K-Means Clustering Algorithm

K-Means聚类算法

INFORMATION ENGINEERING UNIVERSITY

中国人民解放军战略支援部队信息工程大学—李响副教授

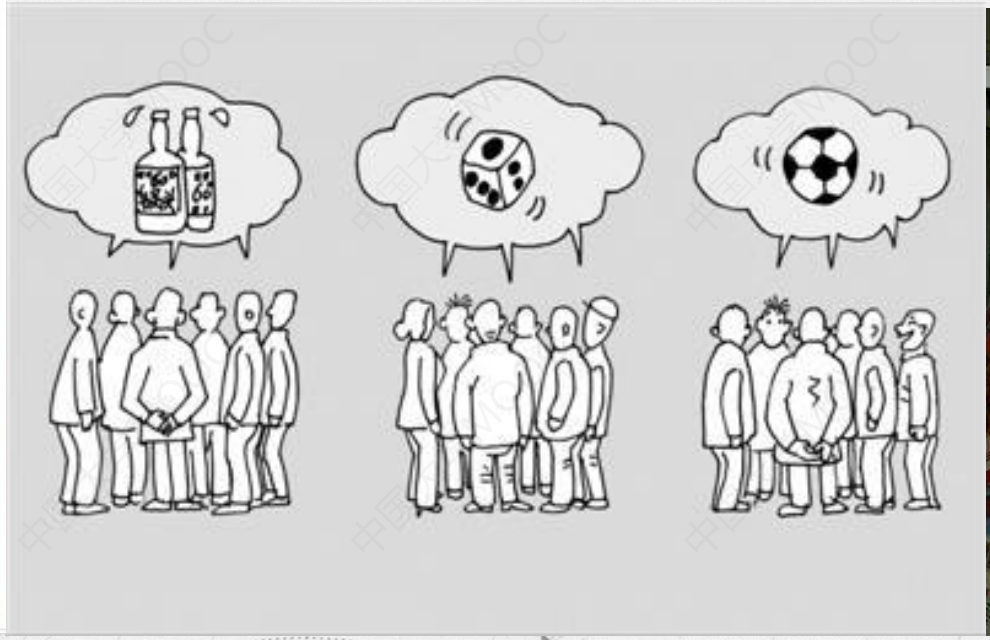
PLA Strategic Support Force Information Engineering University——A/Prof. Xiang Li

- 德国奥格斯堡大学访问学者和青年科学家，地理信息世界特聘审稿专家，测绘学报等核心期刊审稿人，高校GIS论坛十大新锐人物。
- 主要研究方向地理信息系统平台及其应用，主持国家自然科学基金，国家重点研发（子课题）等课题多项，获省部级科技进步二等奖2项，三等奖1项，部门理论成果一等奖1项，高校GIS论坛“优秀教学成果”奖1项。
- 出版和翻译著作6部，近5年，以第一作者或通讯作者发表论文16篇，发明专利2项，软件著作权3项。

