web Data Mining 第四次读书报告

2020/2/17-2020/2/23

汇报人：71117415-喻泽弘

读书进度：chapter 4.1-chapter 4.3

问题与解答

- 个人提出的问题：
  1. 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？

     对于4.2.3节提出的计算good initial seed的方法，包括一下步骤：

     - 对所有的data point 计算均值m
     - 遍历所有的data point，选取距离m最远的data point作为第一个seed---x_1
     - 遍历所有的data point，选取同前i个seed距离和最远的节点作为第i+1个seed，即x_{i+1}

     根据讨论，选取的seed和之前选取的seed重合或者距离很近主要有两个原因：

     （1） 数据点本身分布出现集群的现象，或者某一个聚类的数据点十分多，那么，选取seed的时候就有可能选取到相同的data point

     （2） k-means方法中k值设置的过大，即，数据集中并没有这么多的聚类，那么，就会出现聚类很近，甚至重合的现象

  2. 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 intuition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(这个问题被别人提出，但是我思考过)

     对于这个问题，我认为，书中想表达的意思应该是：聚类出的结果可以通过领域内的一些常识或者知识判断出聚类结果是否有意义。比如：某服装厂对男性的身高进行聚类，在一个聚类结果中，把男性身高160和190聚类在了一起。从而得出，169和190的男生需要穿同一尺寸的衣服，但是我们根据常识，显然可以发现这么结论是错误，不符合生活常识。

- 别人提出的问题：
  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？

     首先，如果属性中出现了连续值，那么该属性的表现就是数值类型的值，而不是categorical data，从而对于连续值的区间，我们可以选取区间的中点或者其它值，讲区间离散化，也就是将区间视为离散值继续做，从而继续运用k-means方法取聚类。

  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？

     对于这个问题，我和群里面的意见有点分歧。从字面上进行理解，outliers，即离群值，这说明这个值首先是十分少的，并且距离其它的聚类中心十分的远。读书讨论群中认为threshold value是该聚类距离其它聚类中心的距离，但是我认为不能仅仅考虑距离，还需要考虑数量，如果一个聚类虽然离其它的距离中心很远，但是这个聚类内部的点非常多的话，那么也不能将这个聚类视为outliers进行处理。当然，threshold value具体是什么内容，书中并没有详细说明，上述的内容是我的个人观点。

3. 使用k-means算法的时候，为什么全局最小值对于大规模数据来说是不可行的?

   类比于数学中的函数，k-means算法实际运用的过程中，会出现局部最优与全局最优的问题，对于聚类这个问题，我认为这是个NP问题，k-means是一种启发式的算法，先随机取点，再生成聚类，从而得到最优值，但这些最优值往往都是局部最优，如果需要寻找到全局最优，是需要对所有的数据点进行遍历，对于大规模数据来说，这需要耗费大量的时间以及计算资源，显然是可行的。此外，全局最小值也不一定会明显由于目前发现的局部最优值，因此，性价比并不是非常高

4. 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?

   空聚类问题，是当k值设置较大时，此时无法寻找出一个新的data point作为聚类中心，或者两个聚类重合在了一起。对于这个问题，选择离一个含有大量数据的聚类的聚类最新最远的数据点的目的在于，该点是最有可能的潜在聚类中心。

下周读书计划

完成web data mining第五章的阅读，并且尽快投入统计学习这本书的阅读。

读书收获

- k-means方法：这是一种聚类的方法。该方法主要分为三步：
  - 方法开始时，随机选取数据集中的一些点作为聚类的中心
  - 将数据集中的其它点按照同聚类中心的距离分配距离最小的聚类中心
  - 重新计算各个聚类的中心，并且继续执行第二步直至到达了阈值
- 了解到了无监督学习与监督学习之间的差别：
  - 监督学习用于发现具有相同类别的数据点的属性之间的一些模式，并且利用这些模式，根据数据的属性去预测该数据的类别
  - 无监督学习：在一些领域中，数据往往没有类别。无监督学习就是通过探索数据，发现数据中的一些潜在特征，将相似的数据聚集在一块。同时，无监督学习不需要人工添加标签，从而节省了人工的成本
- 认为到了k-mean方法的优点以及缺点：

  优点：

  - 简单、高效
  - k-means方法可以视为一种线性的算法

  缺点：

  - 对于categorical data这种数据，我们无法计算均值，此时k-means方法便不再使用，从而，提出了一种用于处理categorical data的一种方法，即，k-modes。该方法的距离计算并不是通过计算距离，而是计算数据点同mode相匹配的value的个数。同时k-modes方法不再使用means，而是modes，modes为聚类中同一属性中出现的最频繁的值，将该值作为聚类中心，我认为这样处理是比较合理，它反映了聚类中最普遍的情况
  - 用户需要提前设置好聚类的数量k,对于同一数据集，不同的k值,结果的差异可能很大
  - 算法对outliers这种现象十分的敏感。outliers可能是数据集中的错误数据或者是一些拥有特殊值的数据，由于k-means方法使用均值作为每个聚类的聚类中心，因此outliers便有可能导致不理想的聚类出现