机器学习Lab3实验报告

DPC聚类实验

PB19020499 桂栋南

实验原理

DPC聚类根据两个特征来选择聚类簇的中心：

- 局部密度（local density）：$\rho_i = \sum_j \chi(d_{ij} - d_c)$
  - 其中d_c是超参数；
  - $\chi(x)$满足：$x < 0$时$\chi(x) = 1$；$x > 0$时，$\chi(x) = 0$；
  - d_{ij}是第i个点与第j个点之间的距离。
- 与高局部密度点的距离：$\delta_i = \min_{j:\rho_j > \rho_i} d_{ij}$
  - 如果该点是最高密度的点，即不存在$\rho_j > \rho_i$，则设$\delta_i = \max_j d_{ij}$；
- 聚类中心的选取即选择那些同时具有高ρ_i和δ_i的点。
- 异常点即为具有高δ_i但是低ρ_i的点。

实验实现

实验思路

1. 求出点的距离矩阵。
2. 根据距离矩阵获得所有点的ρ_i和δ_i值。
3. 画出决策图（横纵坐标分别为ρ_i和δ_i），选取合适的划分阈值。
4. 画出分类结果，求得最终的DBI。

具体实现

- 求距离矩阵（发现调库运算比自己手写的快速超级多）

```
from scipy.spatial.distance import pdist
from scipy.spatial.distance import squareform

path1 = './datasets/Aggregation.txt'
res_list = process_data(path1)
n = len(res_list)
dist = pdist(res_list, metric='euclidean')
dist = squareform(dist)
```

- 求ρ_i

```
def get_rou(dist, dc):
    return np.count_nonzero(dist - dc < 0, axis=0)
```

- 求δ_i

```
def get_sigma(dist, rou):
    n = len(rou)
    res = []
    for i in range(n):
        temp = dist[i, rou - rou[i]>  0]
        if len(temp) > 0 :
            res.append(np.min(temp))
        else:
            # 最大密度点的情况
            res.append(np.max(dist[i]))

    return np.array(res)
```

- 可视化

```
def plot_pic2(x_list, y_list, center_points, cluster=None):

    print("共有{}个聚类中心".format(len(np.unique(cluster))))
    #根据聚类结果着色
    plt.scatter(x_list, y_list, s=2, c=cluster)
    #对于聚类中心点着色
    plt.scatter(x_list[center_points], y_list[center_points], s=5, c='r')

    plt.show()
```

