【学习】kmeans聚类理论篇K的选择（轮廓系数）

机器学习研究组订阅号 2016-06-11

摘要 数据挖掘DW

kmeans是最简单的聚类算法之一，但是运用十分广泛。最近在工作中也经常遇到这个算法。kmeans一般在数据分析前期使用，选取适当的k，将数据分类后，然后分类研究不同聚类下数据的特点。

本文记录学习kmeans算法相关的内容，包括算法原理，收敛性，效果评估聚，最后带上R语言的例子，作为备忘。

算法原理

kmeans的计算方法如下：

1 随机选取k个中心点

2 遍历所有数据，将每个数据划分到最近的中心点中

3 计算每个聚类的平均值，并作为新的中心点

4 重复2-3，直到这k个中线点不再变化（收敛了），或执行了足够多的迭代

时间复杂度：O(I*n*k*m)

空间复杂度：O(n*m)

其中m为每个元素字段个数，n为数据量，I为跌打个数。一般I,k,m均可认为是常量，所以时间和空间复杂度可以简化为O(n)，即线性的。

原文链接：

http://mp.weixin.qq.com/s?__biz=MzA3MDg0MjgxNQ==&mid=2652389773&idx=1&sn=22c9d96b8fce6db9231638df5b9193c1&scene=0#wechat_redirect