数据处理中的矩阵方法-思考题

李厚华-202418019427056

一、数据描述

数据是 1400 个二维点（x, y），从图像上看，数据具有如下特点：

- 明显的簇结构：数据聚集成多个紧密的簇，大致呈放射状分布。
- 中心密集，外围分散：中间有一个比较密集的中心簇，外围有多个分散的子簇。
- 各簇间间隔明显：各个簇之间有较大的间距，适合聚类分析。

这类数据很适合使用密度或基于距离的聚类方法。

二、聚类方法及原理

1. DBSCAN（基于密度的聚类）

- **原理**：
  - 通过定义邻域半径（`ε`）和最小点数（`min_samples`），将高密度区域划分为簇，低密度区域视为噪声。
  - 能识别任意形状的簇，无需预先指定簇数量，适合复杂分布。
- **适用场景**：
  - 数据中存在不同密度的簇（如星系中心密集、外围稀疏）。
  - 需要自动排除噪声点。
- **参数建议**：
  - 通过尝试不同的`ε`和`min_samples`（例如，`ε=1.5`，`min_samples=5`），观察簇的分离效果。

2. K-Means（基于质心的聚类）

- **原理**：
  - 预先指定簇数量（K），通过迭代优化质心位置，使簇内距离最小化。
  - 假设簇为凸形且大小相似，计算效率高。
- **适用场景**：
  - 数据分布呈球形或椭球形（如多个独立星团）。
  - 已知或能预估簇数量（例如，通过手肘法或轮廓系数确定K值）。
- **参数建议**：
  - 使用手肘法或轮廓系数确定最佳K值（例如，K=5或6）。

三、实施步骤

1. **数据预处理**：
   - 标准化数据（使用`StandardScaler`），避免特征尺度差异影响距离计算。
2. **可视化探索**：
   - 绘制散点图，观察数据分布形态，辅助选择聚类方法（如是否含噪声、簇形状）。
3. **参数调优**：
   - **DBSCAN**：通过尝试不同`ε`和`min_samples`，结合轮廓系数评估质量。
   - **K-Means**：人工确定最佳K值。
4. **聚类与评估**：

- 可视化聚类结果，验证是否符合预期（如颜色标记不同簇和噪声）。

四、实验结果

1. **DBSCAN**：
   - 设置邻域半径分别为3、3.2和5，最小点数为10，其中：
     - 邻域半径选择为3.2时准确分出了31类；
     - 邻域半径选择为3时分出了32类，在中心区域点处出现了错误聚类结果；
     - 邻域半径选择为5时分出了7类，当继续增大或缩小邻域半径时仍不能形成准确聚类结果。