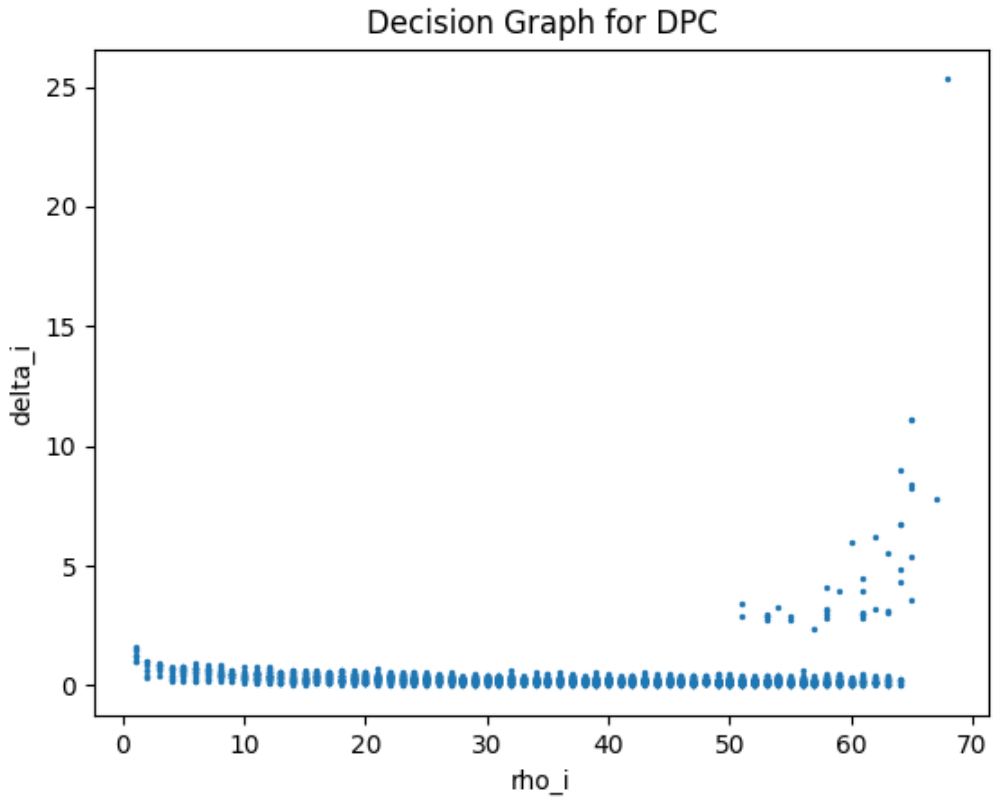
- 调用函数获得聚类结果

```
rou = get_rou(dist, dc=3)
sigma = get_sigma(dist, rou)
plot_pic(rou, sigma)

# 需要根据结果手动修改的参数
rou_threshold = 38
sigma_threshold = 8

temp1 = np.argwhere(rou - rou_threshold > 0)
temp2 = np.argwhere(sigma - sigma_threshold > 0)
# 获取所有满足两个条件的聚类点
center_points = np.intersect1d(temp1, temp2)

cluster = []
for i in range(n):
    # 根据最近的聚类中心点确定每个点所在聚类
    cluster_i = center_points[np.argmin(dist[i, center_points])]
    cluster.append(cluster_i)
```

