**聚类学习笔记总结**

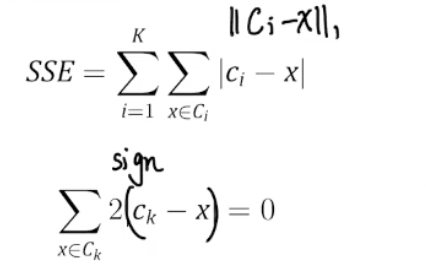
## 1、概念

**聚类分析：简称聚**类(clustering)，是一个把数据对象划分成子集的过程，每个子集是一个簇(cluster)，使得簇中的对象彼此相似，但与其他簇中的对象不相似。

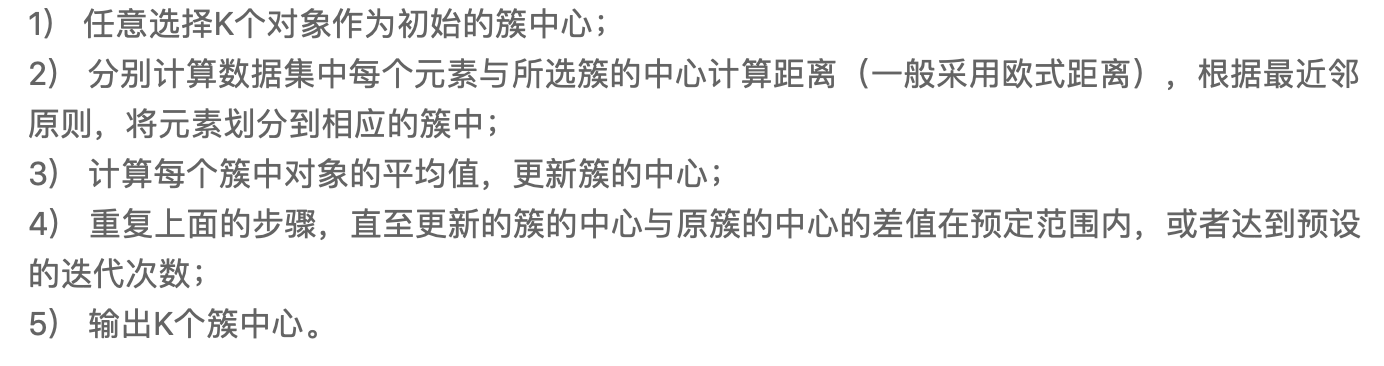
## 2、算法

1. **划分**

K-means:

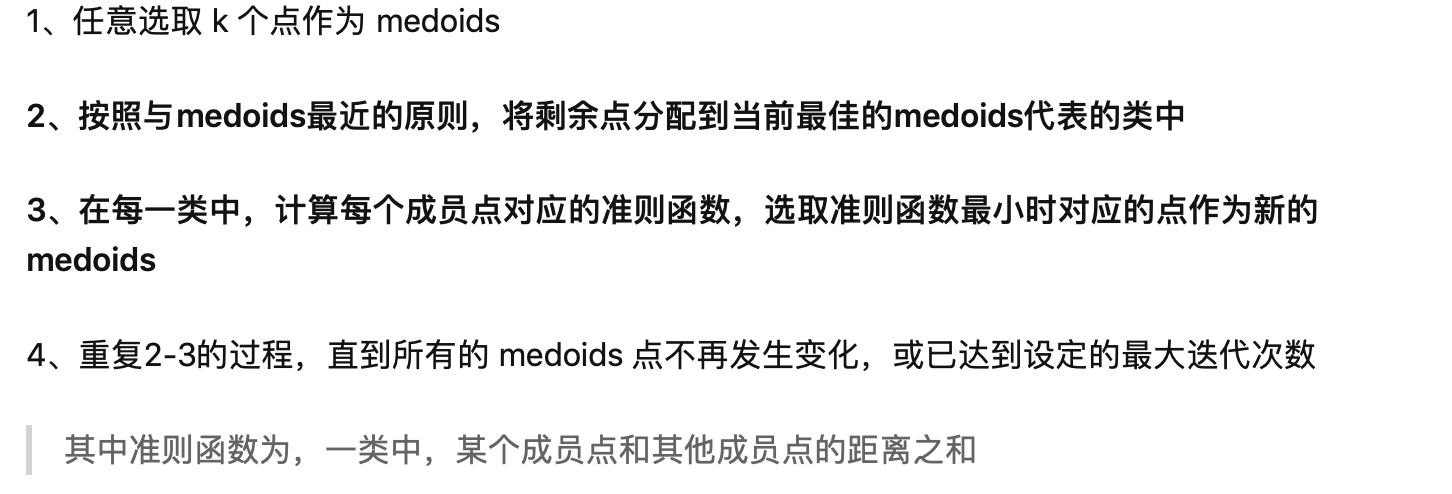


K-means 算法的基本思想是，将制定样本划分为 K 个簇，用一个聚类的中心来代表一个簇，这个中心就是各个簇中所有数据样本的均值（means）（欧式距离-2范数，曼哈顿距离-1范数，对非球形数据、噪音数据不够友好）。[1]



K-Medoids:

K-Medoids（中心点）算法不选用平均值，转而采用 簇中位置最中心的对象，即中心点（medoids） 作为参照点，算法步骤也和 K-means 类似，其实质上是对 K-means算法的改进和优化。[2]



KNN-Kmeans:K-means 属于聚类算法，无监督学习;KNN 属于分类算法，有监督学习。

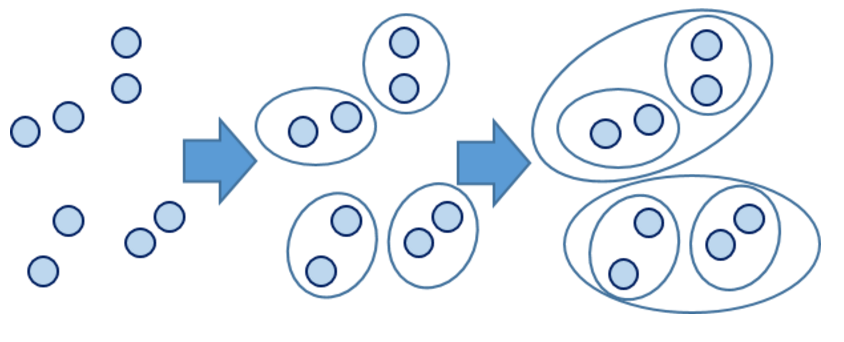
K-medoids :K-means

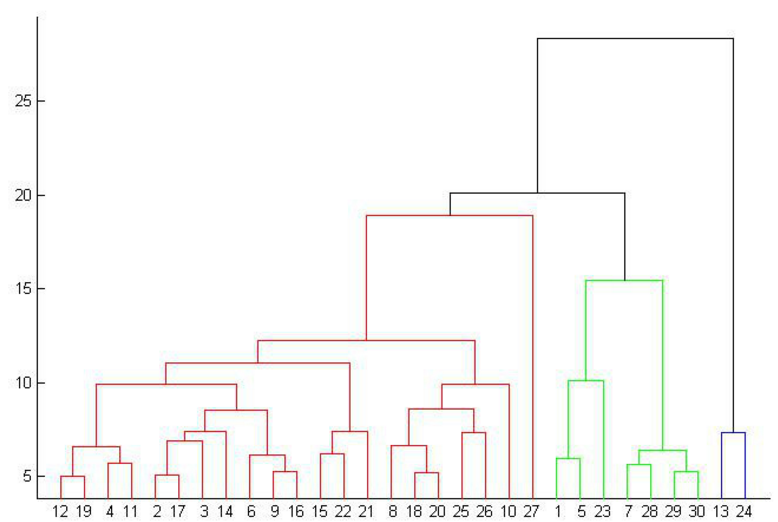
优点：当存在噪音和孤立点时, K-medoids 比 K-means 更健壮。

缺点：K-medoids 对于小数据集工作得很好, 但不能很好地用于大数据集，计算 质心的步骤时间复杂度是O(n^2)，运行速度较慢

1. **层次[4]**

**层次聚类(Hierarchical Clustering)**是聚类算法的一种，通过计算不同类别数据点间的相似度来创建一棵有层次的嵌套聚类树。在聚类树中，不同类别的原始数据点是树的最低层，树的顶层是一个聚类的根节点。创建聚类树有自下而上合并和自上而下分裂两种方法。





