1. 概念

k均值聚类算法（k-means clustering algorithm）是一种典型的基于距离的聚类算法。所谓聚类，可以理解为把具有某种相同或者相似特征的样本归为同一类。k均值聚类算法可以对给定的一个样本集，按照样本之间的距离大小，将样本集划分为K个簇，让簇内的点尽量紧密的连在一起，而让簇间的距离尽量的大。另外，k-means算法属于无监督学习的一种聚类算法，它是在不知数据所属类别及类别数量的前提下，依据数据自身所暗含的特点对数据进行聚类。

2. 原理

k-means算法是基于迭代求解的思想，下面将以一个点集为例进行说明。如图2.1所示是一个样本个数为N的点集P，现在我们需要把它分为4个簇。原始的k-means算法首先是在点集P中随机选取4个点作为初始的聚类中心，每个聚类中心表示一个簇，用表示。然后计算点集P中每个点到各个聚类中心的距离，把每个点归到离它自己最近的那个聚类中心所在的簇。当所有点都归类完毕后，重新计算并更新簇的聚类中心。如果前后两次的聚类中心没有发生变化，说明算法已经收敛，得到四个分好的簇。否则根据新的聚类中心，对点集中的每个点进行重新归类，也即重新计算点集P中每个点到各个新聚类中心的距离，把每个点归到离它自己最近的那个新聚类中心所在的簇。对上述过程进行迭代，直至算法收敛。k-means算法分类后的结果如图2.2所示，不同的颜色区域表示一个簇，每个簇中最大那个黑点就是聚类中心。

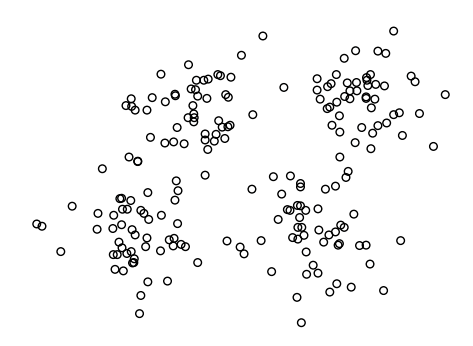


图2.1 初始需要分类的数据集

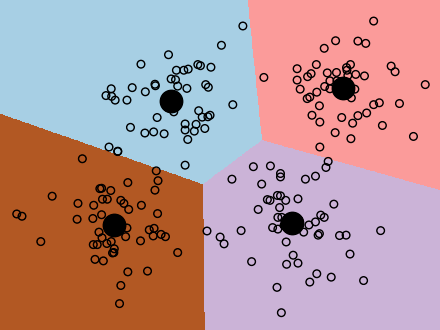


图2.2 根据k-means算法分类后的4个簇

# 3. 算法流程图

k-means算法流程图如图3.1所示：



图3 k-means算法流程图

# 4. C++实现的仿真结果

此次仿真数据集为随机产生的200个点，点的横纵坐标范围均为[0,30]内的整数，聚类个数为4。实验代码和需要用到的数据集文件已上传到群文档，文件名为“k-means.cpp”的文件为源代码，文件名为“k-means\_input.txt”的文件为待分类的数据集，文件名为“k-means\_output.txt”的文件为分类后的结果。图4.1为未分簇时数据集的分布，图4.2为运行时随机选取的初始聚类中心以及最终分类后的聚类中心，图4.3为分簇后各个簇的分布，用不同符号表示不同区域内的点，O为每个区域内的聚类中心。

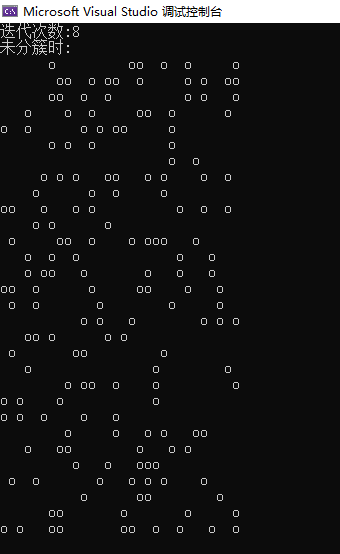


图4.1 未分簇时数据集的分布

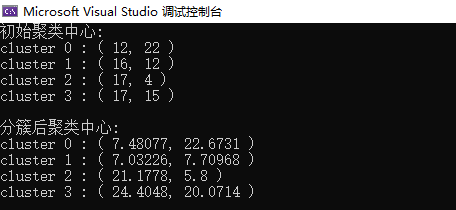


图4.2 随机选取的初始聚类中心以及最终分类后的聚类中心



图4.3 分簇后各个区域点集的分布

# 4. 参考资料

https://blog.csdn.net/qq\_41251963/article/details/81603169?biz\_id=102&utm\_term=k-means%E8%81%9A%E7%B1%BB%E7%AE%97%E6%B3%95&utm\_medium=distribute.pc\_search\_result.none-task-blog-2~all~sobaiduweb~default-1-81603169&spm=1018.2118.3001.4187

https://blog.csdn.net/qq\_39742013/article/details/81675050?ops\_request\_misc=%257B%2522request%255Fid%2522%253A%2522159619781819195265936298%2522%252C%2522scm%2522%253A%252220140713.130102334..%2522%257D&request\_id=159619781819195265936298&biz\_id=0&utm\_medium=distribute.pc\_search\_result.none-task-blog-2~all~top\_click~default-1-81675050.first\_rank\_ecpm\_v3\_pc\_rank\_v2&utm\_term=k-means%E8%81%9A%E7%B1%BB%E7%AE%97%E6%B3%95&spm=1018.2118.3001.4187

<https://baike.baidu.com/item/K%E5%9D%87%E5%80%BC%E8%81%9A%E7%B1%BB%E7%AE%97%E6%B3%95/15779627?fr=aladdin>

https://blog.csdn.net/k76853/article/details/50440344?ops\_request\_misc=%257B%2522request%255Fid%2522%253A%2522159624894319195265935920%2522%252C%2522scm%2522%253A%252220140713.130102334..%2522%257D&request\_id=159624894319195265935920&biz\_id=0&utm\_medium=distribute.pc\_search\_result.none-task-blog-2~all~top\_click~default-1-50440344.first\_rank\_ecpm\_v3\_pc\_rank\_v2&utm\_term=k-means%E8%81%9A%E7%B1%BB%E7%AE%97%E6%B3%95C%2B%2B&spm=1018.2118.3001.4187