華中科技大学

课 程 实 验 报 告

课程名称： 大数据分析

专业班级： 计科 2003 班

学　　号： U202015374

姓　　名： 张隽翊

指导教师： 崔金华

报告日期： 2022 年 12 月 28 日

计算机科学与技术学院

目录

实验四 kmeans 算法及其实现

4.1 实验目的

1、加深对聚类算法的理解,进一步认识聚类算法的实现；
2、分析 kmeans 流程,探究聚类算法院里；
3、掌握 kmeans 算法核心要点；
4、将 kmeans 算法运用于实际，并掌握其度量好坏方式。

4.2 实验内容

提供葡萄酒识别数据集(WineData.csv)，数据集已经被归一化(normalizedwinedata.csv)。同学可以思考数据集为什么被归一化，如果没有被归一化，实验结果是怎么样的，以及为什么这样。

同时葡萄酒数据集中已经按照类别给出了 1、2、3 种葡萄酒数据，在 csv 文件中的第一列标注了出来，大家可以将聚类好的数据与标的数据做对比。

编写 kmeans 算法，算法的输入是葡萄酒数据集，葡萄酒数据集一共 13 维数据，代表着葡萄酒的 13 维特征，请在欧式距离下对葡萄酒的所有数据进行聚类，聚类的数量 K 值为 3。

在本次实验中，最终评价 kmean 算法的精准度有两种，第一是葡萄酒数据集已经给出的三个聚类，和自己运行的三个聚类做准确度判断。第二个是计算所有数据点到各自质心距离的平方和。请各位同学在实验中计算出这两个值。

实验进阶部分：在聚类之后，任选两个维度，以三种不同的颜色对自己聚类的结果进行标注，最终以二维平面中点图的形式来展示三个质心和所有的样本点。效果展示图可如图 1 所示。

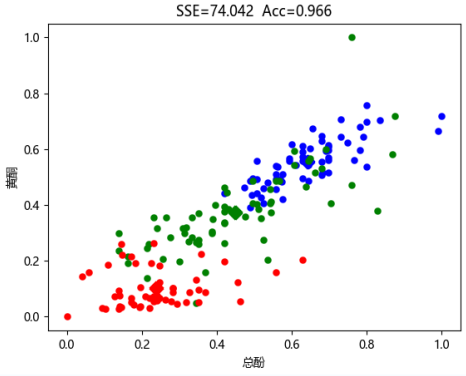

图 1 葡萄酒数据集在黄酮和总酚维度下聚类图像（SSE 为距离平方和，Acc 为准确率）

4.3 实验过程

本实验采用 kmeans 算法对葡萄酒数据进行分析，设置 3 个质心进行聚类操作。kmeans 算法的流程图如图 1.1 所示。

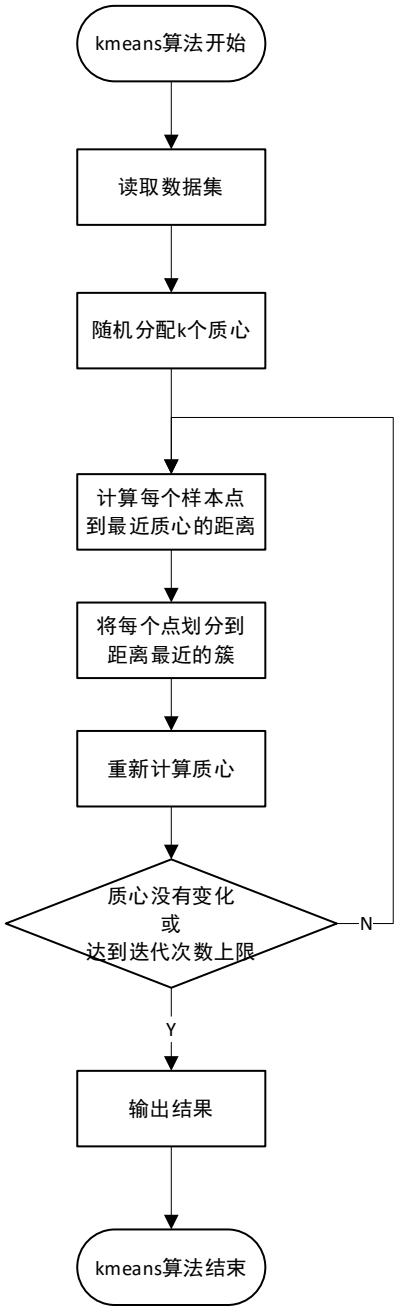

图 1.1 kmeans 算法流程图

4.3.1 编程思路

（1）数据读取

采用 numpy 二维矩阵存放每个样本点的数据，第一个数据固定为类型，后面数据为属性。

（2）类簇更新

初始化 k=3 个随机质心，使用 numpy.random 生成。由于原始数据已经进行了归一化处理，可以将质心的各个维度随机范围设置在 0~1 之间。

迭代循环前，建立二维数组存放每个样本点到所属质心的距离平方和以及所属的类簇编号。循环过程设置更新标志，在更新过程中若质心发生了变化则将该标志设置为 true，否则保持为 false。当没有发生更新或者已经达到事先设置的最大迭代次数时，退出循环过程。