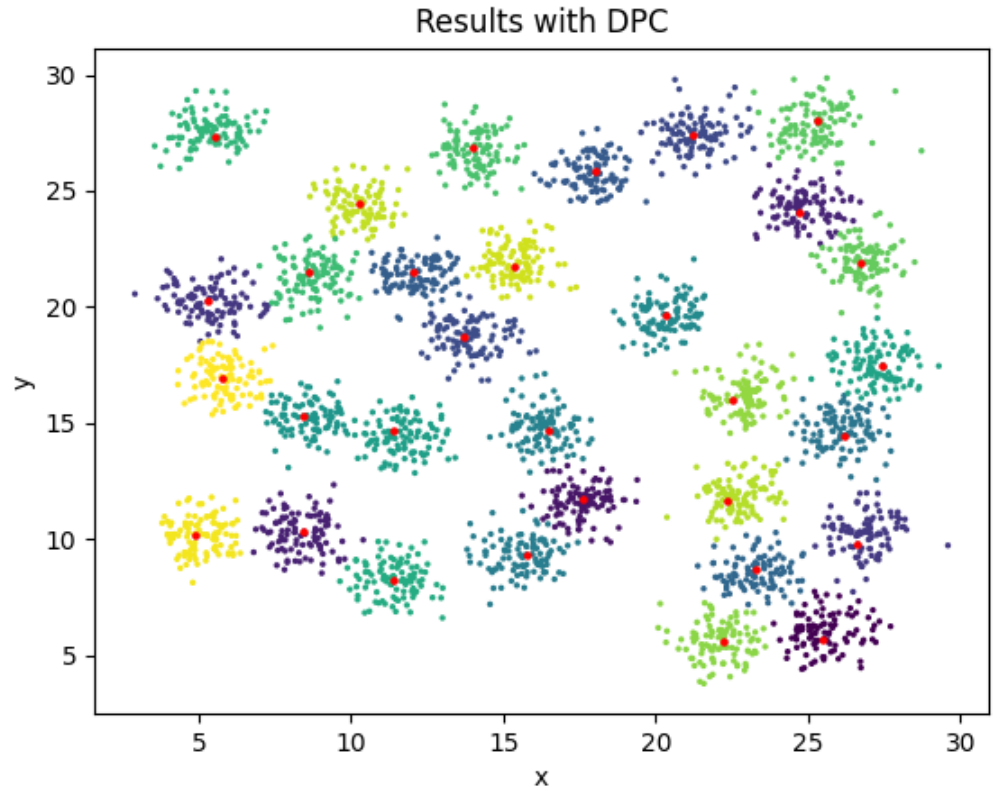
实验结果

D31

- 选取$d_c = 1, \rho^* = 50, \delta^* = 2$
- 决策图为



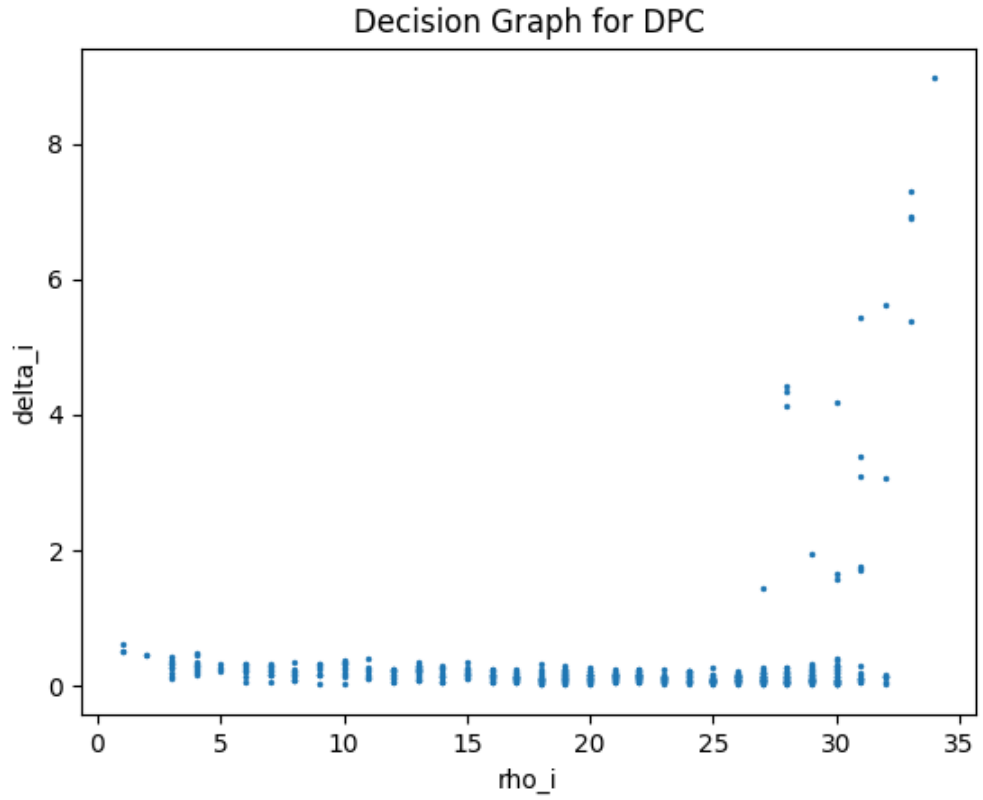
- 结果为(红色为聚类中心，下同)



