第十章 层次聚类

一、基本思想

1. 层次聚类的目的
   层次聚类会生成一系列具有层次结构的嵌套簇，为了展示这一层次关系，通常采用树状图的图形，树状图是一种树形结构，根据图形的方式记录了聚类过程中簇的合并或者拆分。进一步，层次聚类可以被划分为两种方式，一种是自底向上的合并式，一种是自顶向下的拆分式
2. 树状图介绍
   树状图是一种便于划分聚类划分分类层次的图形，它是一个简单的树，在这个数中：
   每个节点代表一个组
   每个叶节点代表一个单数据点
   根节点是包含整个数据集的集合
   每个内部节点都有两个子节点，代表了这个节点是从这两个子节点聚合得到的
   其中连接方式决定了划分不同组别数据点之间差异性的方式
3. **层次聚类的特点**
   - 对簇的数量不做任何假定
     通过在适当的层次切割树状图，我们可以得到任何想要的簇的大小
   - 层次聚类可能和有意义的分组相契合

二、合并聚类

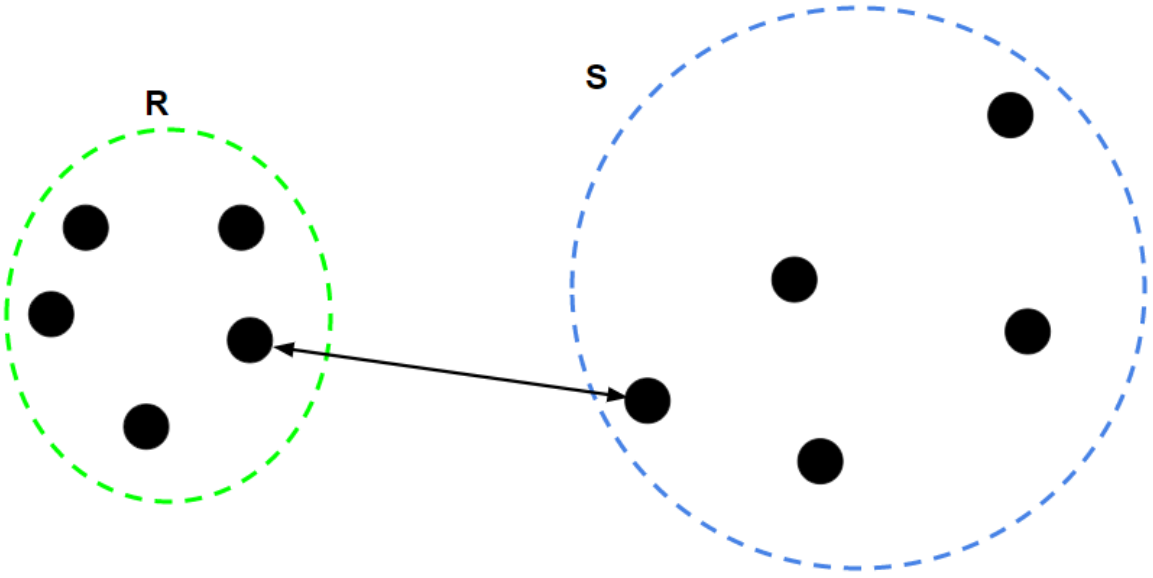
1. 两种层次聚类介绍
   - **合并式聚类**
     这种采用自底向上的策略，首先将每个对象划分为一个簇，然后将这些簇划分为更大的簇，直到所有的对象在同一单个簇之中，或者说终止条件被满足。大多数层次聚类都属于合并式聚类
   - **分裂式聚类**
     这种自顶向下的策略和合并式聚类方法截然相反，它是从包含所有对象的一个簇开始，将簇划分为越来越小的簇，直到每个簇中只有一个对象，或满足特定的终止条件，例如达到预设的聚类数量，或者每个簇的直径处于特定的阈值
2. **合并式聚类算法流程**
   - 计算数据点之间的距离矩阵
   - 让每个数据点单独成簇
   - 合并两个最近的簇
   - 更新距离矩阵
   - 直到只剩下一个簇，或者达到终止条件
     其中关键操作是计算两个簇之间的距离，使用不同方法计算距离会导致不同的聚类结