K-Means聚类算法

K-Means Clustering Algorithm

物以类聚，人以群分

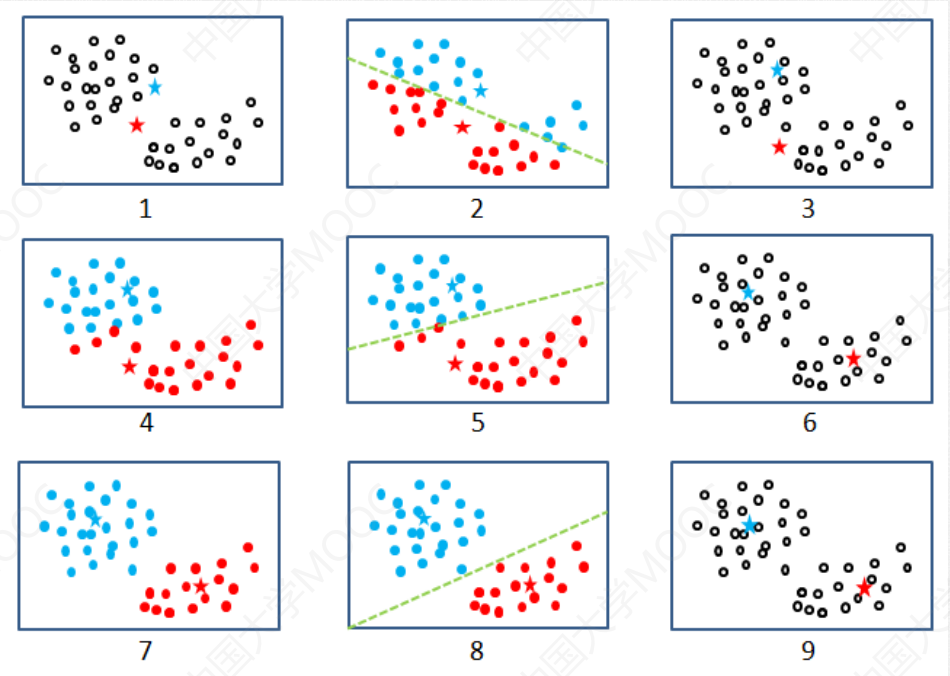


K-Means聚类算法

K-Means Clustering Algorithm

K-means聚类算法

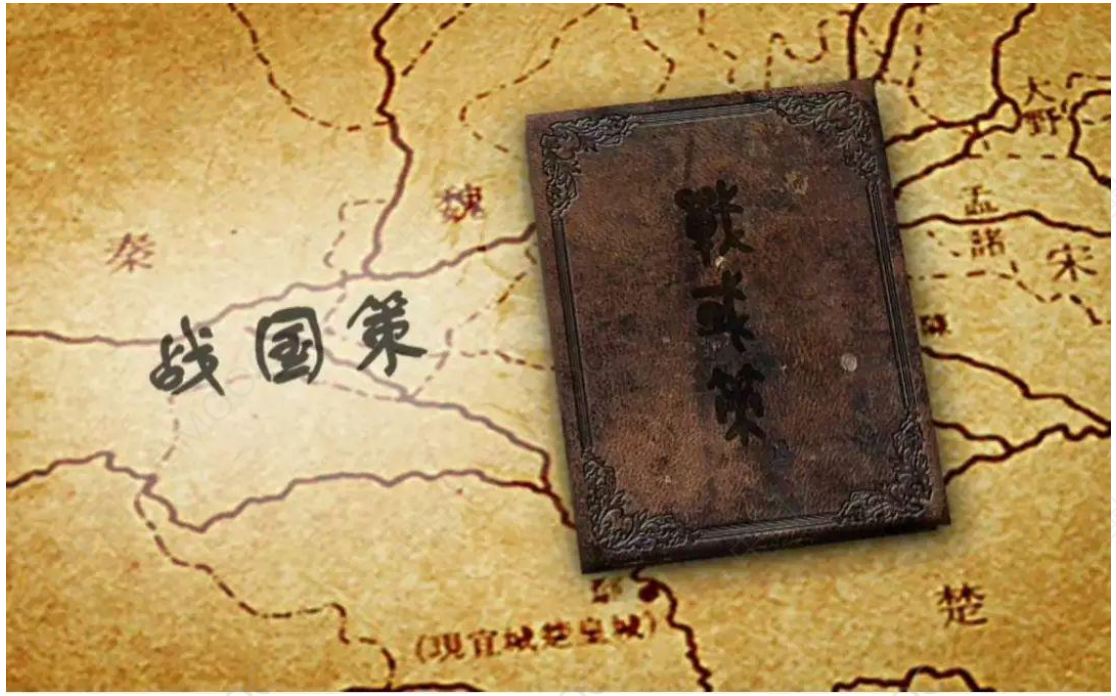
