k-means聚类算法的优缺点，以及有没有什么改进的方法？

原创 空字符 月来客栈 7月1日

收录于话题
#《跟我一起机器学习》

51个

这是最近在某呼收到的一个提问“k-means聚类算法的优缺点，以及有没有什么改进的方法？”下面就来谈谈自己的观点。

优点： 应用广泛，速度快，鲁棒性强；对于未知特性的数据集都可以先用Kmeans去试试。

缺点： 有倒是有，只是题主并没有指明哪一类缺点，所以这里就说一个方向的缺点 ”Kmeans在聚类过程中同等的看待每个特征维度”，当出现下列情况的数据集时就不能很好的处理：

当数据集中存在噪音维度。假定某个数据集有5个特征维度，但是其中一个是噪音维度。但是Kmeans在聚类过程中仍旧将其看成是正常的特征维度进行利用，而不能加以区分。因此就诞生了WKmeans聚类算法，这种方法在聚类时会给每个特征维度赋予一个权重，使得噪音维度的权重会尽可能的趋于0，以此来去除对聚类结果的影响。

但是WKmeans算法就完美了吗？当然没有，在新闻类数据集聚类中仍旧存在一个问题，即不能在不同簇中区分不同的有效维度。假定某个新闻数据集有8个特征维度，同时包含娱乐，财经，体育三个簇。对于娱乐这类新闻来说，其有效的特征维度为1，2，5维度；对财经新闻来说，其有效的特征维度为3，4，6；余下两个维度为体育新闻的有效特征维度。也就是说，对于某个簇无效的特征维度实际上就是噪音维度，但是由于每个簇的噪音维度又不一样，因此不能用WKmeans来解决。故而又出现了EWKmeans聚类算法。

总结就是，上述两种改进都可以说是朝着一个方向的，整体上还是基于Kmeans这个聚类框架来进行改进的。且同时基于Kmeans框架改进的算法还有很多，此处就当是抛砖引玉了。

引用

- K-means聚类算法的优化？
  https://www.zhihu.com/question/277214861/answer/1107649204

收录于话题 #《跟我一起机器学习》 51个

上一篇 下一篇