Density Peak Clustering 实验报告

PB20010429 侯相龙

2022 年 12 月 11 日

1 实验目的

k-means 算法无法处理非球形聚类的情况，而 DBSCAN 的实现对阈值选择的要求很高且计算代价大，为了克服这两种算法的缺点我们采用了 Density Peak Clustering 算法（以下简称 DPC），它对于超参数的选择要求不高，且能够处理非球形聚类的问题。

2 实验原理

实验基于以下假设：

1) 聚类中心周围密度较低，中心密度较高

2) 聚类中心与其它密度更高的点之间通常都距离较远

我们分别基于这两点假设引入了两个量来度量点的性质：局部密度 ρ 和与更高密度点的距离 δ_i.

$$\rho_i = \sum_j \chi\left(d_{ij} - d_c\right) \qquad \delta_i = \min_{j:\rho_j > \rho_i} d_{ij}$$

其中，$\chi(x)$ 为 x<0 的示性函数。

于是，基于假设我们应该找密度大且与其他簇类中心远的点作为中心点，即 ρ 和 δ 都较大的点

3 实验步骤

3.1 构建 DPC 类

1) 生成决策图、根据实验原理中的公式初始化 ρ, δ。

2) 确定中心点：这里有两种方式确定（分别作为两个函数）。其一，选择 $\rho * \delta$ 按从大到小排序，选择前指定数量的下标作为中心点；其二，选择 ρ和δ 在不小于指定值的下标作为中心点。

3) 分类。初始化中心点类别。对于未分类的值，选取密度大于它的最近邻的类别作为它的类别。这一点可以递归实现。

4) 计算 DBI。直接调用 sklearn 中的函数实现。

3.2 读取数据及训练模型

三组数据处理类似，且步骤同上一节**构建 DPC 类**。增加的是根据分类后的 label（对应颜色）作散点图

3.3 参数调试及比较

本模型的超参数有阈值 d_c、范数类型、中心点数目 num、确定中心点的 $\rho\delta$ 范围。依照 DBI 值和分类图像确定超参数选择。具体见**实验结果**部分。

4 实验结果

4.1 最优模型实验结果：

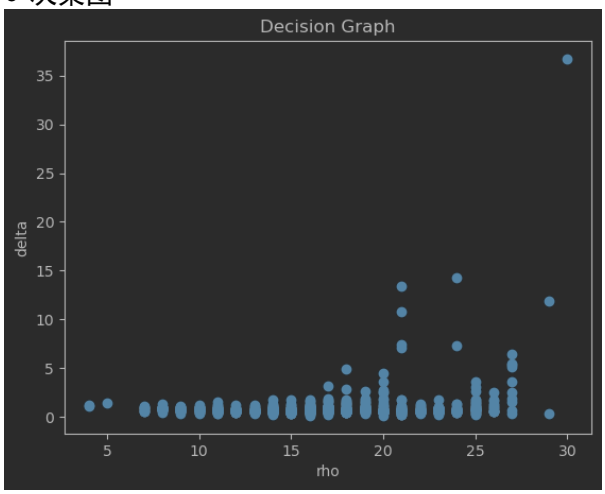
在用一些超参数进行测试后，得到了如下的实验结果

Aggregation

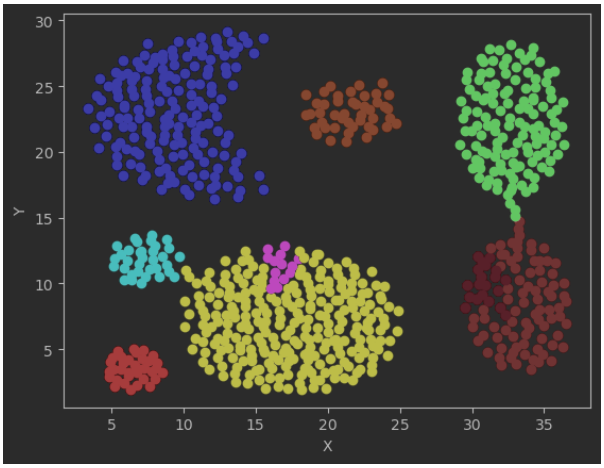
- 中心点及 DBI

```
中心点：  [768 340 602  47 721 191 254 553 552]
DBI= 0.8733462606576252
```