两个组合数据点间的距离

　　计算两个组合数据点间距离的方法有三种，分别为Single Linkage，Complete Linkage和Average Linkage。在开始计算之前，我们先来介绍下这三种计算方法以及各自的优缺点。

Single Linkage：方法是将两个组合数据点中距离最近的两个数据点间的距离作为这两个组合数据点的距离。这种方法容易受到极端值的影响。两个很相似的组合数据点可能由于其中的某个极端的数据点距离较近而组合在一起。

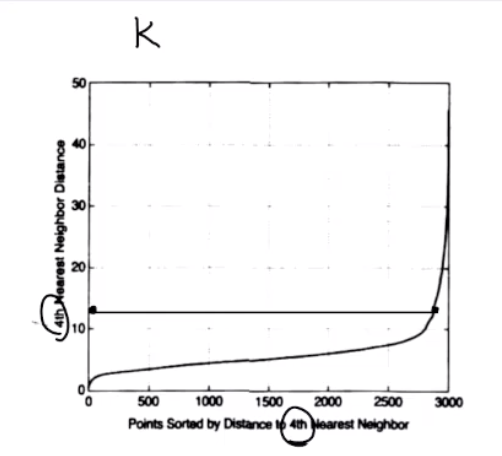
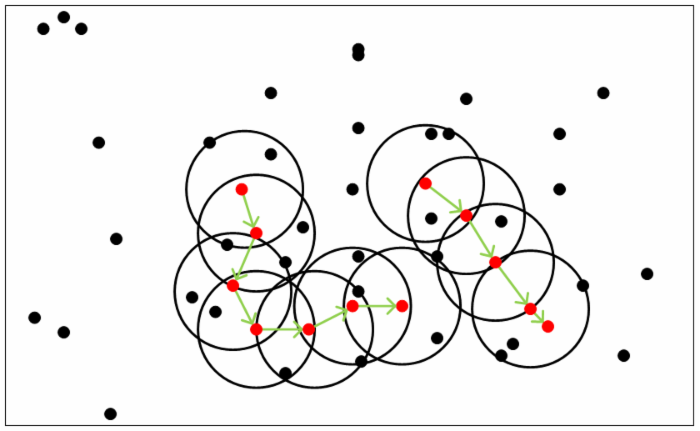
Complete Linkage：Complete Linkage的计算方法与Single Linkage相反，将两个组合数据点中距离最远的两个数据点间的距离作为这两个组合数据点的距离。Complete Linkage的问题也与Single Linkage相反，两个不相似的组合数据点可能由于其中的极端值距离较远而无法组合在一起。

