**数据挖掘第二次作业**

15020510034-肖德祥

# 一、K-means算法概述

动态聚类是一种普遍采用的方法，具有以下 3 个要点：

（1）选定某种距离度量作为样本间的相似性度量。

（2）确定评价聚类结果质量的准则函数。

（3）给定某个初始分类，然后用迭代算法找出使准则函数取得极值的最好聚类结果。

首先对样本集进行初始聚类划分，一般的做法是先选择一些代表点作为聚类的核心，然后把其余的点按某种方法进行划分。代表点的选取，可以随机选取，也可以按照一定的方法选取。 选定一批代表点后，其余的点离那个点近，就被归为那一类。从而得到初始划分。

# 二、PAM算法概述

选用簇中位置最中心的对象，试图对n个对象给出k个划分；代表对象也被称为是中心点，其他对象则被称为非代表对象；最初随机选择k个对象作为中心点，该算法反复地用非代表对象来代替代表对象，试图找出更好的中心点，以改进聚类的质量；在每次迭代中，所有可能的对象对被分析，每个对中的一个对象是中心点，而另一个是非代表对象。对可能的各种组合，估算聚类结果的质量；一个对象o可以被使最大平方-误差值减少的对象代替；在一次迭代中产生的最佳对象集合成为下次迭代的中心点。

对比K-means：K-means是每次选簇的均值作为新的中心，迭代直到簇中对象分布不再变化。其缺点是对于离群点是敏感的，因为一个具有很大极端值的对象会扭曲数据分布。那么我们可以考虑新的簇中心不选择均值而是选择簇内的某个对象，只要使总的代价降低就可以。PAM算法比k-menas对于噪声和孤立点更鲁棒，因为它最小化相异点对的和而不是欧式距离的平方和。一个中心点可以这么定义：簇中某点的平均差异性在这一簇中所有点中最小。

# 三、k-means算法流程

问题：将N个样本{x1,…,xN}划分到k个类{C1,…,Ck}中，k为一正整数

目标：使得各个数据与其对应聚类中心点的误差平方和最小

其中：

Ji 为第 i 类聚类的目标函数

k为聚类个数

x是划分到类Ci的样本

m1,…,mk 是类C1,…,Ck的质心（均值向量）

算法流程：

1. 初始化：随机选择k个样本点，并将其视为各聚类的初始中心m1,…,mk ；

2. 按照最小距离法则逐个将样本x划分到以聚类中心m1,…,mk为代表的k个类C1,…,Ck中；

3. 计算聚类准则函数J，重新计算k个类的聚类中心m1,…,mk ；

4. 重复Step 2和 3直到聚类中心m1,…,mk无改变或目标函数J不减小。

# 四、PAM算法流程

输入：簇的数目k和包含n个对象的数据

输出：k个簇，使得所有对象与其距离最近中心点的相异度总和最小

1. 初始化：随机挑选n个点中的k个点作为中心点。

2. 将其余的点根据距离划分至这k个类别中。

3. 当损失值减少时：

1)对于每个中心点m，对于每个非中心点o：

I)交换m和o，重新计算损失（损失值的大小为：所有点到中心点的距离和）

ii)如果总的损失增加则不进行交换

# 五、实验要求

1.编程实现K-means算法针对UCI数据集的waveform数据集中的每类数据100个

2.编程实现K-means算法对一副无噪声图像进行分割

3.编程实现PAM对waveform数据集加20%的高斯噪声

4.编程实现PAM对一副噪声图像进行分割

# 六、实验结果

## 1.K-means数据聚类

每次随机选取300个数据，并对选出的300个数据进行聚类，设定终止条件为迭代的聚类中心不再改变。

选取6次实验结果展示：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 正确率 | 0.273333 | 0.506667 | 0.543333 | 0.323333 | 0.49 | 0.283333 |
| 迭代次数 | 12 | 6 | 6 | 12 | 18 | 9 |

这六次的平均正确率为40%。

## 2.K-means图像分割

对jpg格式图像处理，在python中先要把图像转化为三维矩阵，第三个维度对应于RGB值，为了简化计算，我将图像首先处理成灰度图像，再进行矩阵运算，这比直接对三维RGB矩阵操作时间要快许多。将图像转换成矩阵后，操作和1是一致的。

以下是一副小孩自信满满的图像：



以下是运行3次的效果图，我用灰度图像展示：

  

## 3.PAM数据聚类

PAM算法运行时间较长，我选取了前500个数据点作为样本点进行聚类，总共运行了六次，统计如下：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 正确率 | 0.468 | 0.276 | 0.276 | 0.45 | 0.45 | 0.488 |

这六次的平均正确率为40.1%。

## 4.PAM图像处理

以下是一幅动漫图像：



设定图像类别数为4，进行聚类划分。

展示三次运行结果：

  

PAM大致能把图像分好，然而程序运行的时间太长，求解欧氏距离的耗费时间比较多。