2006年IEEE国际挖掘大会中被评为十大最具影响力的数据挖掘算法之一。



K-Means聚类算法

K-Means Clustering Algorithm

自然的聚集现象



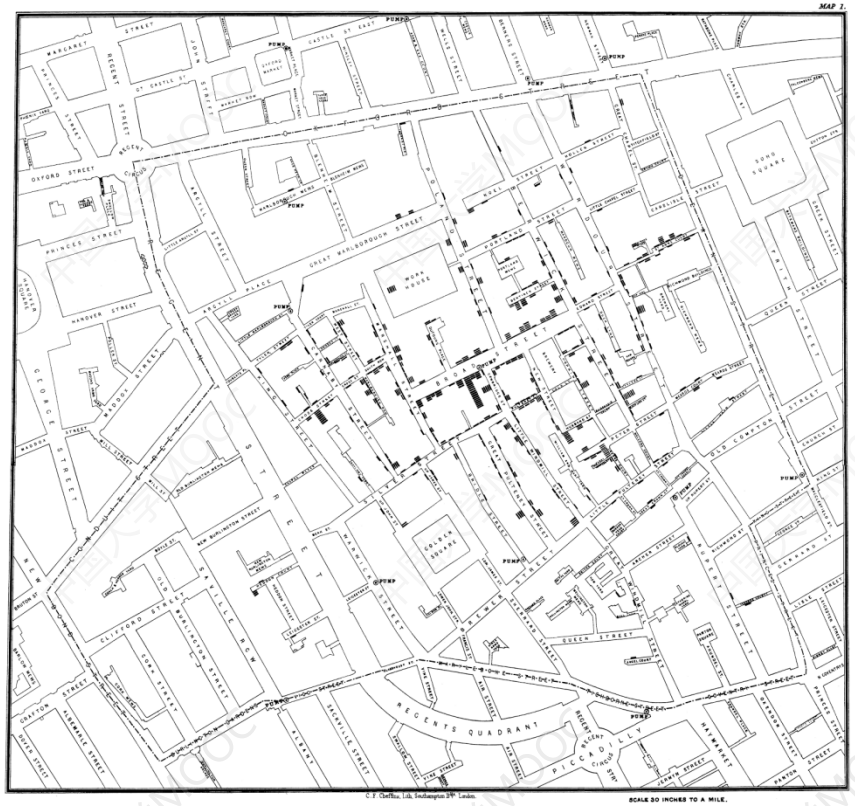
《战国策·齐策三》

“人们要寻找柴胡、桔梗这类药材，如果到水泽洼地去找，恐怕永远也找不到，要是到梁文山的背面去找，那就可以成车地找到。”