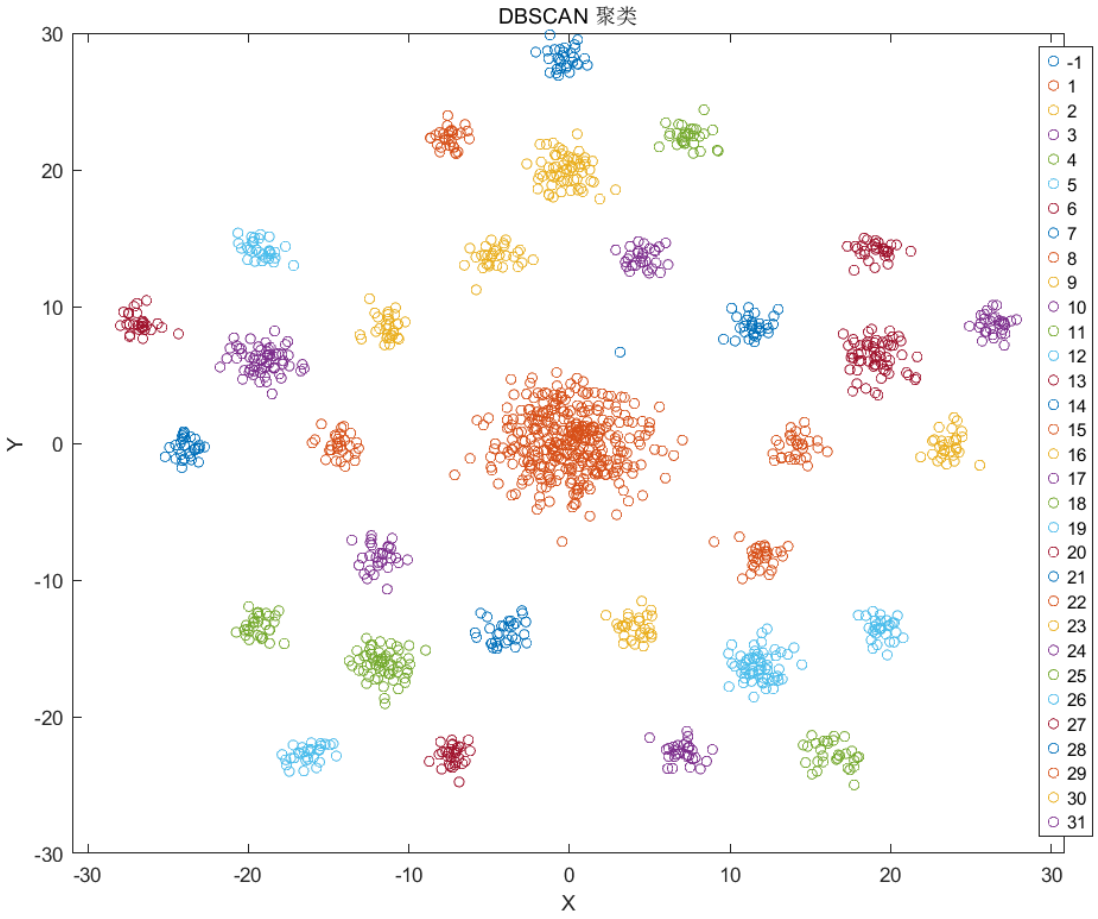

图1：邻域半径分别为3

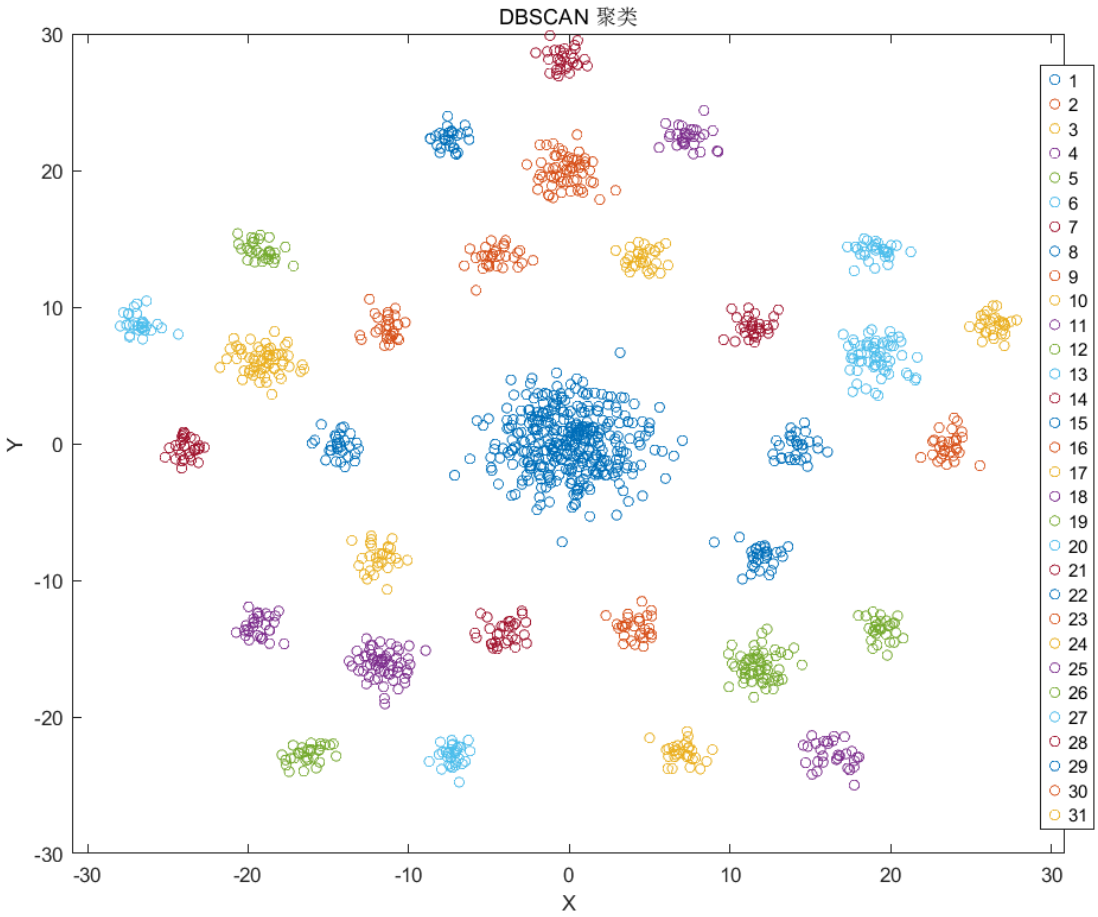

图2：邻域半径分别为3.2