- 决策图



- 分簇结果

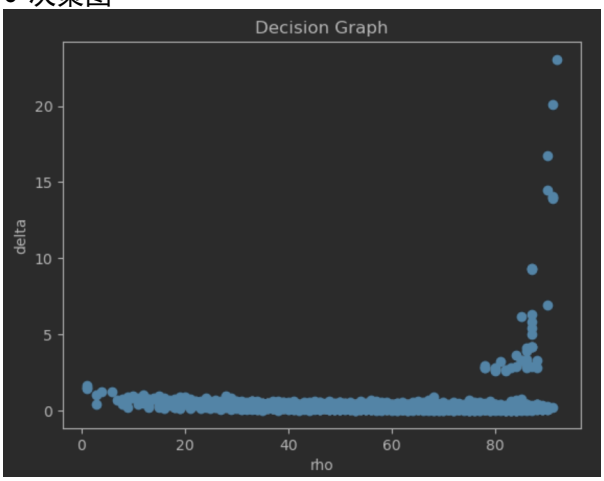


D31

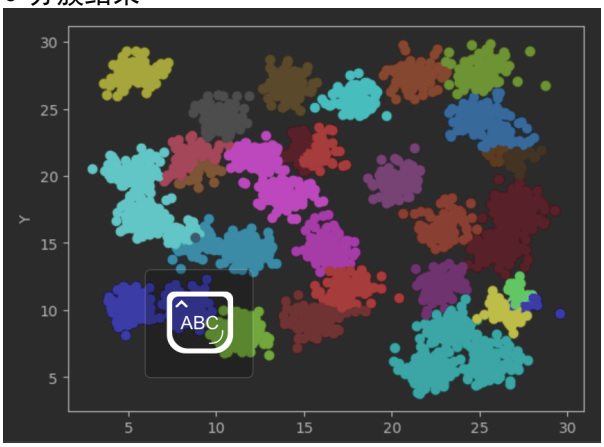
- 中心点及 DBI

```
中心点： [ 179.  512.  547.  588.  675.  895.  925. 1153. 1270. 1381. 1444. 1532.
 1568. 1593. 1830. 1933. 2069. 2098. 2111. 2240. 2324. 2330. 2343. 2428.
 2530. 2675. 2740. 2804. 2828. 2972.]
中心点数目： 30
DBI= 0.7788822311065383
```

- 决策图



- 分簇结果

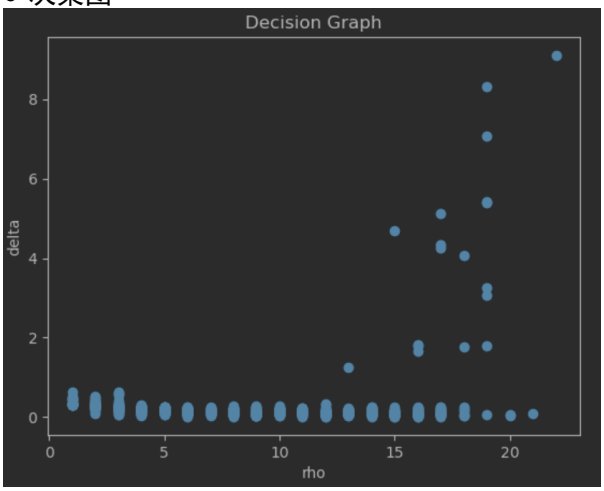


R15

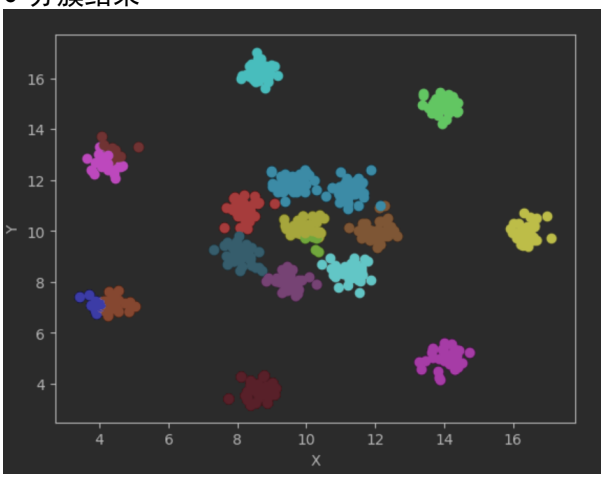
- 中心点及 DBI

```
中心点： [179 344 368 517 496 432 478 521 463 594 279 108 210 299  17  38  72]
中心点数目： 17
DBI= 0.6326347611755493
```

- 决策图



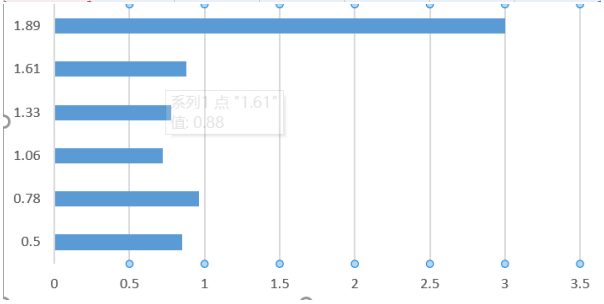
- 分簇结果



4.2 参数调试及比较

本实验测试了多组数据的多种不同超参数，为简便起见，下面只列出 D31 数据的部分调试结果。

d_c	0.5	0.78	1.06	1.33	1.61	1.89
DBI	0.85	0.96	0.72	0.78	0.88	3



5 实验分析

i) 实验的主要难点在于超参数的选取，我们可以通过“观察”与 DBI 指标结合的方式，使得分类结果既符合科学评估标准，也和人眼预测相同。

ii) 我们也可以根据观察 DBI 图进行参数的调试，即选取 DBI 图中的满足条件的孤立点，这样也能很好地提高分类的准确性。

iii) 由于时间有限，本实验还有很多超参数可以调试选取，（如选择中心点的方法），以达到更好的实验结果。