K-Means聚类算法

K-Means Clustering Algorithm

自然的聚集现象



伦敦霍乱图-John.Snow（1854）

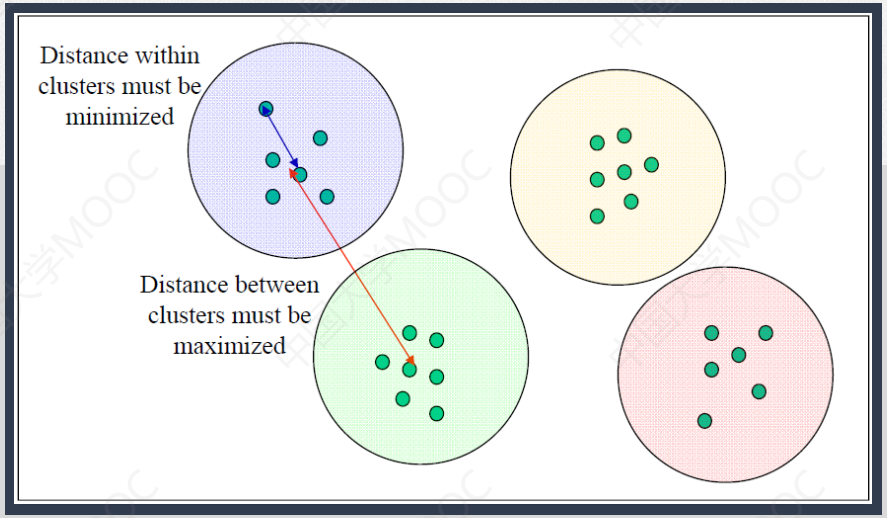
空间聚类分析最早的成功应用

K-Means聚类算法

K-Means Clustering Algorithm

聚类

在同一个集合中的所有对象尽可能的相似。



» 簇（cluster）

K-Means聚类算法

K-Means Clustering Algorithm

基于划分的空间聚类算法

基于层次的空间聚类算法

基于密度的空间聚类算法

K-Means聚类算法

K-Means Clustering Algorithm

基于划分的空间聚类算法是历史最为悠久，也是应用最为广泛的聚类算法之一。K-Means就是其中最具代表性的算法。

核心思想

对于包含n个对象的集合，给定聚类数k ($k \leq n$)，通过一定的目标划分准则，不断优化，直到将整个数据集划分为k 个划分，每个划分即为一个簇。

K-Means聚类算法

K-Means Clustering Algorithm

具体的算法流程

（1）随机选取K个点，作为K个簇的中心点；

（2）对于集合中的每个点，分别计算该点到K个簇的中心点的距离；

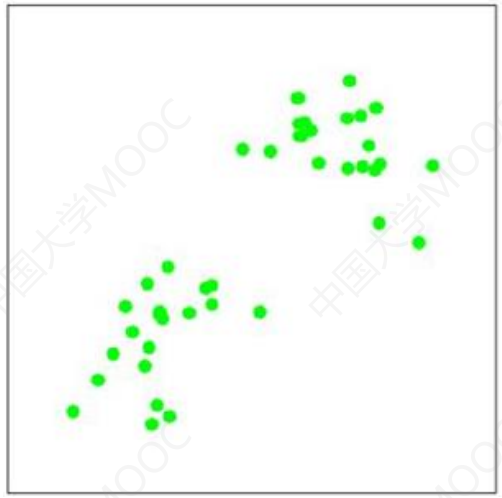
（3）按照距离最近的原则，将每个点归为不同的簇；

（4）重新计算每个簇的中心点（比如将每个簇的所有点的位置求取平均来计算中心点）

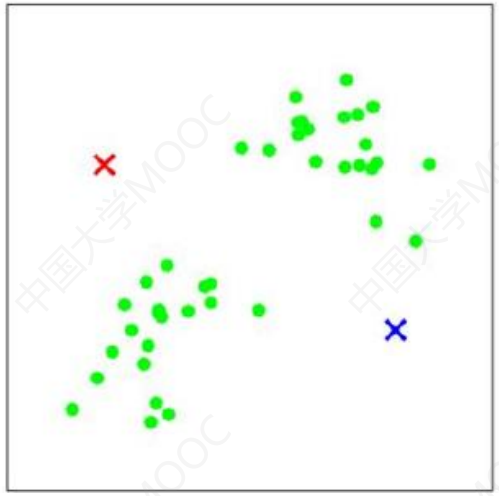
（5）如果每个簇的中心点，不再发生变化，那么该算法结束，否则跳转到第（2）步，继续执行该算法。

