k-means算法开发文档

# 1 算法原理

## 1.1 算法概述

K均值聚类算法是先随机选取K个对象作为初始的聚类中心。然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个聚类。每分配一个样本，聚类的聚类中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。终止条件可以是没有（或最小数目）对象被重新分配给不同的聚类，没有（或最小数目）聚类中心再发生变化，误差平方和局部最小。

## 1.2 算法组成与步骤

算法步骤：

对于K-Means算法，首先要注意的是k值的选择，在确定了k的个数后，我们需要选择k个初始化的质心。由于我们是随机生成质心方法，k个初始化的质心的位置选择对最后的聚类结果和运行时间都有很大的影响，因此需要选择合适的k个质心，最好这些质心不能太近。

K-Means算法流程：

输入是样本集D={x1,x2,...xm}，聚类的簇数k,最大迭代次数N，新老质心变动距离的阈值Q

输出是簇的划分C={C1,C2,...Ck}

1、从数据集D中随机选择k个样本作为初始的k个质心向量： {μ1,μ2,...,μk}

2、对于n=1,2,...,N：将簇划分C初始化为Ct=∅t=1,2...k

3、对于i=1,2...m,计算样本xi和各个质心向量μj(j=1,2,...k)的距离：dij=||xi−μj||，将xi标记最小的为dij所对应的类别λi。此时更新Cλi=Cλi∪{xi}

4、对于j=1,2,...,k,对Cj中所有的样本点重新计算新的质心μj=1|Cj|∑x∈Cjx

5、如果所有的k个质心间的新老距离小于输入的阈值或者达到最大迭代次数，则转到步骤6。否则重复执行步骤2-4）

6、输出簇的划分结果C={C1,C2,...Ck}

## 1.3 算法核心参考文献

<https://www.cnblogs.com/pinard/p/6164214.html> k-means聚类算法原理

<https://blog.csdn.net/huangfei711/article/details/78480078> 5分钟带你弄懂 k-means 聚类

# 2 算法开发

## 2.1算法封装（输入与输出）

|  |  |  |
| --- | --- | --- |
|  | 名称 | 属性 |
| 输入：样本集 | points | double[][] |
| 输入：聚类的簇数k | k | int |
| 输入：最大迭代次数N | MAX\_ATTEMPTS | int |
| 输入：新老质心变动距离的阈值Q | MIN\_CRITERIA | double |
| 输出：簇的划分结果 | res | double[] |

## 2.2算法核心函数及说明

组成关系图参考



## 2.3 算法组成与执行流程

算法由Kmeans、KmeansDate、KmeansParam、KmeansResult、KmeansTest五个类组成，其中KmeansTest为测试类，KmeansDate为数据类、KmeansParam为参数结构类、KmeansResult为测试结果类、Kmeans为算法的主体。

# 3 算法验证

## 3.1 验证算例说明

https://blog.csdn.net/keseliugeizuori/article/details/53162171 k-means测试数据

## 3.2 验证结果说明

输入的测试数据为：

1.0000 1.0000

2.0000 1.0000

1.0000 2.0000

2.0000 2.0000

3.0000 3.0000

8.0000 8.0000

8.0000 9.0000

9.0000 8.0000

9.0000 9.0000

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

初始质心坐标为：

9.0000 8.0000

9.0000 9.0000

第1次迭代质心坐标为：

3.7143 3.5714

8.5000 9.0000

第2次迭代的质心坐标为：

1.8000 1.8000

8.5000 8.5000

第3次迭代的质心坐标为：

1.8000 1.8000

8.5000 8.5000

迭代完毕

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

聚类个数=2

迭代次数=3

退出迭代时的新旧质点间最大距离（小于阈值）=0.0

每个聚类包含的点的个数:

5 4

原始数据对应的分类为:

属于0簇的数据：

[1.0, 1.0]

[2.0, 1.0]

[1.0, 2.0]

[2.0, 2.0]

[3.0, 3.0]

第0簇数据归纳完毕

属于1簇的数据：

[8.0, 8.0]

[8.0, 9.0]

[9.0, 8.0]

[9.0, 9.0]

第1簇数据归纳完毕

计算完毕

流程图参考

