71Y18103 王然

**14.2 层次聚类**

层次聚类中分为**聚合(agglomerative)**或**分裂(divisive)**聚合聚类需要预先确定三个要素: **①距离或相似度 ②合并规则 ③停止条件**

**聚合聚类算法：**

输入：n个样本组成的样本集合及样本间的距离

输出：对样本集合的一个层次化聚类

1. 计算n个样本两两之间的欧式距离 {dij}，集作==记作矩阵D=[dij]n×n
2. 构造n个类，每个类只包含一个样本
3. 合并类间距离最小的两个类，其中最短距离为类间距离，构建一个新类
4. 计算新类与当前各类的距离.若类的个数为1，终止计算，否则回到步骤3.

算法复杂度：第(1)步计算距离的时间约为O(m)，两两距离共有O(n²)组，这一步总共的时间复杂度是O(n²m).第(1)步的时间约为O(n).第(3)步涉及一次扫描，时间约为O(n).第(4)重算的距离共有O(n)组.(3)(4)步共要经历(n/2+n/4+...+1)=O(n),总共的时间复杂度是O(n²m)

**算法的复杂度是O(n³m)?**

**14.3 k均值聚类**

属于样本集合划分的聚类算法，将样本划分为k个子集，k均值聚类是硬聚类

结果的k个类形成**对样本集合X的划分**

划分或聚类可以用函数l=C(i)表示，模型是一个从样本到类的函数

k均值函数的策略：**损失函数最小化**

定义**样本与其所属类的中心之间的距离的总和**为损失函数

函数W(C)也成为能量，表示相同类中样本的相似程度，相似的样本被聚类到同类时，损失函数值最小

k均值的最优解是NP问题，现实中采用迭代的方法求解

算法部分和之前在Web data mining上看到的基本一致