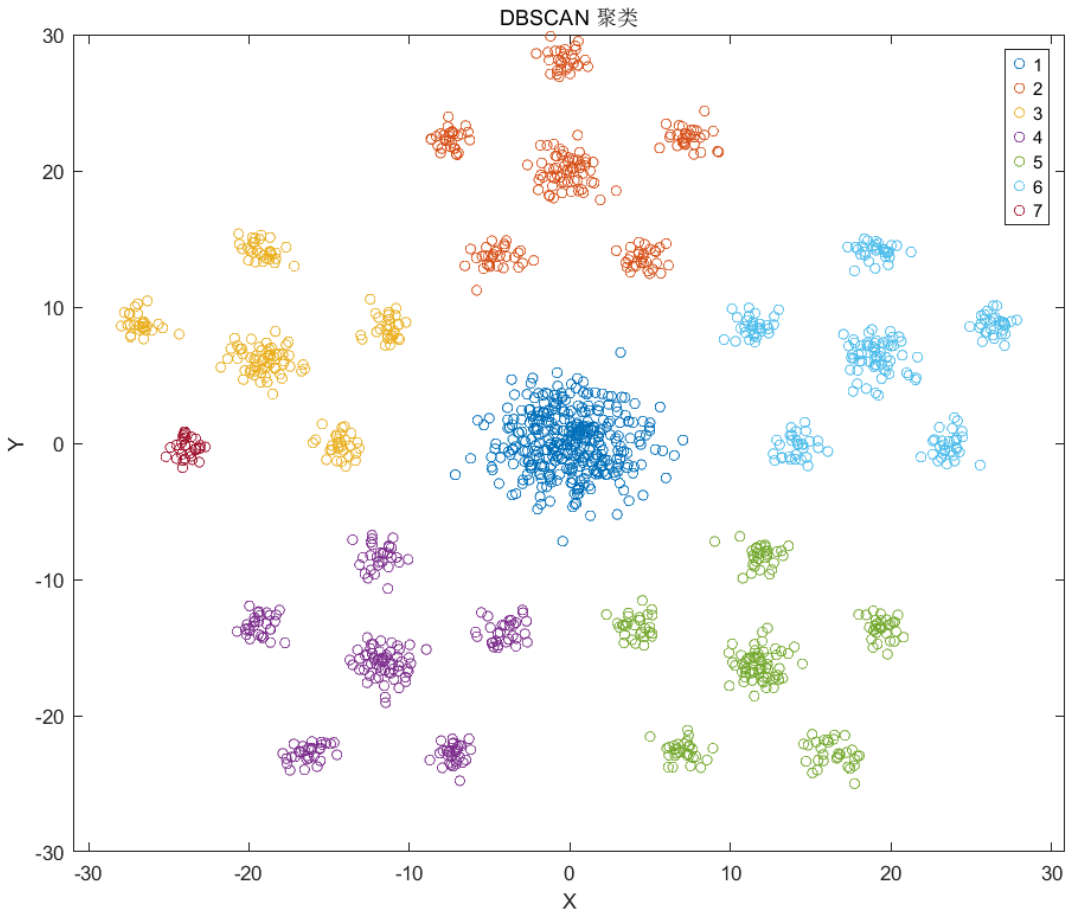

图3：邻域半径分别为5

1. **K-means**：
   - 分别设置簇数量为6和31，K-Means都能准确聚类。

