# Report 2--进度报告

### K-means算法的缺陷

1. 聚类中心的个数K 需要事先给定，但在实际中这个 K 值的选定是非常难以估计的，很多时候，事先并不知道给定的数据集应该分成多少个类别才最合适
2. Kmeans需要人为地确定初始聚类中心，不同的初始聚类中心可能导致完全不同的聚类结果。（可以使用Kmeans++算法来解决）

针对上述第2个缺陷，可以使用Kmeans++算法来解决

### K-means++介绍

#### k-means++算法选择初始seeds的基本思想就是：初始的聚类中心之间的相互距离要尽可能的远。

1. 从输入的数据点集合中随机选择一个点作为第一个聚类中心
2. 对于数据集中的每一个点x，计算它与最近聚类中心(指已选择的聚类中心)的距离D(x)
3. 选择一个新的数据点作为新的聚类中心，选择的原则是：D(x)较大的点，被选取作为聚类中心的概率较大
4. 重复2和3直到k个聚类中心被选出来
5. 利用这k个初始的聚类中心来运行标准的k-means算法

 从上面的算法描述上可以看到，算法的关键是第3步，如何将D(x)反映到点被选择的概率上，一种算法如下：

1. 先从我们的数据库随机挑个随机点当“种子点”
2. 对于每个点，我们都计算其和最近的一个“种子点”的距离D(x)并保存在一个数组里，然后把这些距离加起来得到Sum(D(x))。
3. 然后，再取一个随机值，用权重的方式来取计算下一个“种子点”。这个算法的实现是，先取一个能落在Sum(D(x))中的随机值Random，然后用Random -= D(x)，直到其<=0，此时的点就是下一个“种子点”。
4. 重复2和3直到k个聚类中心被选出来
5. 利用这k个初始的聚类中心来运行标准的k-means算法

### 目前存在的问题

K-means算法和K-means算法进行对比，主要在时间运行对比这一块有问题，时间的记录有问题，重新用时间戳记录下。还有k值选取时的问题，多次进行测试。