循环结束后，计算准确度 acc 和距离和 sse。

（3）结果生成

使用 matplotlib.pyplot 库函数中的 scatter 函数绘制散点图，将每个类簇内的样本点设置成相同的颜色。绘图与效果展示图一致，选择总酚和黄酮这两个维度，前者为横坐标，后者为纵坐标。

4.3.2 遇到的问题及解决方式

（1）循环初始化

每次进入循环，需要将最小距离和对应的聚类中心初始化，否则进行的是无效更新。

（2）绘制散点图时中文乱码

使用 matplotlib 绘制散点图时，需要设置额外的中文字体，解决乱码问题。本人的设置如下。

```
# ? 画图字体设置
plt.rcParams["font.sans-serif"] = ["SimHei"] # 设置字体
plt.rcParams["axes.unicode_minus"] = False # 该语句解决图像中的“-”负号的乱码问题
```

4.3.3 实验测试与结果分析

运行 kmeans.py 文件，计算输出如图 1.2 所示。

```
/home/Asuna/anaconda3/bin/python /home/Asuna/Documents/VSCode Files/BDA2022/LAB4-Kmeans/src/kmeans.py
iteration times:  6
The  1 41.15008926751599
The  2 23.180457222344923
The  3 25.090542046973688
All sse:  89.42108853683459
acc:  0.9943820224719101
All done.
```

图 1.2 kmeans.py 输出结果

上述结果表明，一共迭代了 6 次，总的 SSE 为 89.24，ACC 为 0.9943。

葡萄酒数据集在黄酮和总酚维度下的聚类图像如图 1.3 所示。

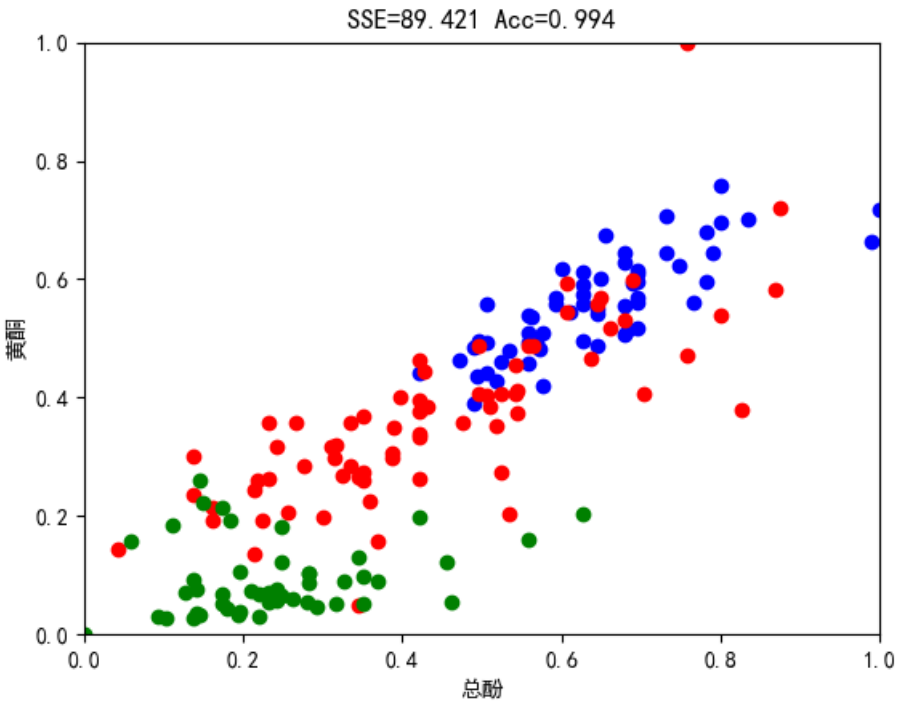

图 1.3 葡萄酒数据集再黄酮和总酚维度下聚类图

综合上述结果，生成的聚类图像与示例图较为相似，满足要求。

4.4 实验总结

本次实验实现了 kmeans 聚类算法。该算法具有以下优点：

- 容易理解，聚类效果不错，保证局部最优；
- 处理较大数据集合时可以保证较好的伸缩性；
- 当簇近似于高斯分布时效果很好；
- 算法的复杂度相对低。

当然，kmeans 算法也存在缺点：

- K 值需要人为设定，不同的 K 值得到的结果不一样；
- 对初始的簇中心敏感，不同的选取方式会得到不同的结果；
- 对异常值比较敏感；
- 不适合过于离散的分类、样本类别不平衡的分类、非凸形状的分类。

在本实验中对数据进行归一化处理主要考虑两个因素。

①随机分配质心时，可以在 0~1 之间随机。如果不进行归一化操作，数据间的间隔太大，随机维度数据的间隔不好把握，误差较大；

②减少迭代次数。如果不进行归一化操作，随机步长较大，可能在很多次循环之后才能保证质心稳定，程序效率较低。