

答:

提交分析任务完成后，历史记录中会有一条对应的结果记录。同时在 一致性聚类 可视化模块中可以选择对应的数据记录，可以对数据进行可视化。