# 聚类分析

2023年11月29日 21:54

## 聚类分析的一般步骤

1. 选取合适的变量
2. 缩放数据(标准化)
   scale()
3. 寻找异常点
   outliers包中函数筛选一场单变量离群点
4. 计算距离(默认使用欧几里得距离)
   欧几里得距离适用连续型数据的距离度量
   若存在其他数据类型，可以使用cluster包中的daisy()函数
   的欧几里得距离定义为：

$$d_{ij} = \sqrt{\sum_{p=1}^{p}(x_{ip} - x_{jp})^2}$$

   这里 $i$ 和 $j$ 代表第 $i$ 和第 $j$ 个观测值，
   $p$ 是变量的个数。

   dist()计算距离并形成形成下三角矩阵
   as.matix()将下三角矩阵转换为标准矩阵
5. 选择聚类算法
   a. 层次聚类对于小样本来说很实用(150以下)
   b. 划分聚类能处理更大的数据量
6. 确定类的数目
7. 获得最终的聚类解决方案
8. 结果可视化
   聚类结果通常表示为树状图
9. 解读类
10. 验证结果
    fpc、clv、clValid包包含了评估聚类解的稳定性的函数

## 层次聚类

每一个观测值自成一类，两两合并，最终全部合成一类

算法步骤：

1. 定义每个观测值为一类
2. 计算每类和其它各类的距离
3. 把距离最短的两类合并为一类，这样类的个数就减少一个
4. 重复步骤(2)和步骤(3)，直到包含所有观测值的类合并为单个的类

各种层次聚类算法的区别是对类的定义不同

| 聚类方法 | 两类之间的距离定义 |
| --- | --- |
| 单联动 | 一个类中的点和另一个类中的点的最小距离 |
| 全联动 | 一个类中的点和另一个类中的点的最大距离 |
| 平均联动 | 一个类中的点和另一个类中的点的平均距离（也称作UPGMA，即非加权对组平均） |
| 质心 | 两类中质心（变量均值向量）之间的距离。对单个的观测值来说，质心就是变量的值 |
| Ward法 | 两个类之间所有变量的方差分析的平方和 |

单联动法(single)倾向于发现细长的类，展示链式现象
全联动(complete)倾向于发现大致相等的直径紧凑类
平均联动(average)提供了两种方法的折中

hclust(d,method=)
d:dist()函数产生的距离矩阵
例子：

```
data(nutrient,package='flexclust')
row.names(nutrient) <- tolower(row.names(nutrient))#将行名转换为小写
nutrient.scale <- scale(nutrient)#数据标准化
d <- dist(nutrient.scale)#计算每个观测间的距离
fit.average <- hclust(d,method='average')#层次聚类
plot(fit.average,hang=1,cex=0.8,main='Average Linkage Clustering')
```

## 划分聚类

首先指定类的个数为K，将观测值随机分成K类，再重新聚合