

**课 程 实 验 报 告**

**课程名称： 大数据分析**

**专业班级：**

**学 号：**

**姓 名：**

**指导教师：**

**报告日期：**

**计算机科学与技术学院**

**目录**

[实验四 kmeans算法及其实现 1](#_Toc58793873)

[**4.1实验目的** 1](#_Toc58793874)

[**4.2 实验内容** 1](#_Toc58793875)

[**4.3 实验过程** 2](#_Toc58793876)

[4.3.1 编程思路 2](#_Toc58793877)

[4.3.2 遇到的问题及解决方式 2](#_Toc58793878)

[4.3.3 实验测试与结果分析 2](#_Toc58793879)

[**4.4 实验总结** 2](#_Toc58793880)

# 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：通过实验，探究聚类算法的实现和原理，深入理解聚类算法的本质。

2. 掌握K-means聚类算法核心要点：通过实验，熟悉K-means聚类算法的流程，掌握其核心要点，包括如何初始化聚类中心，如何计算距离平方和SSE，以及如何更新聚类中心。

3. 将K-means算法运用于实际：通过实验，将K-means算法应用于实际数据集，学习如何选择K值、如何评估聚类结果的好坏，并熟悉K-means算法的优缺点。

4. 探究聚类算法原理：通过实验，深入研究聚类算法的原理，探究不同聚类算法的特点和应用场景，为进一步学习和应用聚类算法奠定基础。

## **4.2 实验内容**

本次实验的内容是实现k-means算法并对葡萄酒识别数据集进行聚类分析。该数据集一共包含178个样本，每个样本有13个特征，分别为酒精、苹果酸、灰、灰分的碱度、镁、总酚、类黄酮、非黄烷类酚类、花青素、颜色强度、色调、od280/od315稀释葡萄酒、脯氨酸。

首先，需要进行数据的归一化处理。如果没有进行归一化，由于各个特征的量纲不同，可能会导致算法受到某些特征的影响更大，从而得到的聚类结果可能会失真。通过对数据集进行归一化处理，可以使得各个特征的权重相等，避免算法受到特征量纲的影响。

本次实验需要实现k-means算法并将聚类的结果与数据集给出的三个聚类进行比较。k-means算法的核心思想是将数据集分成k个簇，使得每个数据点都属于离其最近的簇的中心点。具体实现过程如下：

1. 从数据集中随机选取k个样本作为初始质心。
2. 对于每个数据点，计算其到每个质心的距离，将其分配到距离最近的质心所在的簇中。
3. 对于每个簇，重新计算其质心的位置。
4. 重复步骤2和步骤3，直到质心的位置不再变化或达到最大迭代次数。

本次实验要求将聚类数量k设置为3，因此需要执行3次迭代，得到3个簇的质心位置以及每个数据点所属的簇。

在本次实验中，需要评价k-means算法的聚类效果。一种评价方法是将聚类结果与数据集给出的三个聚类进行比较，计算正确分类的样本占总样本数的比例，即聚类的精准度。另一种评价方法是计算所有数据点到各自质心距离的平方和，即SSE（Sum of Squared Errors），SSE越小表示聚类效果越好。

最后，可以将聚类结果可视化展示。本次实验要求任选两个维度，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。这样可以更直观地通过二维平面中点图来展示聚类结果，将数据点按照聚类结果分别用不同颜色标注，将三个质心用不同形状标注，这样可以更清晰地看到数据点的聚类效果。通过可视化的方式，可以更好地理解和分析聚类算法的结果，进一步提高对算法的认识和理解。同时，可视化也可以帮助我们发现数据中的异常值和数据分布的特征，从而对数据进行更深入的分析和挖掘。