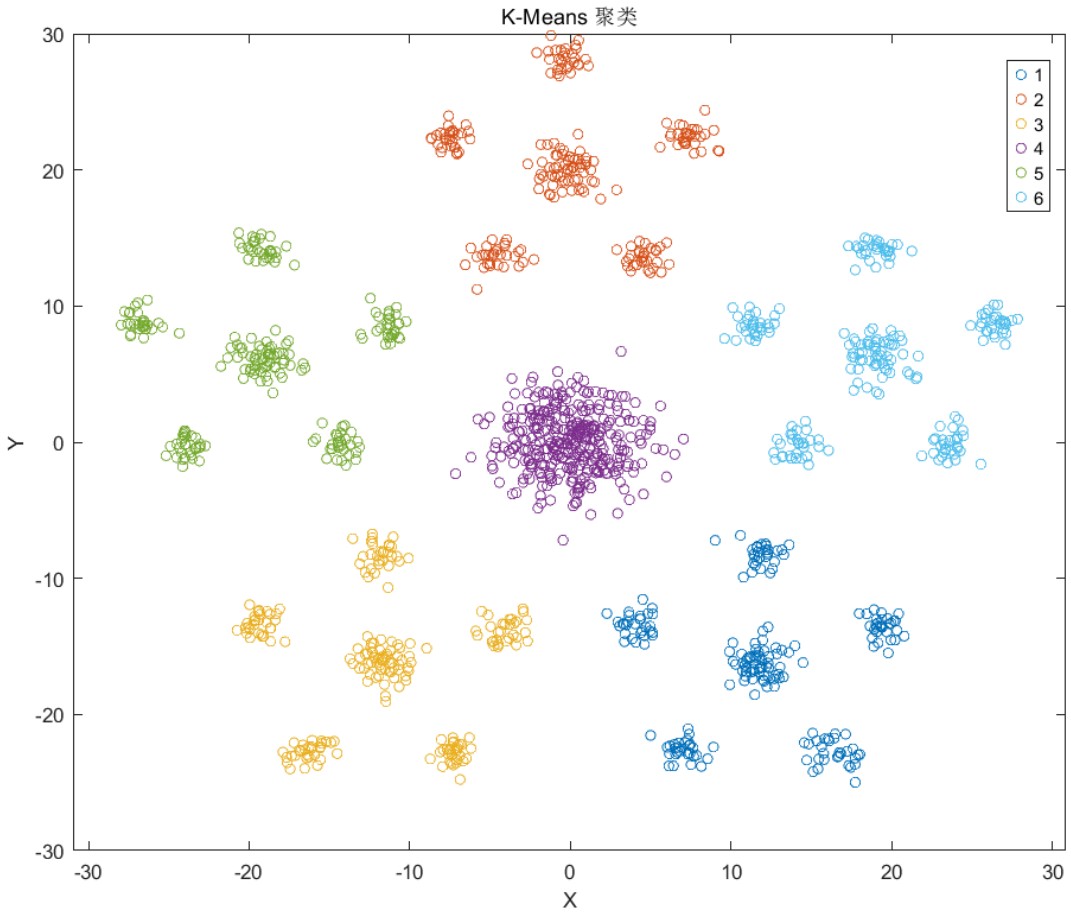

图4：簇数量为6

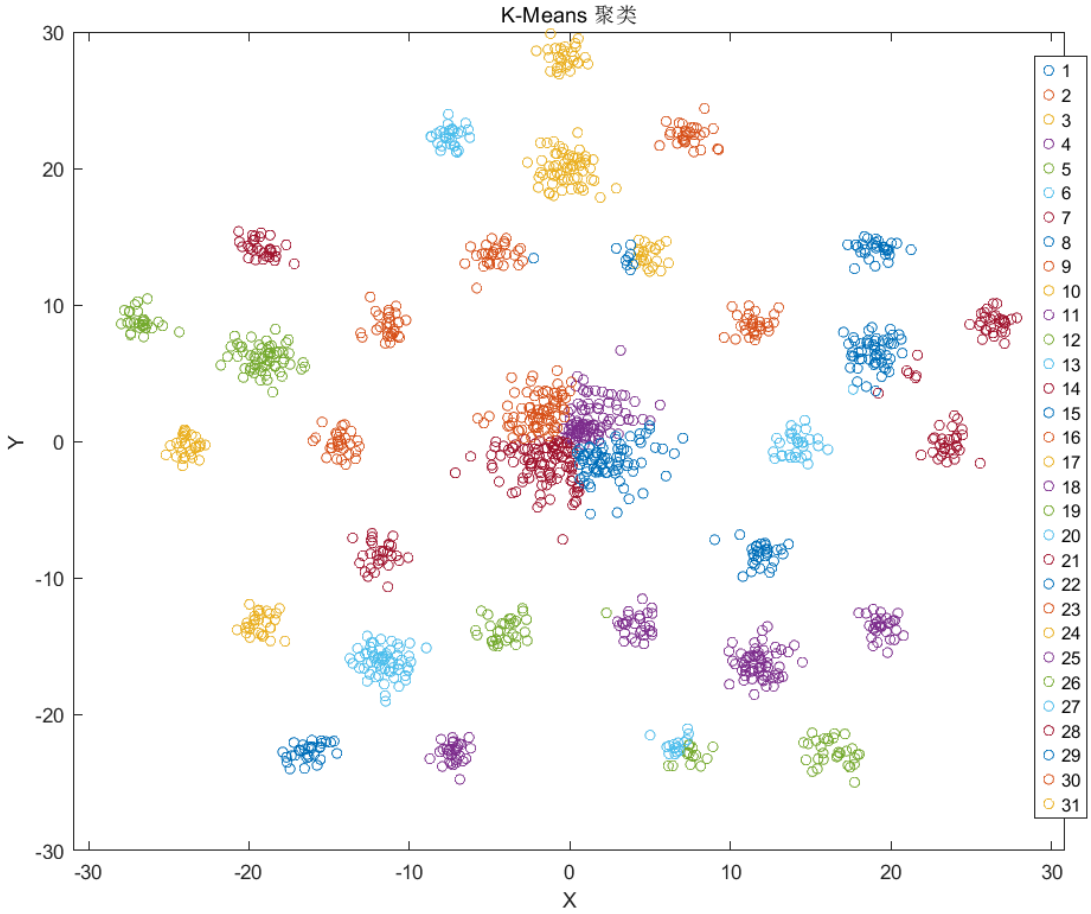
K-Means 聚类
30
20
10
0
-10
-20
-30
Y
-30
-20
-10
0
10
20
30
X
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31

图5：簇数量为31