**读书报告**

61518424 王贵涛

1. **问题与解答**

我提出问题：

1. 聚合聚类将所有样本聚成一类不就没有意义了吗？

讨论结果：给样本做了一个有一层层子目录的分类。

1. 样本间距为什么选取欧式距离平方？如果选用欧式距离会有什么变化？

讨论结果：因为单单选择欧式距离还需要开根号，而k均值聚类时比较的是距离大小，同时平方并不会影响结果。

别人提出的问题：

1. 为什么说k均值聚类中当k较大时平均直径将收敛？另外为什么拐点就是最优的k？

我的解答：当类已经饱和的时候如果再增加类会强行将一个类拆分成两个类，一个优秀的类别的形状应该是近圆形或近球体，两个半圆的外包圆直径仍为原来的圆的直径，因此平均直径不再降低。到达拐点的时候平均直径的下降速度减慢，我们不仅需要优化平均直径，还要保证类的数量不能过多，否则分类将没有意义。

1. 聚合聚类算法中，新类到原类的距离的定义是否欠妥？为何不考虑为类中各点中心到其他各类中的距离？

我的解答：各个类之间是区分开来，没有关联的，类间距的影响不大，无论两个类相聚远近，我们都认为它们是两个类，平均直径已经足够。当然也可以考虑将类间距最大化，如果有一个类距离其它类很远，我们会倾向于将其它类进行合并，这其实并不常见，而且会增加计算复杂度。

1. 既然k均值聚类算法需要选择初值，初值的选择是否会对收敛性有影响？如果有，那么k均值聚类算法对于初值的选择很依赖吗？

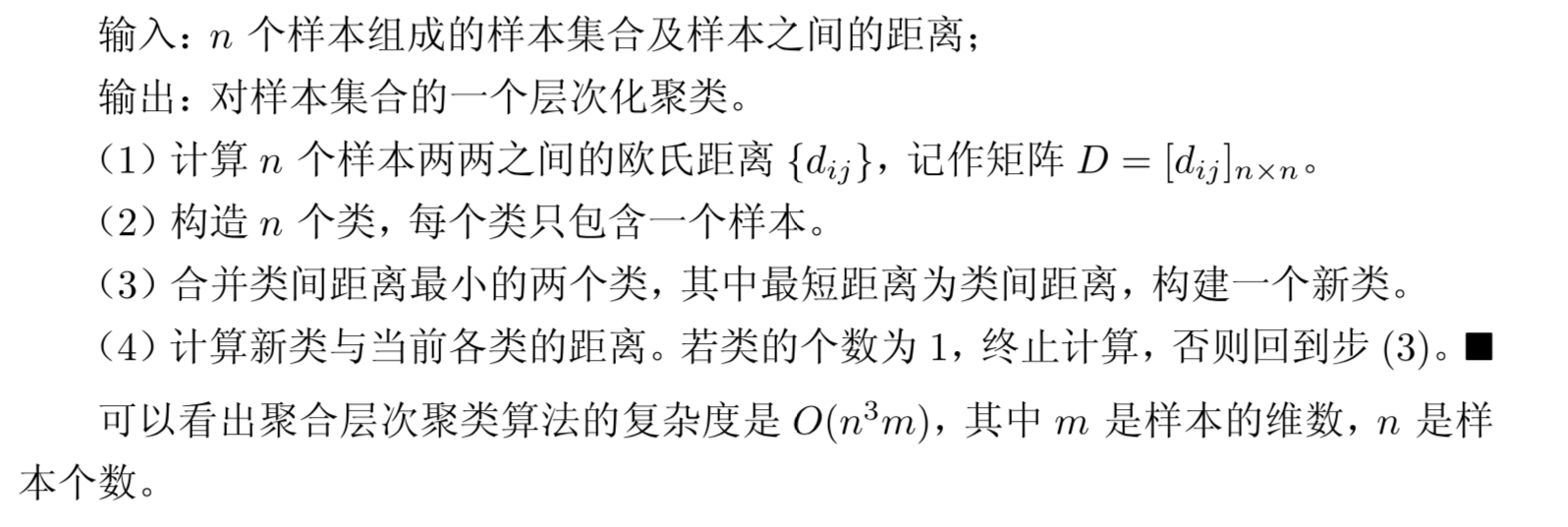
我的解答：会有影响，类中心的移动很小，最终类中心的位置很大程度上取决于初值的选择。我们可以通过随机选取不同初值的方法来找到最优的类中心。

1. **下周计划安排**

看完第十五章并参加讨论。

1. **读书收获**

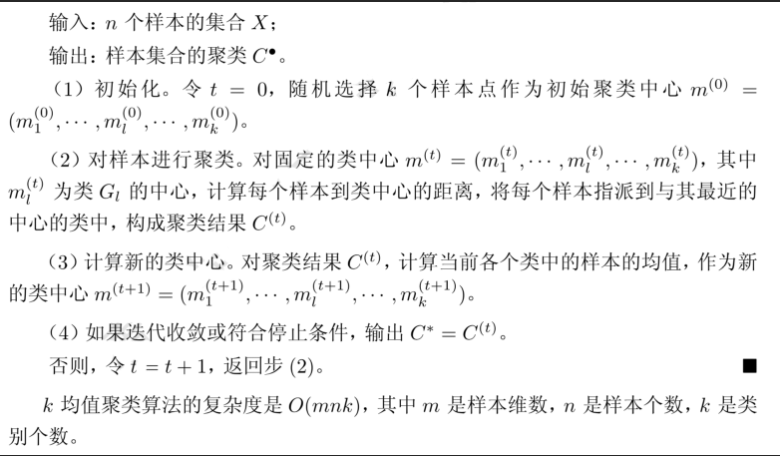
聚合聚类算法：



k均值聚类：

* k均值聚类是基于样本集合划分的聚类算法。
* k均值聚类将样本集合划分为k个子集，构成k个类，将n个样本分到k个类中，每个样本到其所属类的中心的距离最小。
* 每个样本只能属于一个类，所以k均值聚类是硬聚类。

k均值聚类算法：



算法特性：

1. 收敛性

* k均值聚类属于启发式方法，不能保证收敛到全局最优，初始中心的选择会直接影响聚类结果。
* 注意，类中心在聚类的过程中会发生移动，但是往往不会移动太大，因为在每一步，样本被分到与其最近的中心的类中。

1. 初始类的选择

* 选择不同的初始中心，会得到不同的聚类结果。
* 初始中心的选择，比如可以用层次聚类对样本进行聚类，得到k个类时停止。然后从每个类中选取一个与中心距离最近的点。

1. 类别数k的选择

* k均值聚类中的类别数k值需要预先指定，而在实际应用中最优的k值是不知道的。
* 尝试用不同的k值聚类，检验得到聚类结果的质量，推测最优的k值。
* 聚类结果的质量可以用类的平均直径来衡量。
* 一般地，类别数变小时，平均直径会增加
* 类别数变大超过某个值以后，平均直径会不变，而这个值正是最优的k值。实验时，可以采用二分查找，快速找到最优的k值。

