華中科技大学

课 程 实 验 报 告

课程名称： 大数据分析

专业班级： **CS2202**

学　　号： **U202215399**

姓　　名： 吴渝东

指导教师： 崔金华

报告日期： **2024.06.20**

计算机科学与技术学院

目录

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。

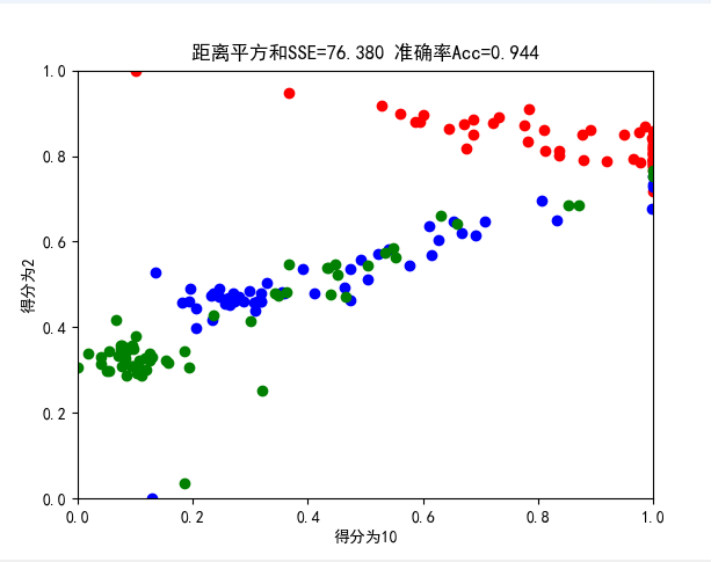
在对数据集进行处理时，按照 Popularity 列进行降序排序，在其中选择 K 类（eg. 选择 Popularity 高、中、低三类），每类选择一定数量的数据（eg. 每类选择 60 个数据），将选出的 K 类数据的 K 作为标签与 Popularity 和 Score2~Score10 组合成一个 11 维的数据，对除 K 以外的数据进行归一化处理。

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

首先读取文件中的数据，选取其中的 popularity,score2-10 作为特征，然后按照 popularity 逆序排序，从中低中高各选取 60 个作为数据集，并记录所属聚类的标签，采用减去平均值再除以方差的形式，对数据集进行标准化处理。

然后将处理好的数据作为 kmeans 聚类的输入，根据输入的数据，用选取 3 类数据的质心来初始化 3 个聚类，然后遍历每一条数据，计算其到 3 个聚类质心的欧氏距离，并将其添加到最近的聚类，在一次迭代结束后，再次用聚类的质心来初始化每个聚类进行下一次迭代，直至所有迭代结束。

将聚类后的数据与初始数据进行对比，计算准确率和 sse 值。

4.3.2 遇到的问题及解决方式

问题：对数据进行逆序排序时出错

解决方式：对数据集进行分析发现，数据集中存在许多非数值型的数据，因此出现报错，所以在数据处理阶段对包含非数值的数据进行处理，删除其对应的行。

4.3.3 实验测试与结果分析

实验运行结果如图 4-1，4-2 所示，观察可以得知，该算法准确率较高，但 sse 值同样较大，结合图 4-2 可以推测，sse 值较大的原因是聚类中的点较为分散，以及出现了部分离群点，可能是由于选取的 k 值对于数据集来说不够合适。

```
D:\Anaconda\python.exe D:\Desktop\大数据分析\实验\gitr
precise:
0.9777777777777777
SSE:
964.3413879154028

Process finished with exit code 0
```

图 4-1

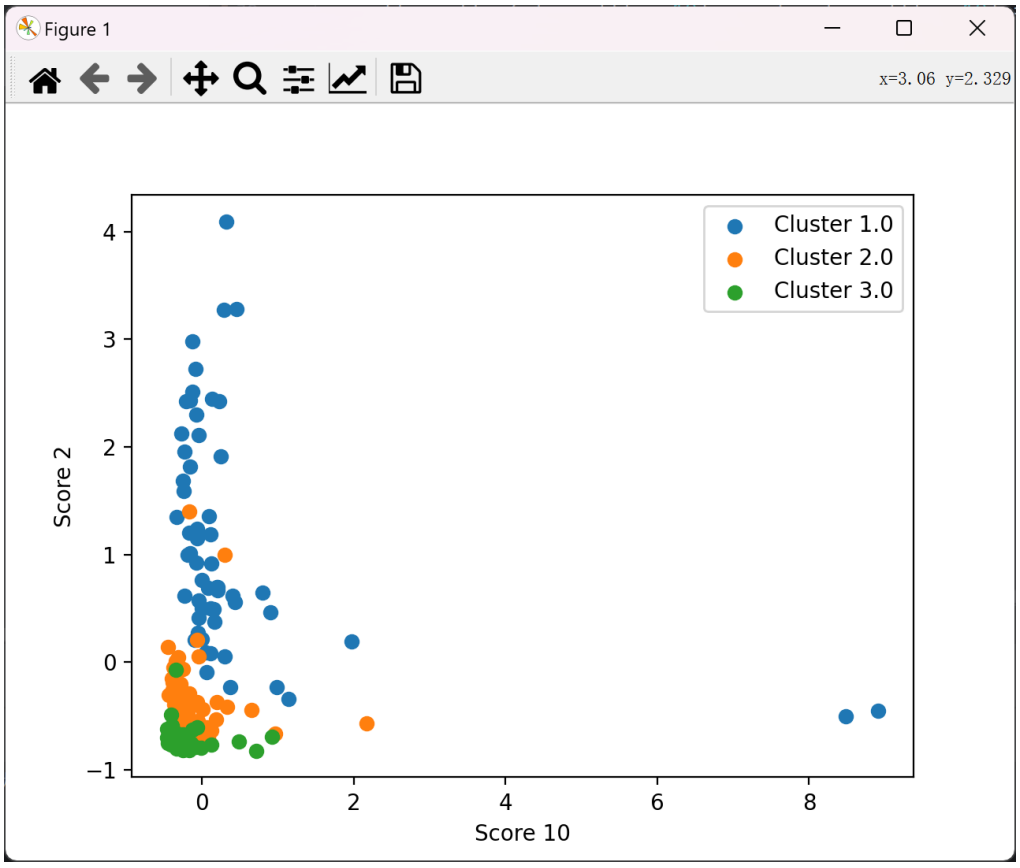


图 4-2

4.4 实验总结

在本次实验的过程中，实现了数据集的选取以及 kmeans 聚类算法，kmeans 算法采用近朱者赤的思想，将相邻的节点划分为一类，但是，通过实验发现，这个算法运行的过程中，可能会受到数据中噪声和离群点的影响，如何减小甚至消除此类影响值得我去思考和尝试。另外，此次实验采用质心的方式来表征每个聚类，是否可以以其他形式来表征每个聚类，以及其相应的聚类效果同样值得去思考和尝试。