果

三、关联——两个簇之间的距离

1. **单链链接**
   对于两个簇R和S，这个单链链接，就是计算两个簇之间距离最近的两个样本点之间的距离

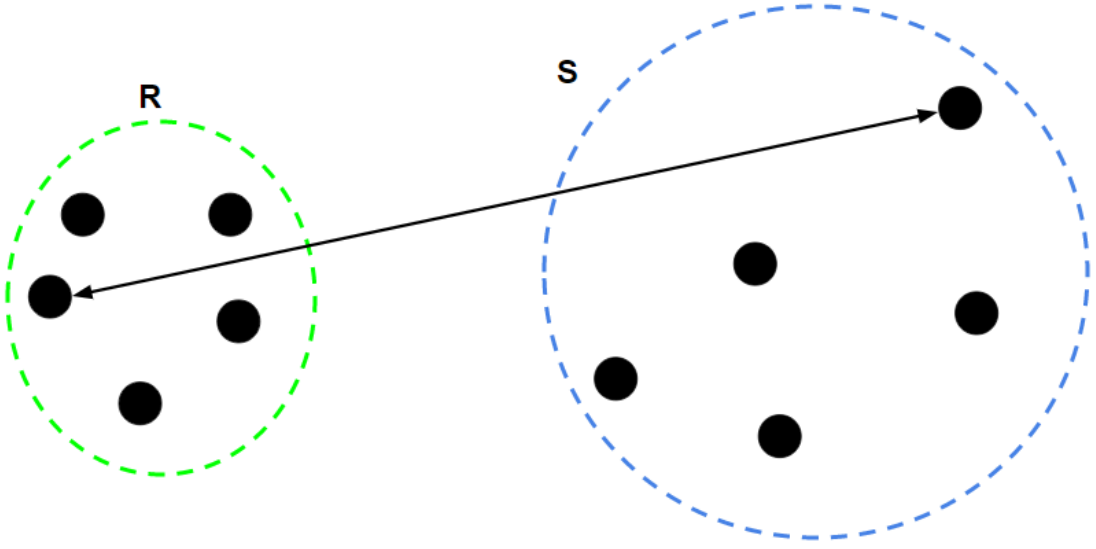
$$L(R,S) = min(D(i,j)), where\ i \in R, j \in S$$



