

**课 程 实 验 报 告**

**课程名称： 大数据分析**

**专业班级： CS1804（交换）**

**学 号： X2020I1007**

**姓 名： 刘日星**

**指导教师： 杨 驰**

**报告日期： 2021.12.25**

**计算机科学与技术学院**

**目录**

[实验四 kmeans算法及其实现 1](#_Toc91148749)

[**4.1实验目的** 1](#_Toc91148750)

[**4.2 实验内容** 1](#_Toc91148751)

[**4.3 实验过程** 2](#_Toc91148752)

[4.3.1 编程思路 2](#_Toc91148753)

[4.3.2 遇到的问题及解决方式 2](#_Toc91148754)

[4.3.3 实验测试与结果分析 2](#_Toc91148755)

[**4.4 实验总结** 4](#_Toc91148756)

# 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,探究聚类算法原理；

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，数据集已经被归一化。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1.1所示。



图4.1 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

葡萄酒数据集需要被归一化，这是因为不同维度属性的度量范围不同，而在分类中默认要均衡每一个属性的影响（当然不同的任务中不同属性也可能权重不同，具体任务具体分析）如果一个属性值范围过大或过小的话，会对分类产偏差。

Kmeans 算法流程说明如下：

首先随机选择n 个初始簇心，给它们分配编号1-n，然后进行循环迭代，计算每个点对每个簇心的距离，将其归类在最近的簇心的编号下。对数据进行每一轮循环后，重新计算不同簇的簇心位置，计算方式为取该簇所有点位置的平均值；循环在所有点的归类均不发生变化或者达到最大循环次数时退出。

Kmeans 算法原理比较简单，但也有很多地方需要注意。例如：初始簇心的

选取非常重要，不同的初始化方式对分类结果影响较大；计算时注意代码的效率问题，尽量使用numpy 进行矩阵加速;结果中很可能分类是比较正确的，但标签给的和原始数据中不匹配，这个也需要后期处理才能获得正确的分类准确率。

作图方面使用matplotlib.pyplot 模块作散点图展示即可。

### 4.3.2 遇到的问题及解决方式

常遇到的问题就是分类的正确性与标签不匹配，解决方法就是当分类的数大于2时，需要对分类结果的标签进行重新分配，直到找到最大的准确率为止。

### 4.3.3 实验测试与结果分析

运行结果如图1所示，经过6个回合的循环后，分类准确率达94.94%

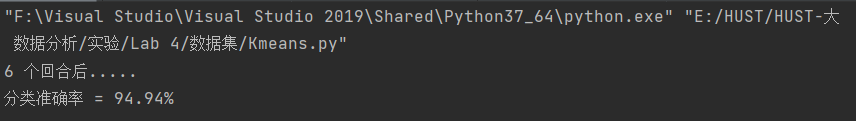


图1

所有数据点到各自质心距离的平方和SSE以及绘制的散点图（含质心），比较了dim1和dim3、dim2和dim4、dim5和dim6不同维度组合之间的分类效果，从下面三图对比中可以看出dim5和dim6这个维度组合的分类效果最为明显。