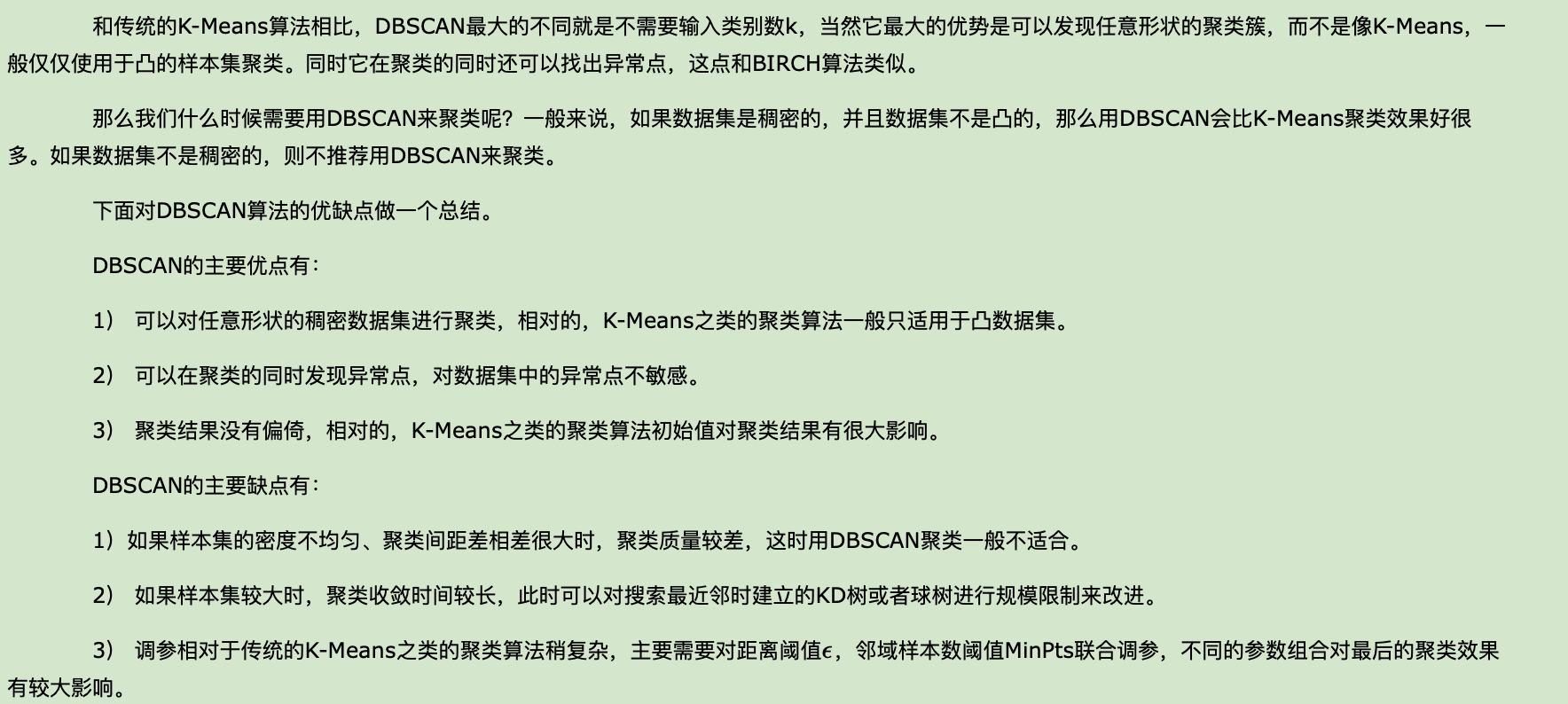
Average Linkage：Average Linkage的计算方法是计算两个组合数据点中的每个数据点与其他所有数据点的距离。将所有距离的均值作为两个组合数据点间的距离。这种方法计算量比较大，但结果比前两种方法更合理。

1. 密度

DBSCAN：

DBSCAN是一种基于密度的聚类算法，这类密度聚类算法一般假定类别可以通过样本分布的紧密程度决定。同一类别的样本，他们之间的紧密相连的，也就是说，在该类别任意样本周围不远处一定有同类别的样本存在。[3]





4）还有高维数灾难。

1. 网格

## 3、评价指标[5]

常见的聚类评价指标: Purity, NMI, RI, Precision(查准率), Recall(查全率), F, ARI, Accuracy(正确率)没咋仔细看；

**好的聚类算法,一般要求类簇具有：**

**高的类内 (intra-cluster) 相似度** (documents within a cluster are similar)

**低的类间 (inter-cluster) 相似度** (documents from different clusters are

dissimilar)

[1].https://zhuanlan.zhihu.com/p/54045059

[2].https://zhuanlan.zhihu.com/p/55163617

[3].https://www.cnblogs.com/pinard/p/6208966.html

[4].https://www.cnblogs.com/zongfa/p/9344769.html

[5].https://zhuanlan.zhihu.com/p/53840697