基于密度的聚类算法

Clustering by fast search and find of density peak

基本步骤

1、对于每一个数据点i，计算局部密度ρ_i：d_c是截断距离，d_c的推荐值是使得平均每个点的邻居数为样本总数的 1%-2%，d_c的选择比较鲁邦；ρ_i相当于距离点i的距离小于d_c的点的个数；

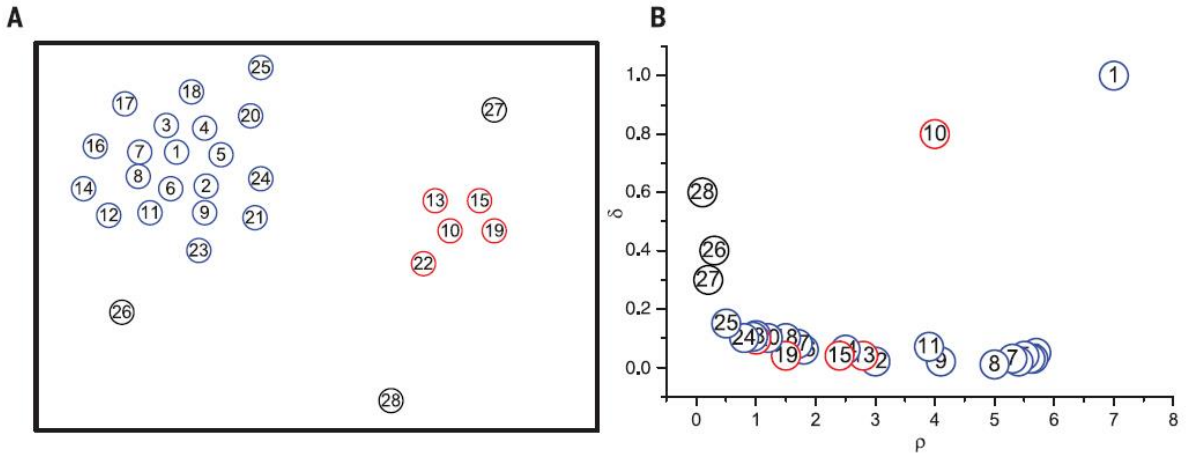
$$\rho_i = \sum_j \chi(d_{ij} - d_c)$$

2、对于每一个数据点i，计算比i点密度高的点到i点的最小距离δ_i；

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

3、根据ρ和δ画出决策图，找出聚类中心；

4、划分剩余数据点到相应的簇。



- 图 A 是所有点在二维空间的分布，所有点的密度值由高到低排列，1 表示密度最高的点；
- 图 B 是以局部密度ρ为横坐标，以到高局部密度点的距离δ为纵坐标的决策图；
- 1 和 10 两个点的ρ和δ都比较大，作为类簇的中心点；
- 26、27、28 三个点的δ也比较大，但是ρ较小，所以是异常点；
- 准则：当前点的类别标签，与高于当前点密度的最近点的标签一致；
- 1 和 10 均为聚类中心，4 号点的类别标签应该和与其距离最近的、密度高于它的点一致，因此 4 号点属于聚类中心 1；
- 由于 5 号点最近的密度比其高的点为 4 号点，因此其类别标签与 4 号相同，也为聚类中心 1。