   单链链接存在“松散连接”的问题。为了将两个群体合并，只需要有一对点彼此靠近即可，而无需考虑其他所有点。因此，聚类结果可能会过于分散，不够紧凑。

2. **完全链接**
   对于两个簇R和S，这个单链链接，就是计算两个簇之间距离最远的两个样本点之间的距离

$$L(R,S) = max(D(i,j)), where\ i \in R, j \in S$$

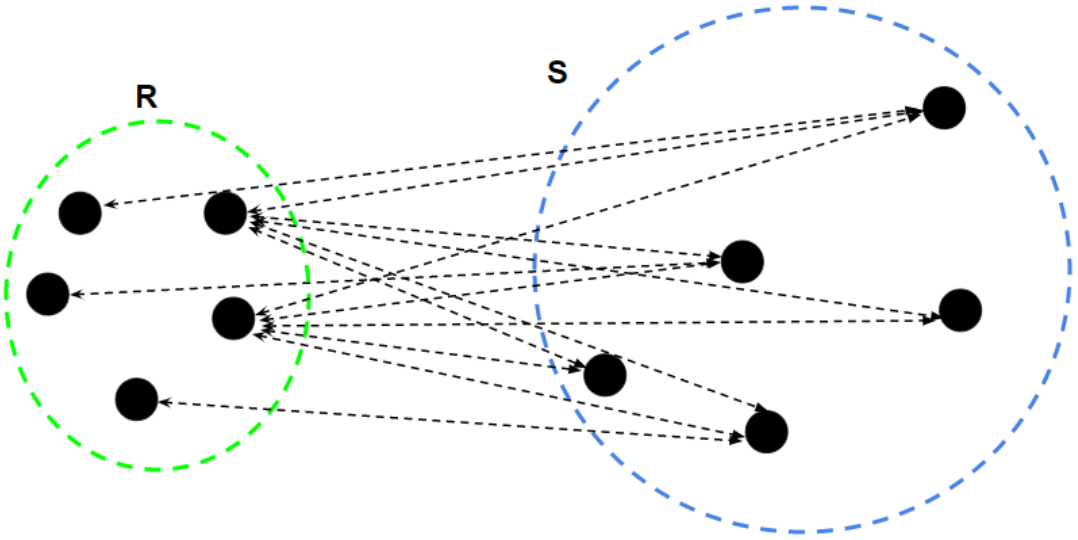


   完全链接方式能避免松散连接问题，但会存在拥挤的问题。 因为其评分是基于各组之间最

差情况下的差异程度得出的，所以一个点可能与其他组中的点距离更近，而非与自身组中的点距离更近。这些组是紧密相连的，但彼此之间的距离又不够远。

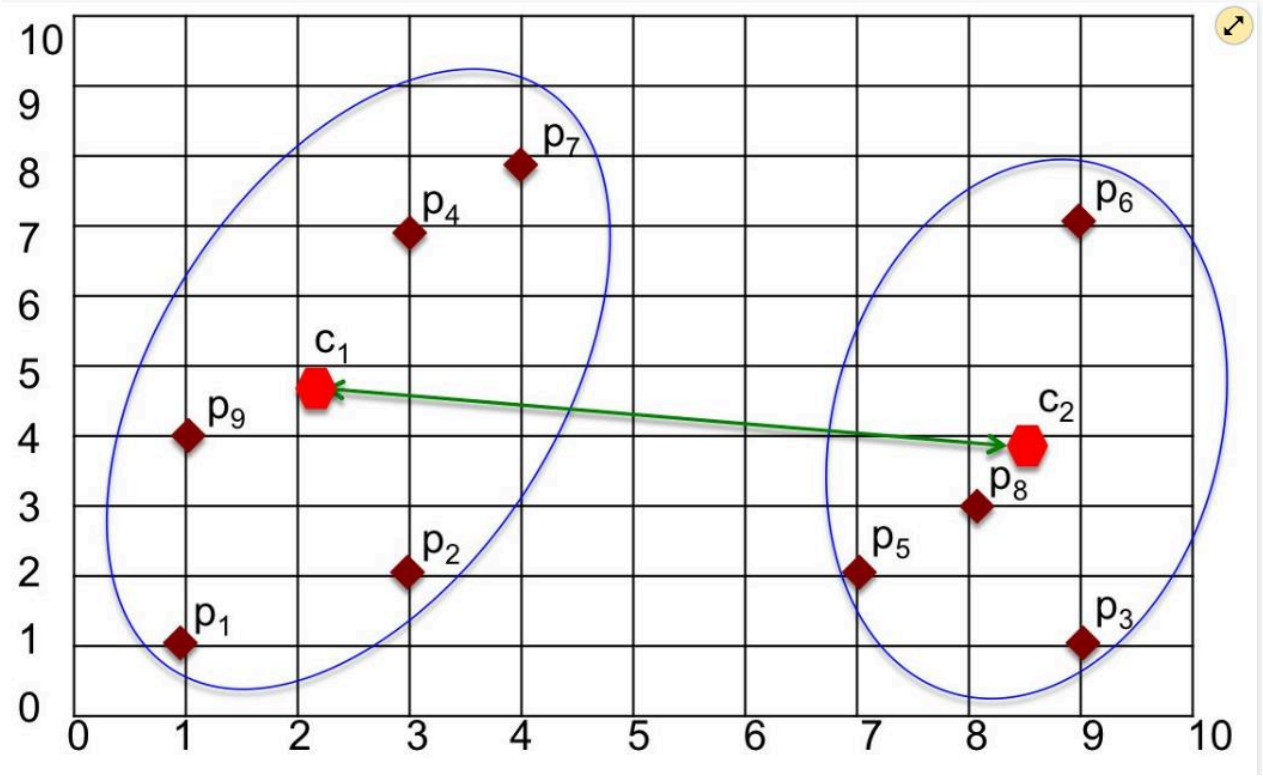
3. **组平均链接**
   对于两个簇R和S，这个单链链接，就是计算两个簇之间每一对两个样本点之间的距离的算数平均值

$$L(R,S)=\frac{1}{n_R\times n_S}\sum_{i=1}^{n_R}\sum_{j=1}^{n_S}D(i,j),where\ i\in R,j\in S$$



4. **组平均距离**
   计算两个簇C_i和C_j的样本均值点或者簇中心点的距离

$$D(C_i,C_j)=d(\mu_i.\mu_j)$$



5. Ward法——新生成的簇的方差增加量是最小的
   通过衡量簇内方差增加量来决定两个簇之间是否需要合并

计算合并后簇的误差与合并前两个簇之间的误差的差值，最后选择让方差增加量最小的

$$\Delta SSE = SSE(C) - (SSE(A) + SSE(B))$$

四、层次聚类与K-means的比较

Feature（特性）	Hierarchical Clustering（层次聚类）	K-Means Clustering（K-Means 聚类）
Type of clustering聚类方式	Agglomerative（自底向上）或 Divisive（自顶向下）	Partitional（基于中心的划分式聚类）
Cluster shape簇的形状	能处理非凸形状、大小不均的簇	假设簇为球状，并且大小相近
Distance metric距离度量	可使用多种距离：欧氏、曼哈顿、余弦等	只能使用欧氏距离
Scalability可扩展性	对大数据集或簇数较多时计算代价高	能高效处理大型数据集和大量簇
Interpretability可解释性	提供层次结构和树状图（dendrogram），解释性更强	只有簇中心和分配结果，不提供层次结构
Outliers对异常点的敏感性	Sensitive（敏感）	Sensitive（敏感）

五、合并层次聚类和K-means聚类

1. 思路来源
   k-means 聚类：一种分区式方法，用于将数据集划分为 *k* 个聚类。
   层次聚类（hierarchical clustering）：一种替代 k-means 的聚类方法，通过使用观测样本之间的成对距离矩阵作为聚类判据来识别数据中的聚类结构。
   然而，这两种标准聚类方法各自都存在局限性
   - **k-means 聚类**要求用户事先指定簇的数量，并且其初始质心是随机选择的。
   - **凝聚式层次聚类**善于识别小型簇，但对大型簇的识别效果不佳。
     对此衍生出了一种新的方法混合层次化K-means聚类(hkmeans)
2. 混合层次化K-means聚类过程
   - 先进行层次聚类，并将聚类树剪切成 *k* 个簇。
   - 计算每个簇的中心（即均值）。
   - 使用步骤 2 中得到的这些簇中心作为初始中心，运行 k-means 聚类。
     通过这种方式，k-means 算法会进一步改进步骤 2 所生成的初始聚类结果。