**2019/02/23读书报告**

学号 71117228 姓名 李泓烨

读书进度：5.1

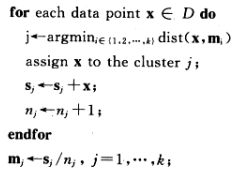
1. **读书报告内容**

**1. 自己提出问题的理解**

① 为什么解决空聚类的时候，选择离一个含有大量数据的聚类的聚类中心最远的数据点？

讨论后的理解：选的是centroid，先选的这个是最有可能的，后面centroid的位置会变动，这样使得可以更快的迭代出结果

② 在4.2.2节中提到，公式(2)的求和运算是分开计算的，那么这个分开计算是如何实现的呢？

讨论后的理解：一个一个点计算，而不是整体计算，具体过程再理解一下图吧

**2. 别人提出问题的理解**

1. **提出的问题1**：3.2.4节中，如果一个聚类任务的数据中，其属性既有离散值又有连续值怎么办？是否可以将连续属性按照数据集中的数据离散化分为几个区间，将区间视为离散的继续做？

**讨论后的理解**：因为聚类问题中聚类的标准是距离的大小，而其实每一个数据的属性数值是确定的，所以距离可以计算；如果数据时范畴数据的话，需要使用k-modes算法，利用模来替代均值作为聚类中心

**引申**：categorical value是绝对性的无先后顺序的值

2. **提出的问题2**：To be safe, we may want to monitor these possible outliers over a few iterations and then decide whether to remove them. It is possible that a very small cluster of data points may be outliers. Usually, a threshold value is used to make the decision.其中，threshold value指的是什么？能具体讲一下怎么监视与这个方法的整体吗？

**自己的理解**：设置一个**阈值**，筛掉一些点。在多次循环中监视，之后再迭代中逐渐筛掉

**引申**：距离指的是更倾向于一点到所有质心的和的阈值

3. **提出的问题3**：使用k-均值算法的时候，为什么全局最小值对于大规模数据集来说在计算上是不可行的？

**自己的理解**：局部最小值指的是在选取这种初始点的情况下能达到的最佳分类，由于初始点是随机的，也很难保证达到全局最小

4. **提出的问题4**：除了人工选取外，有无合理的方法去选择恰当的初始点？

**自己的理解**：暂时没有解答

5. **提出的问题5**：4.3.1节，the centroid representation alone works well if the clusters are of the hyper-spherical shape. If clusters are elongated or are of other shapes, centroids may not be suitable.是因为可以计算出the radius，standard deviation of the cluster吗？我不太清楚

**自己的理解**： 个人认为，超球体是多维的球体，比如一位的超球体是圆形，二维的超球体是球

6. **提出的问题6**：One can use the set of rules to evalute the clusters to see whether they conform to some existing domain knowledge or intuition.怎么理解，怎么操作呢？

**自己的理解**： 聚类用规则分类后，看看是否存在已有的知识模型，可以应用到上面

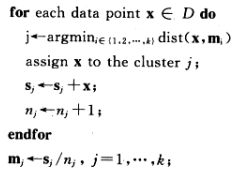
**群里的同学举的例子**：比如服装厂对人身高进行聚类 某一聚类结果把男生身高160和190聚类在了一起，因此得出结论160和190的男生需要穿同一个尺寸的衣服 但是根据领域知识来讲 这显然是错误的

7. **提出的问题7**：4.3.1中提到的one can use the set of rules to evaluate the cluster to see whether they conform to some existing domain knowledge or intution可以利用规则集合来评估聚类是否符合某些已经存在的领域知识或常识

**自己的理解**：和问题6相同，不再赘述

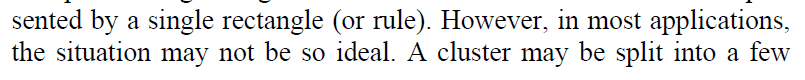
8. **提出的问题8**：在k-均值算法的硬盘版本中，是如何实现了每次for循环中，算法只是简单地扫描全部数据一次，普通版本的k-均值算法呢，扫描了多少次，是如何扫描的？

**自己的理解**：普通方法——扫描两次，一遍聚类，一遍算均值；硬盘版本，扫描一次

9. **提出的问题9**：在4.2.2节中提到，公式(2)的求和运算是分开计算的，那么这个分开计算是如何实现的呢？

**自己的理解**：一个一个点计算，而不是整体计算

10. **提出的问题10**：书上第129页：



这里究竟是什么特殊情况？

**自己的理解**：有一些情况，个人认为需要很多规则去约束，这样就造成了要用好几个矩形来分割

11. **提出的问题11**：对于4.2.3节中提出的计算good initial seed的方法，是否存在除了outliers以外的其它弊端，比如之后选取的seed和之前选取的seed重合或者距离很近？

**自己的理解**：书上给出了改进的办法，先随机选出一部分，然后用均值去最远的x1，x1最远的x2，每次选择时候都会选离之前的尽量远的，应该不存在seed重合或者距离很近的情况，除非k的取值较大或者数据集中的点本身距离很近

12. **提出的问题12**：



为何距离函数需要有平方？是为了保证SSE可以用增量来计算吗？

**自己的理解**：距离函数有平方，其实可以保证在数据量大的情况下，运算起来方便一些；另外，开根号运算的话会造成比较大的误差，不建议使用

**3. 读书计划**

① 本周完成的内容章节：看完了第4章

② 下周计划：看完第5章

**4. 读书摘要、理解&伪代码的具体实现**

1. 读书摘要

这里我以思维导图的形式记录了我的读书笔记

