大数据分析实验
——聚类算法

目录

一 任务背景
二 任务描述
1. 任务一
2. 任务二
三 算法流程
四 验收流程
注意事项

一 任务背景

一 任务背景

- 1、加深对聚类算法的理解；
- 2、分析kmeans流程，探究聚类算法原理，掌握kmeans算法核心要点；
- 3、将kmeans算法运用于实际，并掌握其度量好坏方式。

二 任务描述

二 任务描述

◆任务一

数据集：提供葡萄酒数据集(WineData.csv)，数据集已经被归一化(normalizedwinedata.csv)。

编写kmeans算法，算法的输入是葡萄酒数据集，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。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。

二 任务描述

◆任务二

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。

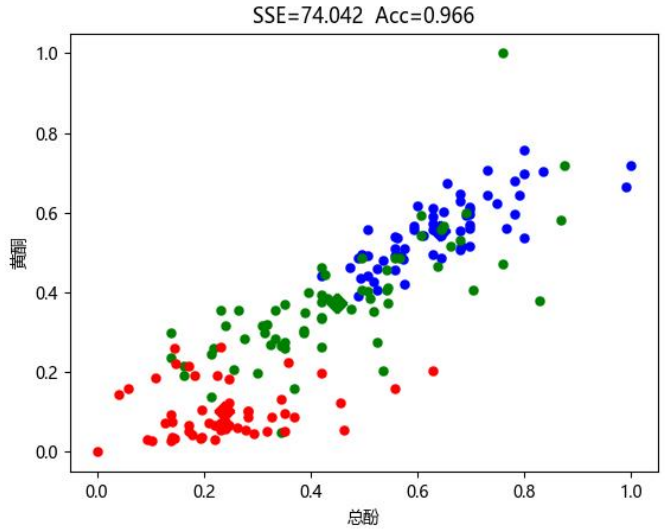
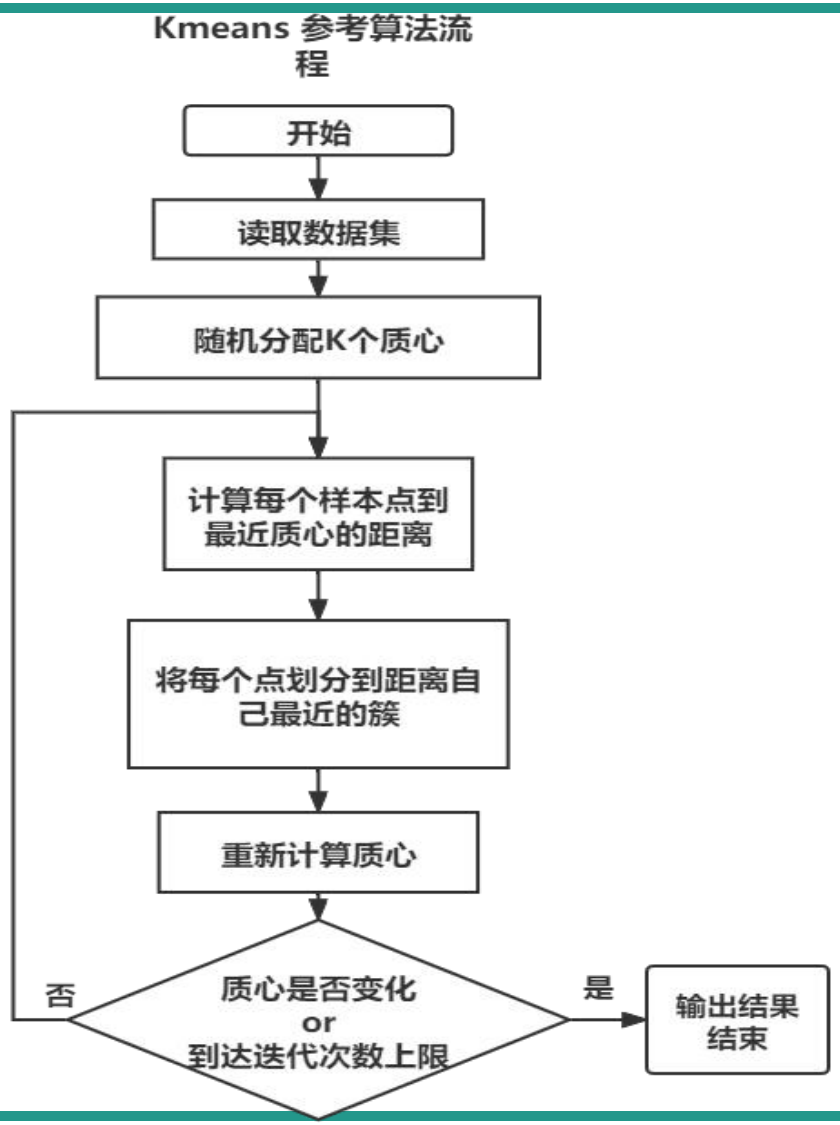


