实验四 kmeans 算法及其实现

实验目的

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。

实验内容

数据集：提供葡萄酒数据集(WineData.csv)，数据集已经被归一化(normalizedwinedata.csv)。葡萄酒数据集一共 13 维数据，代表着葡萄酒的 13 维特征。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。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。

根据给定数据集，编写 kmeans 算法，完成以下任务：

1）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（ACC）判断。第二个是计算所有数据点到各自簇质心距离的平方和（SSE）。请各位同学在实验中计算出这两个值。

2）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。

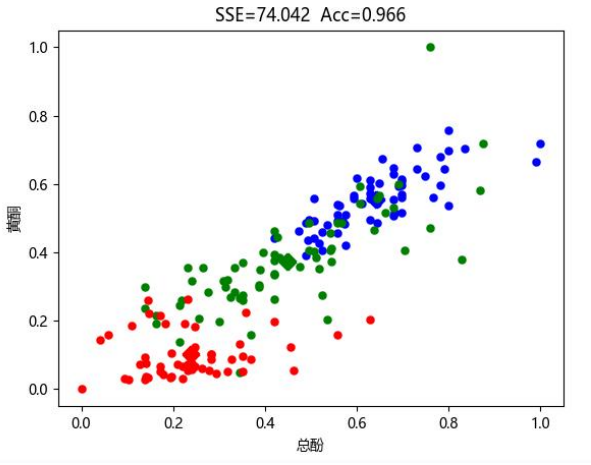

图 1 葡萄酒数据集在黄酮和总酚维度下聚类图像（SSE 为距离平方和，Acc 为准确度）