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： **CS2011**

学　　号： **U202014774**

姓　　名： 王逸

指导教师： 崔金华

报告日期： **2022.12.28**

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 cvs 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。

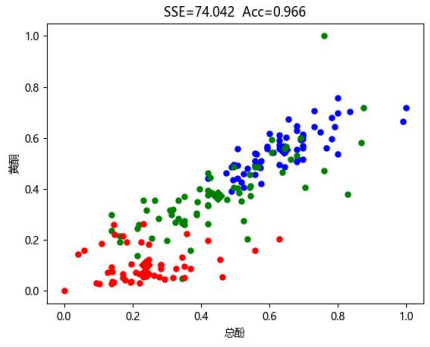
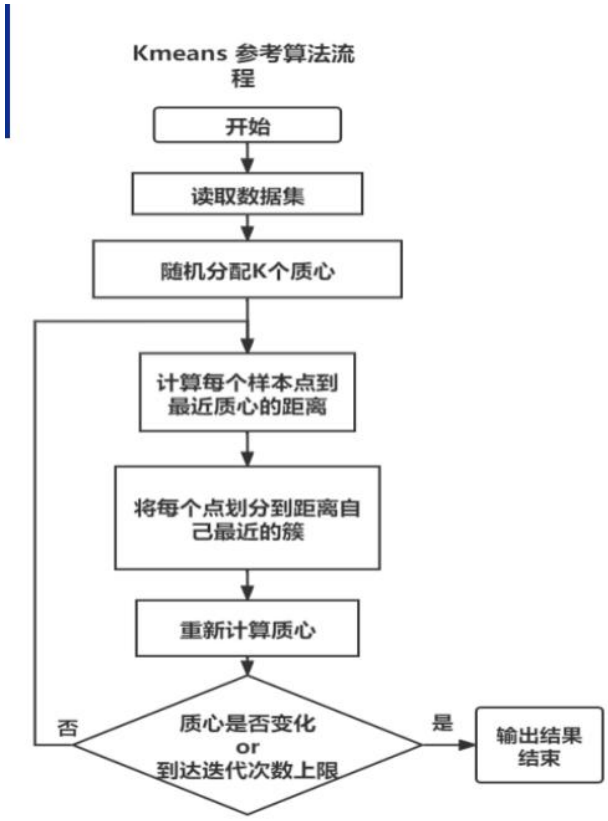


图 1 葡萄酒数据集在黄酮和总酚维度下聚类图像（SSE 为距离平方和，Acc 为准确率）

4.3 实验过程

4.3.1 编程思路

本次实验总体流程如下：



实验数据由一列标签和 13 列特征组成，在训练时只需考虑后 13 列的特征维度，因此初始化质心时，应随机生成 13 个数据构成一个质心向量。在 kmeans.py 中定义了以下几个函数：

1) calculate_dis(centroid, data)：计算一个数据点到给定质心的欧氏距离。
2) read_data(data_file)：读取数据文件，得到特征数组。
3) init_centroids(k)：初始化 k 个质心，采用随机数生成的方法，一个质心有 13 维特征。
4) calculate_sse(d_min, num)：计算每个聚类的 sse。传入的 d_min 数组是一个二维数组，第 0 列记录的该点所处于的 cluster 的质心标签，第 1 列记录了该点到该质心的最短距离。
5) get_cluster_points(data, d_min, num, centroid)：获取给定的质心的聚类里的所有数据点。
6) calculate_acc(data, d_min, num, k)：计算正确率。由于自行自动分类的算法没有将标签作为特征纳入考虑，所以在这里取每个聚类每个标签最大数作为该聚类的 hit 数，最后三个聚类 hit 相加，除以数据总数，获得正

确率。

7) plot_clusters(data, d_min, centers, num, k)：绘图函数。

需要强调的是，由于本次实验并没有给出判断质心点移动的最小距离，所以在判断质心是否变化时采用判断是否有点在类与类之间的移动，若没有移动则训练完成退出循环。

4.3.2 遇到的问题及解决方式

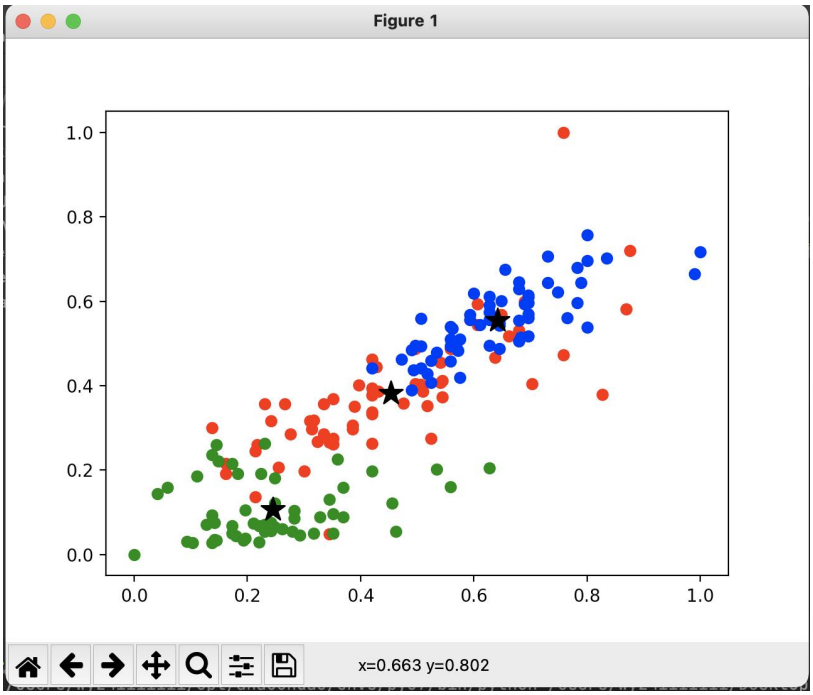
最开始在判断质心是否移动时采用的是通过质心前后距离变化，一开始设置完全相同时则判定循环结束，但由于程序计算时会有舍入误差导致计算结果差距较大；后面改为判断点的移动来判断质心是否变化。

4.3.3 实验测试与结果分析

程序输出结果：

```
第一个聚类的SSE为: 15.055849791999043, 数量为 54.
第二个聚类的SSE为: 13.934869641463003, 数量为 63.
第三个聚类的SSE为: 19.99812500014817, 数量为 61.
SSE为: 48.98884443361022.
准确度为: 0.9438202247191011.
```

绘制结果（与给出的参考事例 x、y 坐标轴相同）：



4.4 实验总结

本次实验主要学习了 kmeans 聚类算法，也对 pyplot 的画图函数进行使用，完成了结果的可视化。同时也掌握了两种判断质心变化的方法，并根据实际情况选择更为适合的一种作为最终的判断依据。