K-Means聚类算法

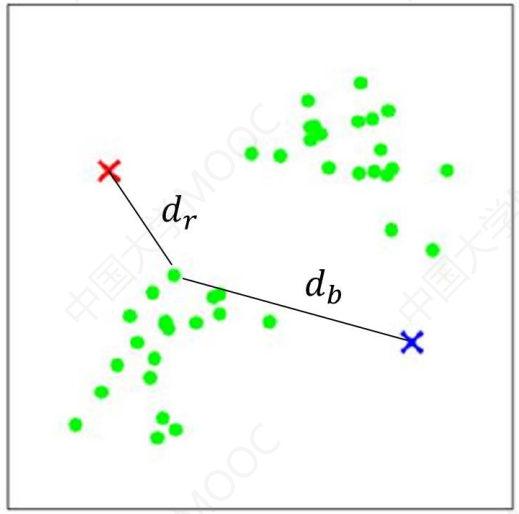
K-Means Clustering Algorithm



原始的数据集合



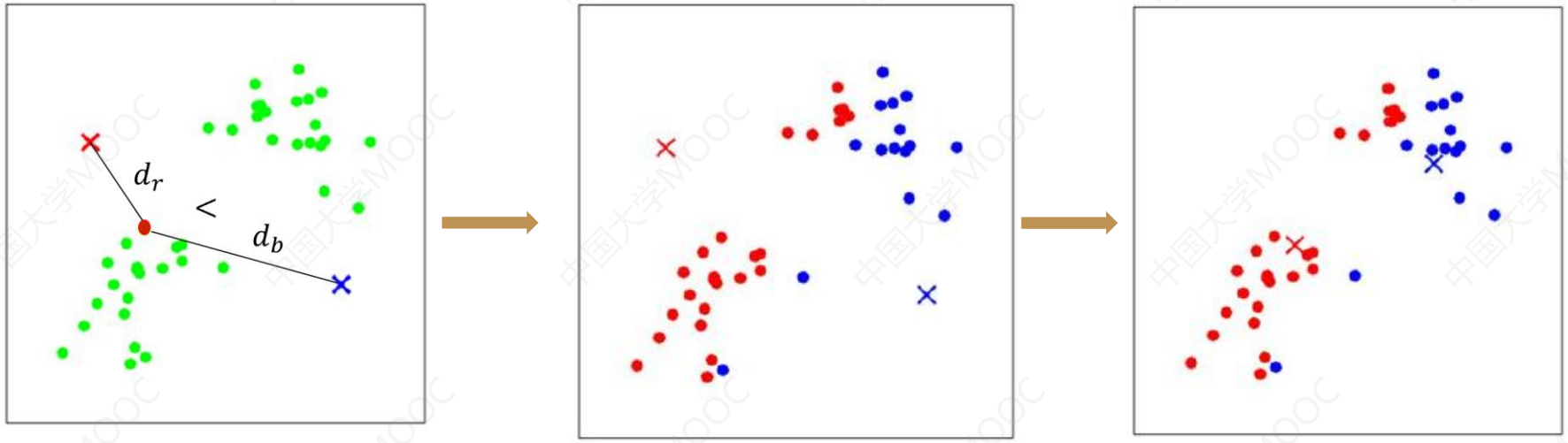
随机选取K个点作为K个簇的中心点，这里K为2



对于集合中的每个点，分别计算该点到K个簇的中心点的距离

K-Means聚类算法

K-Means Clustering Algorithm



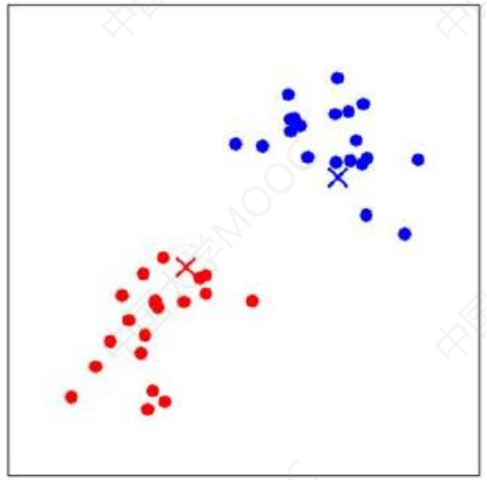
按照距离最近的原则，将每个点归为不同的簇

遍历处理每个点

