大数据分析实验
——聚类算法

目录

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

一 任务背景

一 任务背景

- 加深对聚类算法的理解,进一步认识聚类算法的实现；
- 分析kmeans流程,探究聚类算法原里；
- 掌握kmeans算法核心要点；
- 将kmeans算法运用于实际，并掌握其度量好坏方式。

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

二 任务描述

二 任务描述

◆实验内容

1. 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。

2. 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。

3. 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。

二 任务描述

◆实验要求

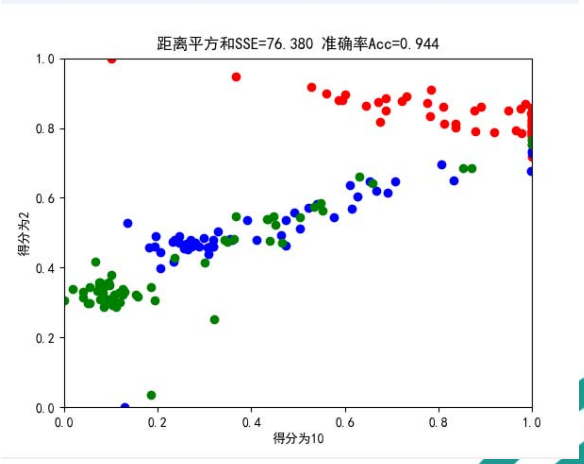
1.以处理后的anime.csv作为输入文件。

2.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。

二 任务描述

◆进阶任务

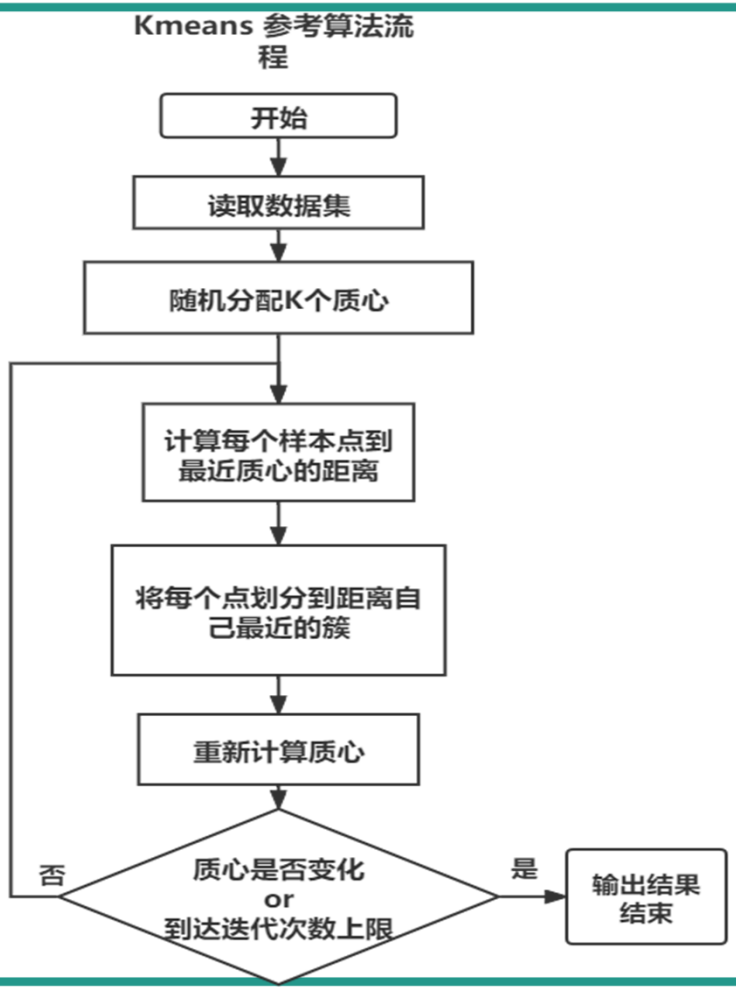
1.在聚类之后，任选两个维度（为了效果良好建议选择Score 10和Score 2列数据进行展示），以K种不同的颜色对自己聚类的结果进行标注，最终以二维平面中点图的形式来展示所有的样本点。效果展示图可如图所示。



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

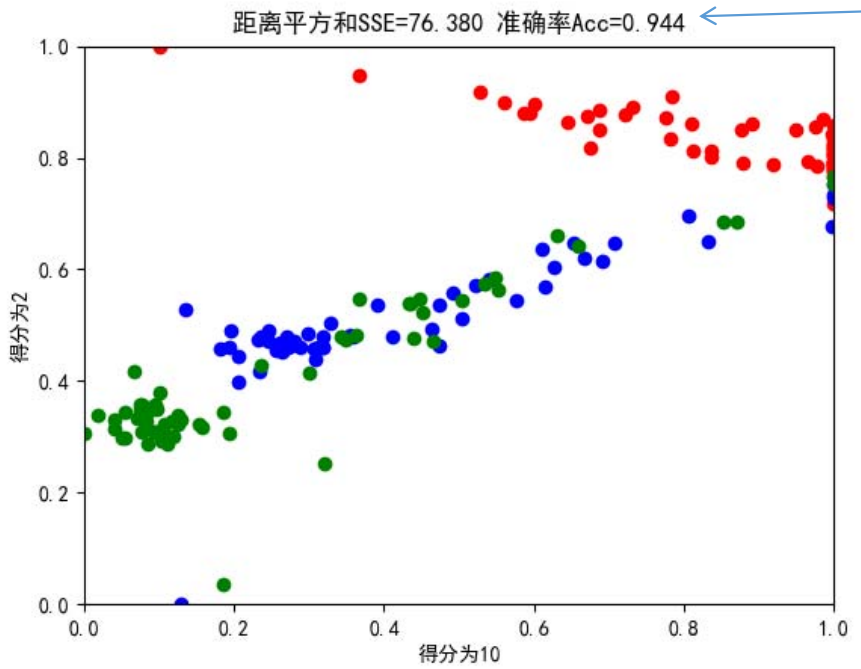
三 算法流程


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


距离平方和SSE=76.380 准确率Acc=0.944
1.0
0.8
0.6
0.4
0.2
0.0
得分为2
0.0
0.2
0.4
0.6
0.8
1.0
得分为10

必做

进阶

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。
- 不要直接调用现有的聚类算法的库。