第三次读书报告

侯宇捷

1. 自己的问题
2. 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？

解答：可以的，但由于部分离散值的无序与彼此之间不存在大小关系，不能将离散值化为类似于连续值的东西在空间中计算之间的距离。

1. 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？

解答：对于threshold value有一定的争议。有人认为是指cluster中data point的数量，有人认为是outlier到距离最近的centroid的距离，还有人认为是到所有centroid距离的和与被评估的outlier的数量。我个人更倾向于最后一种。正如某位读书会成员所提出的，outlier一般具有数量少、离其他data point远的特点。

1. 别人提出的问题
2. 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？

解答：因为算法对于初始的centroid选择十分敏感，而不同的初始选择可能导致最终centroid不同，而导致不同的聚类。因此，对于大规模数据集，通常来说我们是做不到尝试所有的初始可能，而因此就算我们多次尝试，也不过得到一个局部的最小值，达不到全局的范畴。

1. 在4.3.1节的最后，One can use the set of rules to evaluate the clusters to see whether they conform to some existing domain knowledge or intuition.怎么理解，怎么操作呢？

解答：关于这一点，有人提出是根据已有的一定数学知识来看我们得到的聚类成果是否符合其数学模型，也有人提出是依靠于我们的尝试来对于聚类成果判断好坏。我个人是偏向于前一种的，但我认为后一种也未尝没有意义，两者可以结合起来看。

1. 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？

解答：这很大程度上减少了其中存在的误差。面对大数据集时，采取根号相加的方法很容易导致巨大的误差，而平方则没有这样的缺点。

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中心最远的数据点？

解答：面对这个问题，有人提出这样是为了减少有本应被聚类的数据因centroid的集中偏向而导致没有被聚类。此外，我也提出了如果存在大量数据，是否是有可能因为该处centroid分布不好又不密集而导致一些较远的、又不合适的data被分入，在这样的情况下，选择较近的是否也有道理的问题。面对这一问题，也有人提出这种情况经历几次算法的迭代就会逐渐调整centroid的位置而解决，我觉得是有道理的。

1. 读书进度

本周完成内容：4.1-4.3，5.11-5.14

下周读书计划：完成最后的5.15，并阅读有关支持向量机、朴素贝叶斯的内容加深对第五章的理解

四．读书摘要

第四章：

1.unsupervised learning

2.k-means clustering algorithm

Stopping criterion:

(1). no (or minimum) re-assignments of data points to different clusters.

(2). no (or minimum) change of centroids.

(3). minimum decrease in the sum of squared error (SSE)

When meet empty clusters: we can choose a data point as the replacement centroid, e.g., a data point that is furthest from the centroid of a large cluster. If the sum of the squared error (SSE) is used as the stopping criterion, the cluster with the largest squared error may be used to find another centroid.

3.disk version of k-means clustering algorithm

Strength: it does not need to load the entire data set into the main memory, which is useful for large data sets.

Note: we could set a limit on the number of iterations because later iterations typically result in only minor changes to the clusters.

4. Its time complexity is O(tkn), where n is the number of data points, k is the number of clusters, and t is the number of iterations. Since both k and t are normally much smaller than n. The k-means algorithm is considered a linear algorithm in the number of data points.

5.

Weakness1：mean cannot be calculated

Solution1: a variation: k-mode

Weakness2: k should be decided in advance

Solution2: try different k and evaluate the cluster later

Weakness3: existing outliers

Solution3: (1) monitor these possible outliers(that are much further away from the centroids than other data points) over a few iterations and then decide whether to remove them. Usually, a threshold value is used to make the decision.

（2）perform random sampling and use the sample to do a pre-clustering and then assign the rest of the data points to these clusters. Then, you can simply assign each remaining data point to the centroid closest to it or preform supervised learning or semi-supervised learning.

Weakness4: Different initial seeds may result in different clusters. Thus, the algorithm only achieves local optimal. The global optimal is computationally infeasible for large data sets.

Solution4: select good initial seeds.

1. first compute the mean m (the centroid) of the entire data set (any random data point rather than the mean can be used as well). Then the first seed data point x1 is selected to be the furthest from the mean m. The second data point x2 is selected to be the furthest from x1. Each subsequent data point xi is selected such that the sum of distances from xi to those already selected data points is the largest .
2. sample the data and use the sample to perform hierarchical clustering. The centroids of the resulting k clusters are used as the initial seeds
3. manually select initial seeds

Weakness5: The k-means algorithm is not suitable for discovering clusters that are not hyper-ellipsoids (or hyper-spheres).

6.reperesentation of clusters

(1). Use the centroid of each cluster to represent the cluster.

Limits: The centroid representation alone works well if the clusters are of the hyper-spherical shape.

Note: One can also compute the radius and standard deviation of the cluster to determine the spread in each dimension.

(2). Use classification models to represent clusters. The resulting tree or set of rules provide an understandable representation of the clusters.

Limits: A cluster may be split into a few hyper-rectangles or rules.

Note: One can use the set of rules to evaluate the clusters to see whether they conform to some existing domain knowledge or intuition.

(3). Use frequent values in each cluster to represent it.

Limits: This method is mainly for clustering of categorical data (e.g., in the k-modes clustering).

Note: key method used in text clustering, where a small set of frequent words in each cluster is selected to represent the cluster.

第五章：

目的：1.利用标注过的数据与未标注过的数据学习（LU）

2.利用标注为positive的数据与为标注过的数据学习（PU）

针对目的1：

方法1：利用基于朴素贝斯特分类算法的EM算法。

前提：two mixture model assumptions成立，即满足数据由混合模型生成且混合组件与类之间存在一一对应关系。第二个假设若不成立，则未标注的数据不仅不会促进学习，反而会妨碍学习。EM算法本质上是无监督学习，因为被标注的数据很少，起到的是无监督学习中类似于种子的作用，而在第一个迭代后，主要是依靠未被标注的数据来调整算法学习的参数。因此，如果假设不成立，聚类算法不会按照给定的类汇聚，由此妨碍学习。

解决办法：（1）weight the unlabeled data：改变Pr(wt|cj)的计算方式

（2）Finding Mixture Components：直击问题源头，找出类的子类，把类的子类作为无关的新类代替原先的类。

方法2：co-training，即在特征过剩的情况下划分出两个子集，分别训练学习。

前提假设1：classifier要与之兼容，即几个classifier要得出相同的class

前提假设2：两个子集之间的特征要彼此无关

具体内容：先根据已标注的案例学习f1，f2，接着利用f1选出最确定的几个标注为c1的，几个标注为c2的（c1与c2可根据已标注的数据中的比例来，此处不限于二分类）交给f2训练，反之亦然。由于彼此之间条件独立，可看成随机案例，不影响准确性。结束后，针对测试集，利用f1，f2两者的结果结合得出最终的预测。

注：一般来说比EM更有效、准确

方法3：self-training

具体内容：类似于舍去了二分特性集的co-training