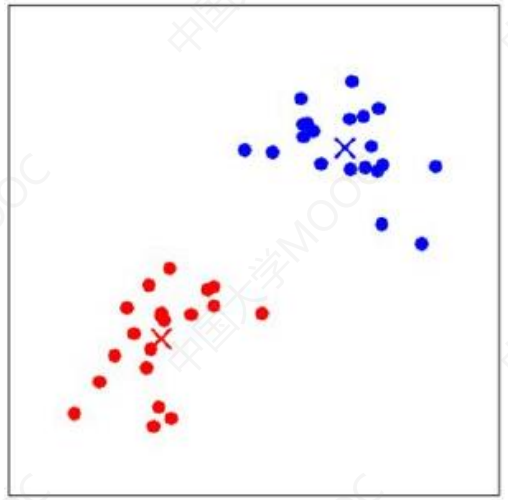
再重新计算每个簇的中心点，至此完成了第1次迭代

K-Means聚类算法

K-Means Clustering Algorithm



由于该簇的中心点明显发生了变化，因此跳转到步骤2，继续重复该过程

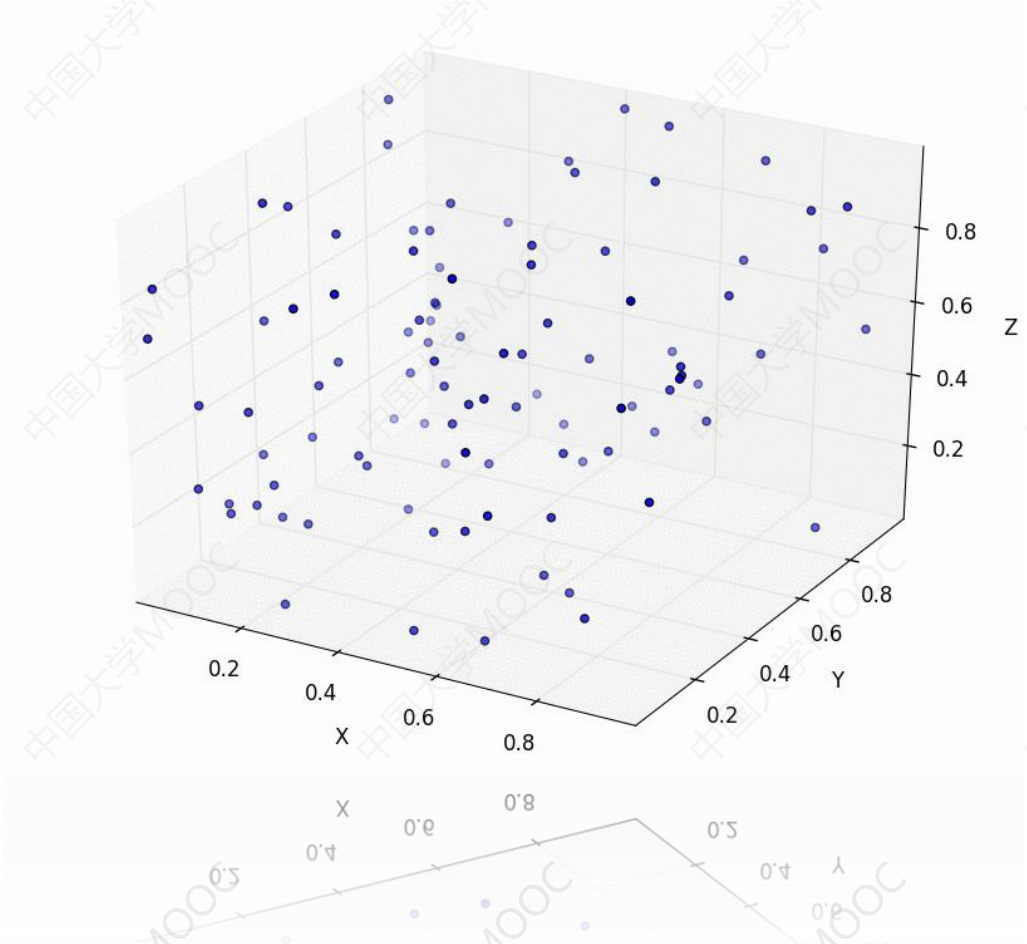


第二次迭代，直到中心点，不发生变化，该算法结束

K-Means聚类算法

K-Means Clustering Algorithm

K-Means空间聚类绝不仅限于二维点，它可以扩展到三维，甚至n维点的空间聚类。



K-Means聚类算法