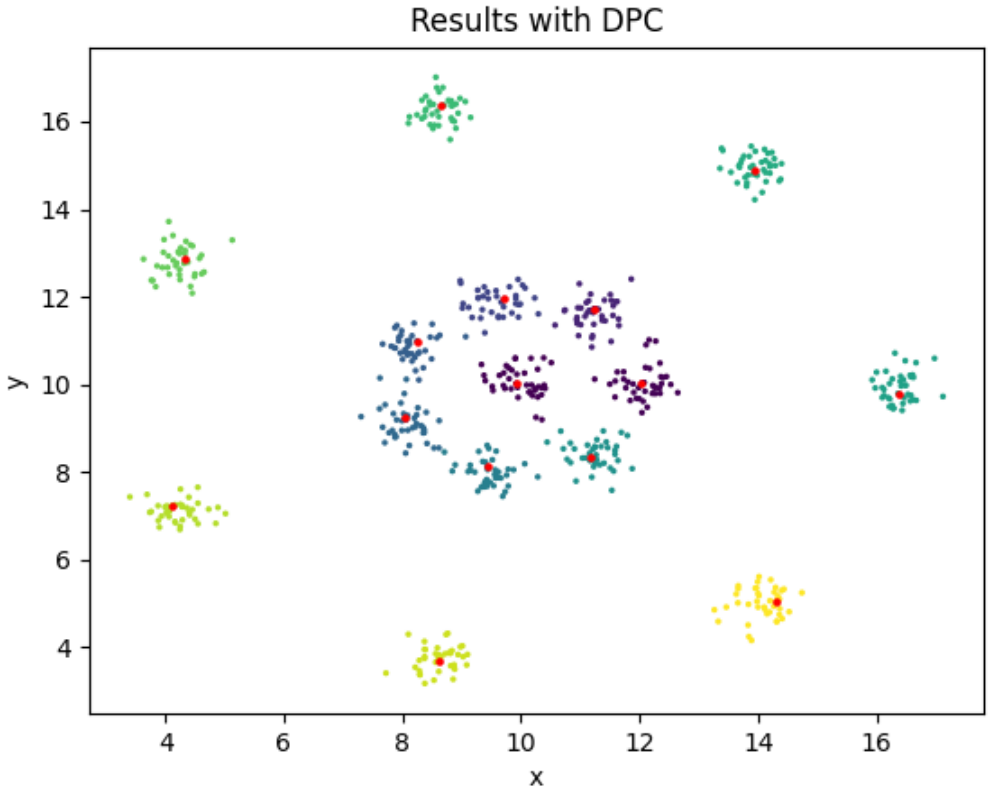
- DBI指数为：0.5479834835731773

R15

- 选取$d_c = 0.5, \rho^* = 25, \delta^* = 1$
- 决策图为



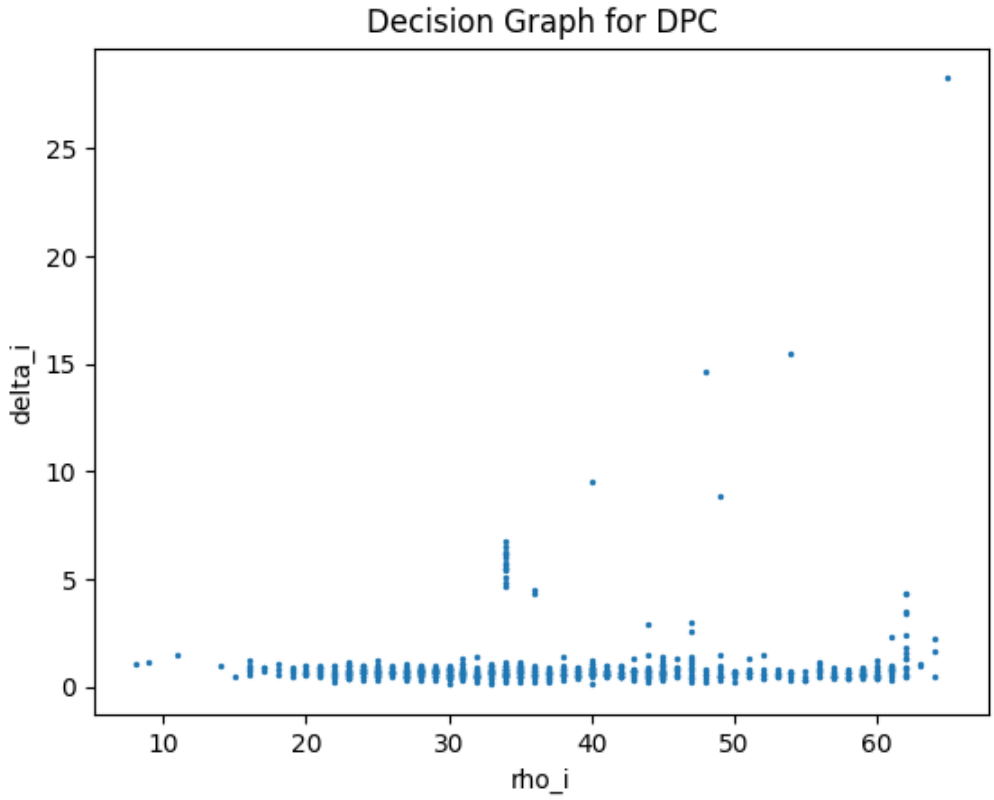
- 结果为