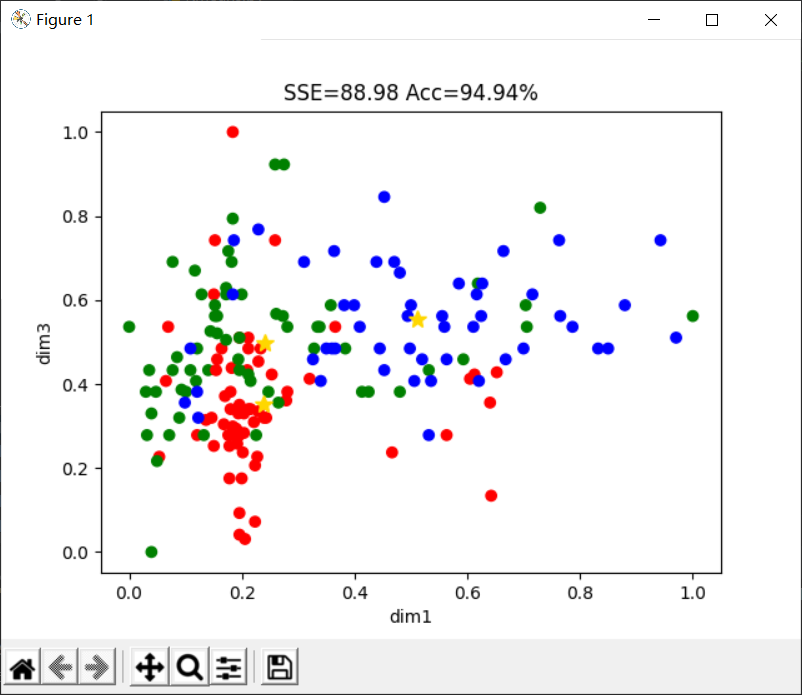


图2 dim1和dim3组合

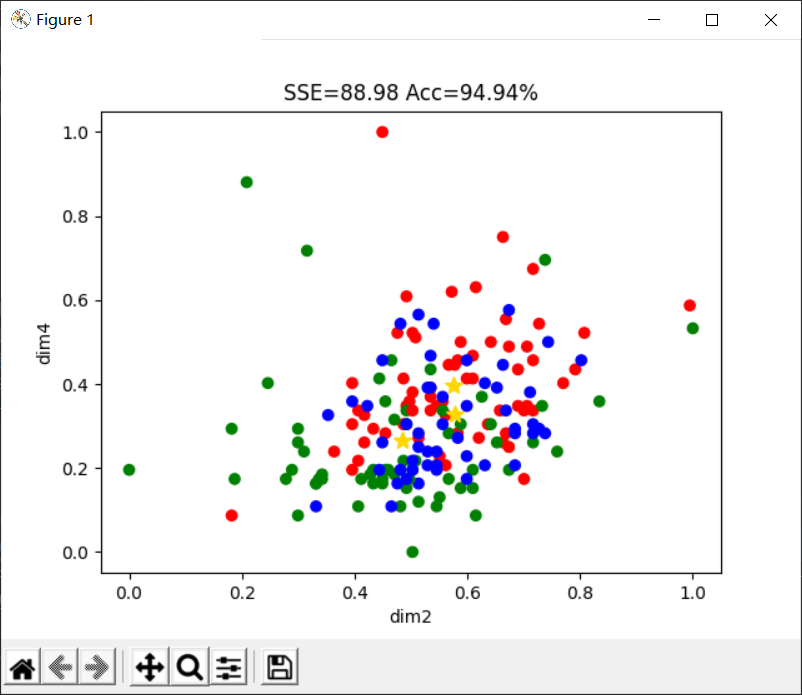


图3 dim2和dim4组合

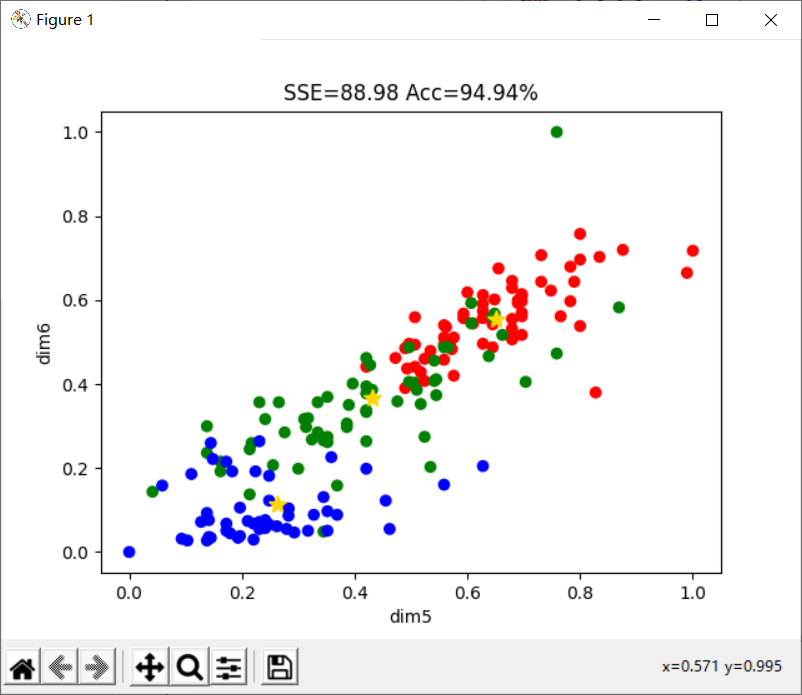


图4 dim5和dim6

## **4.4 实验总结**

本次实验采用的是Kmeans算法——K均值聚类算法，该算法是一种迭代求解的聚类算法，其实现步骤是预将数据分为3组，则随机选取3个对象作为初始的聚类中心（质心），然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。尽管本次实验过程中会遇到一些问题，但在大量查询网络资源后，还是逐步解决问题。最后还实现了一个功能，用户可以自由选择不同维度组合进行对比分类效果。总的来说，这次实验让我对Kmeans算法的实现有了更深的了解，不仅结合了课堂上老师讲的算法原理，以及实验上对算法的实践。