K-means 算法的改进

K-means 算法由于其简单方便的特性成为最常用的聚类算法之一。但是 K-means 算法也存在着一些缺点，比如需要指定 K 值、结果对初始聚类中心非常敏感。目前，很多研究者对 K-means 算法的缺陷提出了改进的方案，改进的方向主要分为两大类：一是如何选取好的初始聚类中心；二是如何确定合适的 K 值

基本思想

用来聚类的数据一般都是紧凑的，对于分布松散的数据集进行聚类是没有意义的，簇内数据对象的分布一般近似地遵循高斯分布。在数据集中，类簇的中心一般基于这样的事实：（1）如果数据集中某个数据对象是一个类簇的聚类中心，那么它应该具有较大的局部密度。（2 假如数据集中某个数据对象是一个类簇的聚类中心，那么该对象到具有比它更高的局部密度的对象之间的欧式距离一定较大。

相关概念

（1）定义数据对象x_i的局部密度为

$$\rho_i = e^{1/d_i}$$

其中，$d_i = \sum_{k_0} d_{i,j}$，表示x_i和与x_i最近的k_0个对象间的距离之和。

k_0的值由下式给出

$$k_0 = 10\% \times n$$

n为数据集中对象的个数。

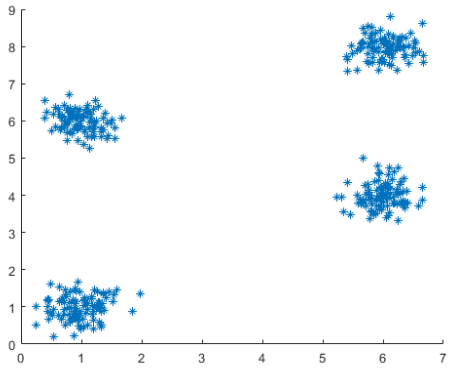
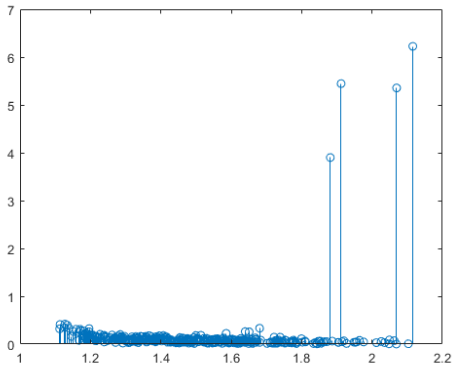
（2）定义数据对象x_i异簇最小距离为:

$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{i,j})$$

若x_i是数据集中局部密度最大的对象，则定义x_i的δ_i为:

$$\delta_i = \max_{x_j \in S} (d_{i,j})$$

回归分析确定聚类中心

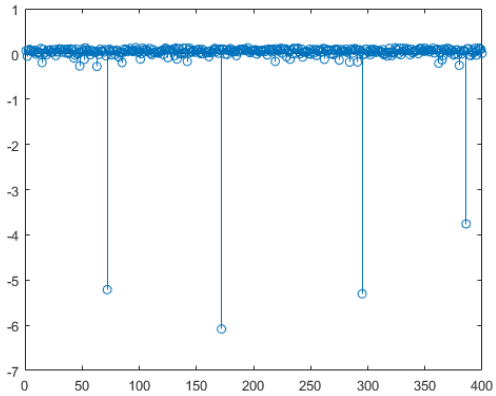
对于如左图所示的数据集，通过右图ρ及δ的数据走势图，可以看出大多数对象的ρ、δ的走势近似反比例函数，而异常点即需要的聚类中心。因此可以通过回归分析及残差分析找出异常点。

采用反比例函数对每个对象的δ值进行拟合，即

$$\delta^* = a_0 + a_1/\rho$$

计算每一个对象的残差

$$\delta_0 = \delta - \delta^*$$



选出残差大于 1.5 的点，即为初始聚类中心，K 值即初始聚类中心的个数。下图即生成的初始聚类中心，在设置迭代终止条件为聚类中心移动小于 0.005 的情况下，只用了一次便完成迭代，这说明初始聚类中心比较理想。