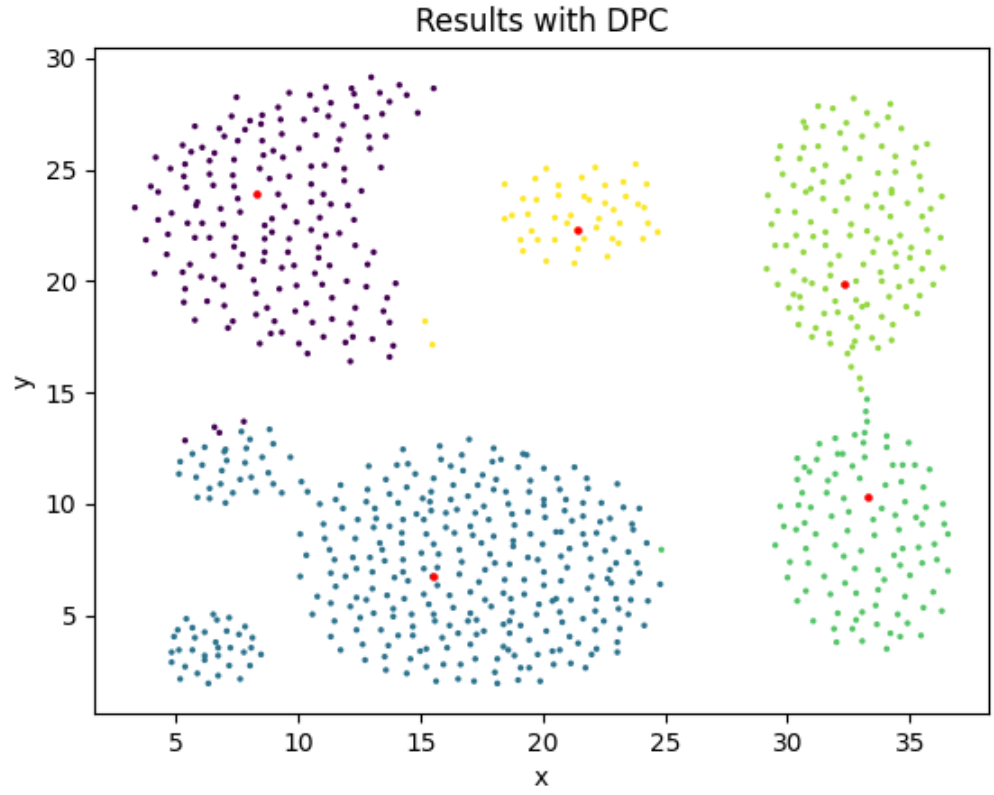
- DBI指数为：0.31496549541789415

Aggregation

- 选取$d_c = 3, \rho^* = 38, \delta^* = 8$
- 决策图为



- 结果为



- DBI指数为：0.5396045963949283