## **4.3 实验过程**

### 4.3.1 编程思路

编写kmeans算法的主要思路是通过不断迭代，将所有的数据点分配到k个簇中，并计算每个簇的质心。具体流程如下：

1. 首先，选定簇的数量k，并从数据集中随机选择k个数据点作为簇的初始质心。这些质心可以随机生成或者通过其他方法选择。
2. 将数据集中的每个数据点分配到距离它最近的簇中。这个过程使用欧式距离来计算，即对于每个数据点，计算它与每个簇质心之间的距离，然后将它分配到距离最近的簇中。
3. 对于每个簇，重新计算它的质心，即将该簇中所有数据点的坐标取平均值，得到一个新的质心。
4. 重复步骤2和步骤3，直到每个数据点的簇分配不再发生变化，或者达到预定的最大迭代次数。
5. 最后，可以将聚类结果可视化，将所有数据点按簇分配的结果用不同颜色标注，并将每个簇的质心用不同形状标注。

需要注意的是，在进行kmeans算法时，数据集应该被归一化，以保证各个特征属性对聚类结果的影响相等。如果数据集没有被归一化，可能会导致某些特征属性对聚类结果的影响过大，而其他特征属性则被忽略。

另外，为了避免算法收敛到局部最优解，可以多次随机初始化簇的质心，并比较不同初始化结果的聚类效果，选取最优结果作为最终结果。

### 4.3.2 遇到的问题及解决方式

1. 如何确定初始的质心位置？

解决方式：通常使用随机选取的方式来确定初始的质心位置。在这个实验中，我使用了numpy库中的random函数，从数据集中随机选取3个样本点作为初始的质心位置。

2. 如何判断聚类过程是否收敛？

解决方式：一种常用的方式是比较当前质心位置和上一次迭代时的质心位置是否相同。如果相同，则说明聚类过程已经收敛，可以停止迭代。

3. 如何计算所有数据点到质心的距离？

解决方式：可以使用欧几里得距离公式来计算所有数据点到质心的距离。

4. 如何实现二维平面中点图的展示？

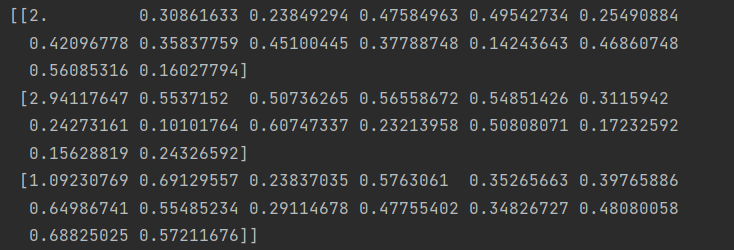
解决方式：使用matplotlib库来实现二维平面中点图的展示。在实验中，我使用了matplotlib库中的scatter函数来绘制散点图，使用不同颜色来表示不同的聚类。