K-Means Clustering Algorithm

算法复杂度

$$O(nkt)$$

- n--所有的对象个数
- k--簇的个数
- t--迭代的次数

如果n远大于k和t的话，这个算法复杂度主要取决于n，这也意味K-Means在处理大型的数据集合时，相当高效。

K-Means聚类算法

K-Means Clustering Algorithm

算法的不足

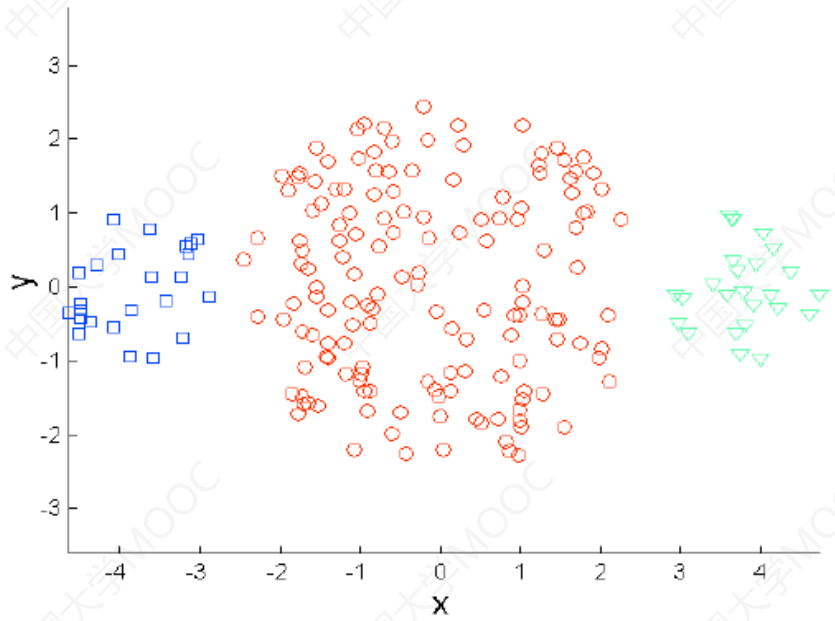
（1）需要预定义k个簇，因此难以处理分类数量不清的数据集合；

（2）K-Means算法对数据聚集的大小、形状以及密度等因素较为敏感；