图1 葡萄酒数据集在黄酮和总酚维度下聚类图像（SSE为距离平方和，Acc为准确率）

華中科技大學
计算机科学与技术学院
School of Computer Science &Technology，HUST

三 算法流程

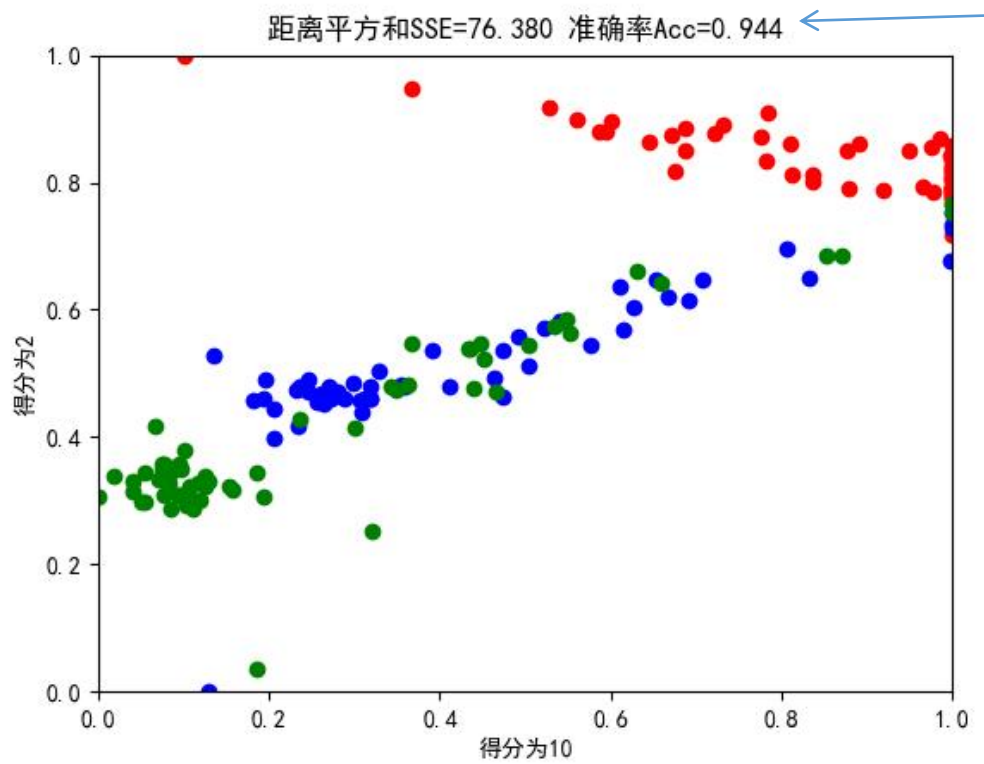
三 算法流程


Kmeans 参考算法流程
开始
读取数据集
随机分配K个质心
计算每个样本点到最近质心的距离
将每个点划分到距离自己最近的簇
重新计算质心
质心是否变化
or
到达迭代次数上限
否
是
输出结果
结束

華中科技大學
计算机科学与技术学院
School of Computer Science &Technology，HUST

四 验收流程

四 验收流程



任务一

任务二

四 注意事项

- 在选择K时可以多选择几组进行实验（3<=K<=10），注意在处理数据时有些列数据是Unknown，注意避免。
- 在选取不同Popularity的数据时，建议选取相隔距离较远的数据。
- 若实验效果不好时可以多进行几次实验选取较好的一次进行检查。
- 可以使用`matplotlib.pyplot`进行画图。
- 不能直接调用现有的聚类算法的库。