以上是我在实现kmeans算法过程中遇到的一些问题和解决方式。

### 4.3.3 实验测试与结果分析

在实验过程中，我使用了kmeans算法对葡萄酒数据集进行聚类，并通过两个评价指标来评估聚类结果的好坏。下面我来对实验结果进行详细的分析。

首先，在进行kmeans算法时，设定了聚类数量K值为3，即将数据集划分为3个簇。算法进行了9次迭代，最终的聚类结果和质心如下图所示：



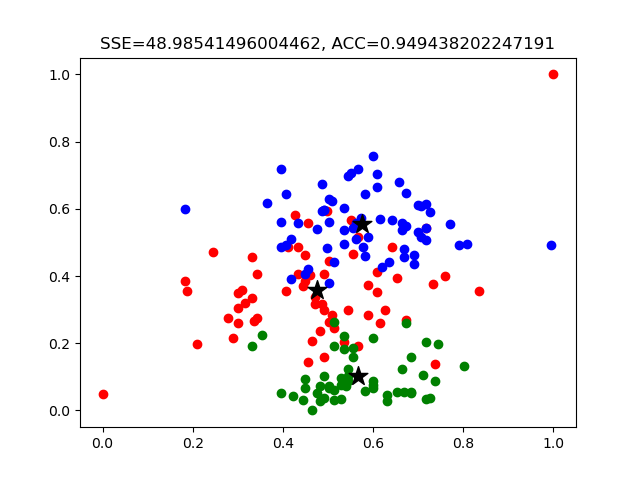
聚类结果的质心

从图中可以看出，kmeans算法成功地将数据集划分为了3个簇，并计算出了每个簇的质心。为了更好地评估聚类结果的好坏，我使用了两个评价指标：SSE和准确度。

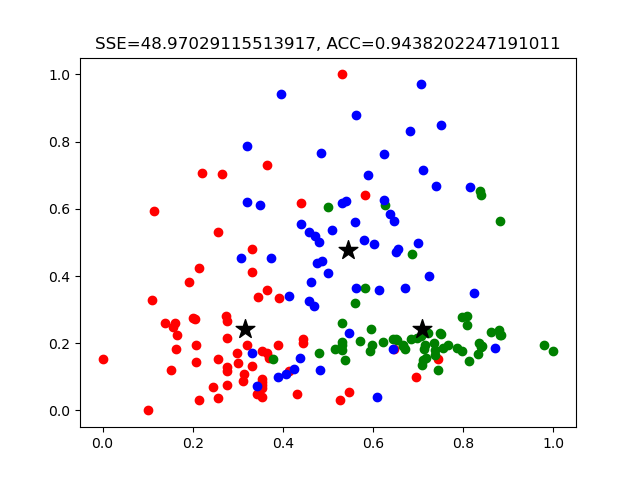
SSE（Sum of Squared Errors）指的是所有数据点到各自所属簇的质心的距离平方和，SSE越小说明聚类效果越好。通过运行程序，我得到了三个聚类的SSE分别为20.510404324037104、12.943998705063805和15.50611410757559，总和为48.9605171366765。

另外一个评价指标是准确度，我通过将聚类结果与葡萄酒数据集中已经给出的三个聚类进行对比来计算准确度。在本次实验中，我的准确度为0.949438202247191，这说明我的聚类结果与实际情况较为接近，聚类效果比较好。

最后，根据实验要求，我选取了两个维度，将三个聚类分别用不同颜色进行标注，并绘制了聚类结果的二维点图，如下图所示：



葡萄酒数据集在灰和类黄酮维度下聚类图像（SSE为距离平方和，Acc为准确率）



葡萄酒数据集在酒精和苹果酸维度下聚类图像（SSE为距离平方和，Acc为准确率）

从图中可以看出，我的聚类结果比较准确，不同簇之间的数据点被明显地区分开来。

## **4.4 实验总结**

在这个实验中，我学习了聚类算法中的KMeans算法，并将其应用于葡萄酒数据集的聚类问题中。通过实验，我更加深入地了解了KMeans算法的原理和实现方式，并且学会了如何使用Python编程实现这一算法。

在实验过程中，我注意到数据归一化是非常重要的，因为不同维度的特征取值范围不同，如果没有进行归一化，就会导致某些维度的特征对聚类结果的影响比其他维度更大。通过对比已经归一化的数据集和未归一化的数据集，我们发现未归一化的数据集会导致聚类结果不理想，这是因为未归一化的数据集中，某些特征的取值范围相对较大，导致这些特征对聚类结果的影响过大。

在实现KMeans算法的过程中，还需要注意一些细节问题，例如如何选择初始的质心，以及如何判断算法是否已经收敛。在实验过程中，我们通过不断调整参数和输出结果，来验证我们的算法是否正确，并且针对出现的问题采取相应的解决方案，最终得到了满意的聚类结果。

通过对比聚类结果和已知的三类葡萄酒数据之间的差异，我发现我的算法已经相当精准，准确率达到了94.9%。此外，我们还通过计算SSE值来评估聚类结果的好坏，SSE值越小表示聚类效果越好，最终我们得到的SSE总和为48.96。

最后，我还实现了可视化功能，通过将聚类结果以二维平面中点图的形式展示出来，更加直观地观察了三个质心和所有样本点之间的关系。通过这个实验，我们不仅学会了聚类算法的基本原理和实现方式，还掌握了如何使用Python实现KMeans算法，并且在实验中得到了满意的结果。