

**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](#_Toc58793874)** [1](#_Toc58793874)

**[4.2 实验内容](#_Toc58793875)** [1](#_Toc58793875)

**[4.3 实验过程](#_Toc58793876)** [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](#_Toc58793880)** [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,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**

提供动漫得分数据集（anime.csv）,包含用户对动漫评分(Score 2~Score 10)、动漫的欢迎程度(Popularity)等数据。

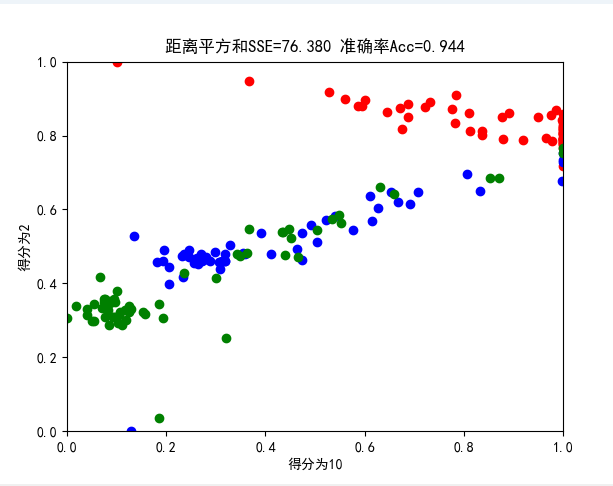
在对数据集进行处理时，按照Popularity列进行降序排序，在其中选择K类（eg. 选择Popularity高、中、低三类），每类选择一定数量的数据（eg. 每类选择60个数据），将选出的K类数据的K作为标签与Popularity和Score2~Score10组合成一个11维的数据，对除K以外的数据进行归一化处理。

编写kmeans算法，算法的输入是归一化后的数据集，动漫数据集一共11维数据，代表着动漫的11维特征，请在欧式距离下对动漫的所有数据进行聚类，聚类的数量为K。

以处理后的anime.csv作为输入文件。

在本次实验中，最终评价kmeans算法的精准度有两种，第一是处理后的动漫数据集已经给出的K个聚类，和自己运行的K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。

 进阶任务：在聚类之后，任选两个维度（为了效果良好建议选择Score 10和Score 2列数据进行展示），以K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所有的样本点。效果展示图可如图所示。



## **4.3 实验过程**

### 4.3.1 编程思路

### 4.3.2 遇到的问题及解决方式

### 4.3.3 实验测试与结果分析

## **4.4 实验总结**