# Report 1--进度报告

### K-means介绍

1967年，James MacQueen提出“K-Means”（K均值），这个被应用得最广泛的基于划分的聚类算法，其实是一种硬聚类算法，属于典型的局域原型的目标函数聚类的代表。算法首先随机选择k个对象，每个对象初始地代表一个簇的平均值或者中心。对于剩余的每个对象，根据其到各个簇中心的距离，把它们分给距离最小的簇中心，然后重新计算每个簇的平均值。重复这个过程，直到聚类准则函数收敛。

该算法是很典型的基于距离的聚类算法，采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似度就越大。该算法认为簇是由距离靠近的对象组成的，因此把得到紧凑且独立的簇作为最终目标。

k个初始类聚类中心点的选取对聚类结果具有较大的影响，因为在该算法第一步中是随机的选取任意k个对象作为初始聚类的中心，初始地代表一个簇。该算法在每次迭代中对数据集中剩余的每个对象，根据其与各个簇中心的距离将每个对象重新赋给最近的簇。当考察完所有数据对象后，一次迭代运算完成，新的聚类中心被计算出来。如果在一次迭代前后，J的值没有发生变化，说明算法已经收敛。

### 算法流程

1）从N个文档随机选取K个文档作为[质心](http://baike.baidu.com/view/139124.htm" \t "http://baike.baidu.com/_blank)

2）对剩余的每个文档测量其到每个[质心](http://baike.baidu.com/view/139124.htm" \t "http://baike.baidu.com/_blank)的距离，并把它归到最近的质心的类

3）重新计算已经得到的各个类的[质心](http://baike.baidu.com/view/139124.htm" \t "http://baike.baidu.com/_blank)

4）迭代2～3步直至新的质心与原质心相等或小于指定阈值，算法结束

### 算法伪代码

k-means算法：

for k = 1, … , K 令 r(k) 为从D中随机选取的一个点；

while 在聚类Ck中有变化发生 do 形成聚类：

For k = 1, … , K do

Ck = { x ∈ D | d(rk,x) <= d(rj,x) 对所有j=1, … , K, j != k}；

End;

计算新聚类中心：

For k = 1, … , K do

Rk = Ck 内点的均值向量

End;

End;

### 目前存在的问题

根据伪代码写的算法主要在训练集上调整数据维度比较困难，高维度下不能难模拟出现实的聚类图像，手动过程选取k值，对不同的k值选取结果进行比较，根据比较挑选出比较好的训练结果。