## 2020.2.23 第四次读书报告

09018330 孙毅远

### 一、自己提出的问题的理解

#### 1.除了人工选取外，有无合理的方法去选择恰当的初始点？

理解：可以通过随机采样先选取一部分点再从中取得初始点

### 二、别人提出的问题的理解

#### 1. 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？

理解：不论是离散值还是连续值，只要是数据值能表示的属性都可以

#### 2. 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？

理解：threshold value指的是一个阈值，通过设置一个值来判断是否存在异常值，去除那些比其他任何数据点离聚类中心都要远的数据点，需要在多次循环中监控再决定是否删除，阈值是一个判断标准

#### 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？

理解：由于初始点的随机选取，算法能做到的最优值有限，只是在这种初始点组合情况下的局部最优值，而总体分类的全局最优值无法保证取到，因为大数据无法遍历所有的初始点组合，只能取随机组合中求得的局部最小值的最小值作为结果。

#### 4. 在4.3.1节的最后，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？

理解：用rule给出的一些判别标准是否适用于已有的经验规律，简便地判断当前分类是否合理

#### 5.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？

理解：改进方法是先随机取样，选出一部分，然后用均值取最远的x1,离x1最远的x2这样依次选取中心点，尽量避免重合或者相近的情况，看最后取得的结果如何，再适当调整k值

### 三、读书计划

#### 本周 结束Web data mining的学习

#### 下周 开始统计学习方法的学习

### 四、读书收获

#### 4 Unsupervised Learning

##### 4.1 Basic Concepts

* 聚类（ ）
* 聚类是一个将数据集中在某些方面相似的数据成员进行分类组织的过程
* 划分聚类（ ） 层次聚类（ ）

##### 4.2 K-means Clustering

* （）
* 设数据点（）集合 为 ，其中 是实数空间的向量，且 表示数据的属性数目， 把给定的数据划分为 个聚类，每个都有一个聚类中心（），即是这个聚类中所有数据点的均值
  + 算法过程如下：
  + 先随机选取 个数据点作为初始聚类中心，然后计算每个数据点与各个种子聚类中心的距离，把每个数据点分配给距离其最近的聚类中心，一旦所有数据点都被分配，每个聚类的中心会重新计算，循环这个过程直到误差平方和（）最小。
  + 聚类的均值可以通过这样计算：
  + 存在的问题：可能出现空聚类
* + 优势：简洁、效率，复杂度为
  + 劣势：
    - 算法只能用于均值可被定义的数据集上
    - 需要事先指定聚类 ，在实践中往往需要测试多个 值产生最优结果
    - 对异常值处理容易出现错误
    - （解决方案：设定阈值提前处理异常值，随机采样预先聚类后进行半监督学习）
    - 对初始的聚类中心要求较高，不同的中心会出现不同结果

##### 4.3 Representation of Clusters

* 用聚类中心表示
* 利用分类模型表示（如规则）

#### 5 Partially Supervised Learning

##### 5.1 Learning from Labeled and Unlabeled Examples

* 由于数据标注的费时，所以研究从少量的已标注数据和大量无标注数据学习的问题（）
* 算法是一种存在不完整数据情况下使用的最大似然估计的迭代算法，步骤分为期望过程（）和最大化过程（）
  + 推导过程中的两个假设
  + $1.数据是由某个混合模型生成的\\2.在混合成员和分类类别之间有着一一对应的关系$
  + 第二个假设不成立的情况下的解决方法
    - 给无标注数据加权（）
    - 在参数估计的过程中给无标注数据赋予较低权重
    - 寻找混合分量
* 这种方法假设数据集中的属性可以被划分为两个集合，其中每一个都足够学习得到目标分类函数。
* 主要思想：每次循环中从已标注数据集 的每个特征集中学习得到一个分类器，然后用这个分类器去标注 中的无标注数据，将标注后分类可信度高的数据加入 中。
* 自学习
* 利用分类器的预测结果来训练自己的方法
* 直推式支持向量机（）
* 基于图的方法
* 基本思想：把已标注数据和无标注数据当作一张图中的顶点，并且使用某种相似度函数来定义边的权重，越相似的权重越大，基于这种想法进行标注
  + 最小割（）
  + 高斯场（）
  + 谱图直推器（）