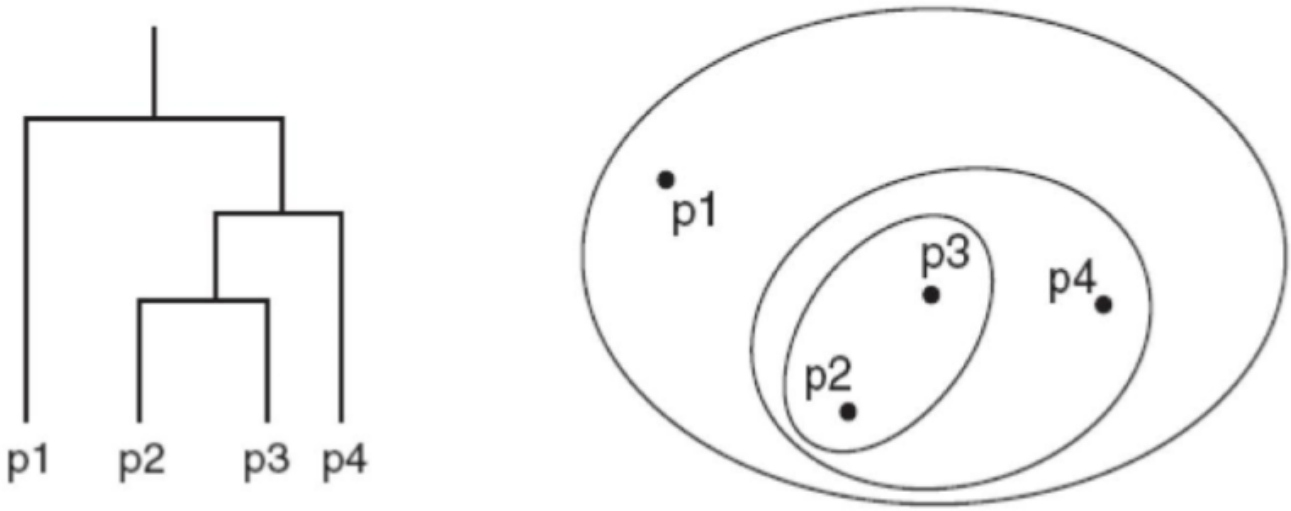
层次聚类

【简介】

层次聚类是一层一层地进行聚类，可以由上往下将一个大类不断分割，也可以由下往上不断对小的类别进行聚合。

层次聚类通常用一个类似树的图案表示，称为树状图。树状图展示了簇和子簇的关系以及簇聚合或者分裂的顺序。



层次聚类有两种基本的方法：

1、凝聚式：将多个小类不断聚合最终聚合成一个大类。

2、分裂式：将一个大类不断分割最终分裂成很多个小类。

下面是对两种聚类方式的阐述：

【凝聚式层次聚类】

凝聚式层次聚类一开始先把每个样本视为个体簇，然后接下来的每一步，合并距离最近的一对簇，最终形成一个大簇。

凝聚式层次聚类的步骤大致总结如下：

- 计算距离矩阵
- 迭代
    - 合并距离最近的两个簇
    - 更新距离矩阵，从而反映新簇和原簇的距离
- 直到只剩下一个簇

从上面的步骤我们可以看出，我们需要计算两个簇之间的距离，对簇间距离的不同定义会导致凝聚式层次聚类的不同版本，这些版本包括

1. Single link（单连接）or MIN

两个簇的距离定义为位于两个簇中任意两个点的最小距离。这种方法可以较好的处理非球状的样本集。

2. Complete link（全连接） or MAX

   两个簇的距离定义为位于两个簇中的任意两个点的最大距离。

3. Group average（组平均）

   两个簇的距离定义为两个簇形成的所有不同点的距离的平均值，另外属于同一个簇的不同点对的距离也计算入内。这种方法是全连接与单连接的折中版本。

$$dis-ga(C_i,C_j)=\frac{1}{(N_i+N_j)(N_i+N_j-1)}\sum_{d_m\in C_i\cup C_j}\sum_{d_n\in C_i\cup C_j,d_n\neq d_m}dis\left(\vec{d}_m,\vec{d}_n\right)$$

clusters i and j

the number of documents in cluster i

4. Centroid Similarity

【分裂式层次聚类】

一开始把所有样本是做一个大簇，然后接下来的每一步都分割一个簇，持续这个过程直到每个簇只包含一个样本或者对象。

分裂式层次聚类简要过程为：

- 初始所有样本都在一个簇中
- 采用非层次聚类算法（如K-means）对每个簇进行分裂（如划分为k个簇）
- 重复步骤2，直到每个簇中只有一个样本或者最近的两个簇之间的距离大于某个人工给定的阈值。

在这种情况下我们需要解决两个问题，一是在每一步我们需要分割哪个簇，二是如何分割簇。

关于簇的选取，通产采用一些衡量松散程度的度量值来比较，例如簇中距离最远的两个数据点之间的距离，或者簇中所有结点相互距离的平均值等，直接选取最松散的一个簇来进行分割。