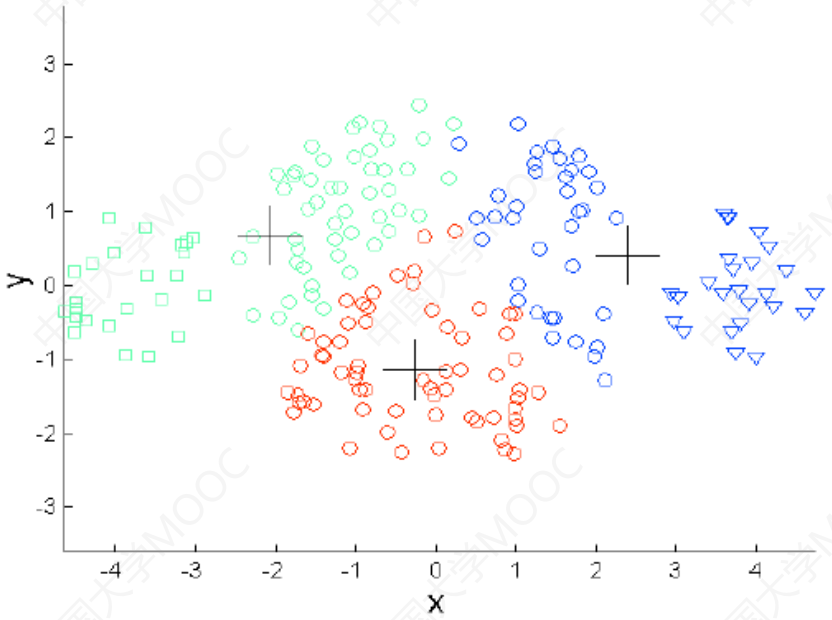
（3）K-Means算法对离群点的数据也较为敏感。

K-Means聚类算法

K-Means Clustering Algorithm



Original Points

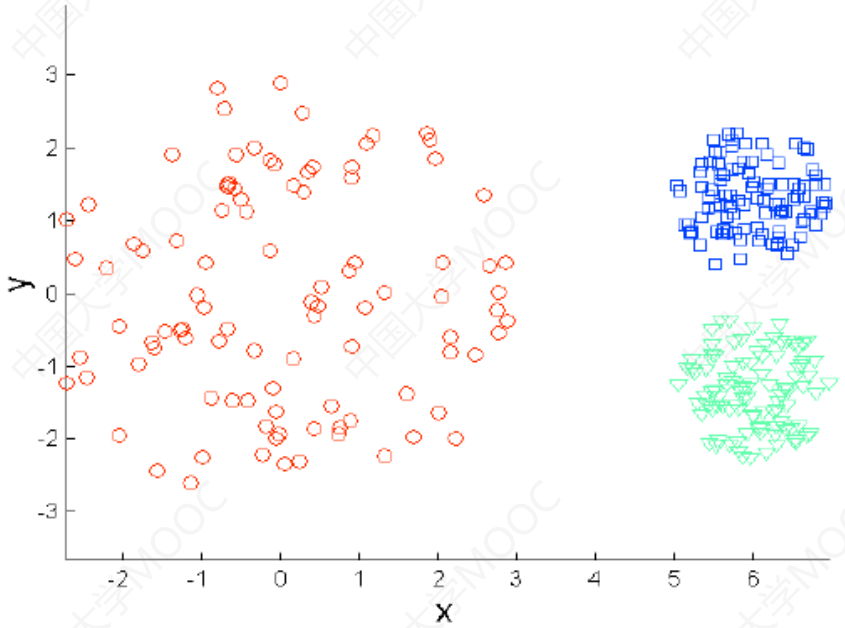


K-means (3 Clusters)

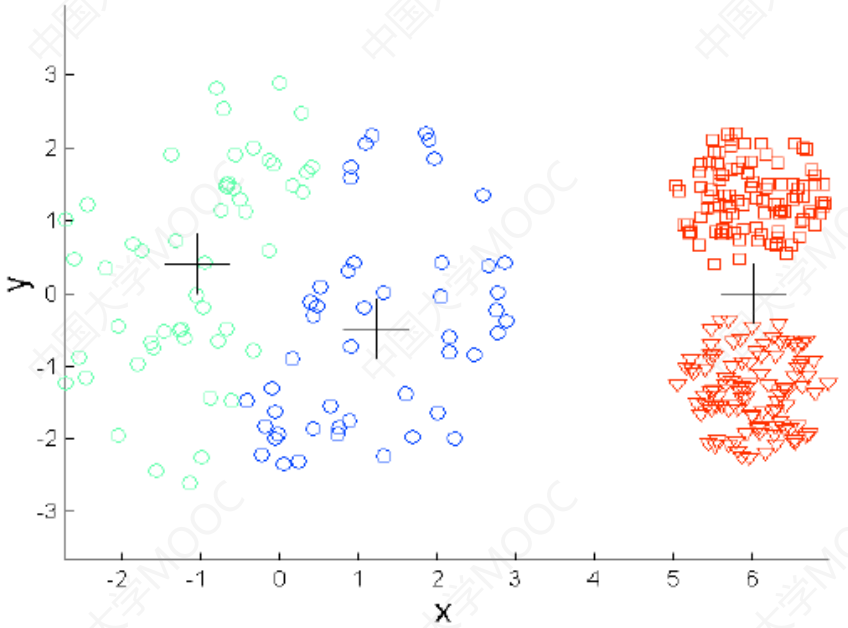
受到数据聚集的大小影响

K-Means聚类算法

K-Means Clustering Algorithm



Original Points

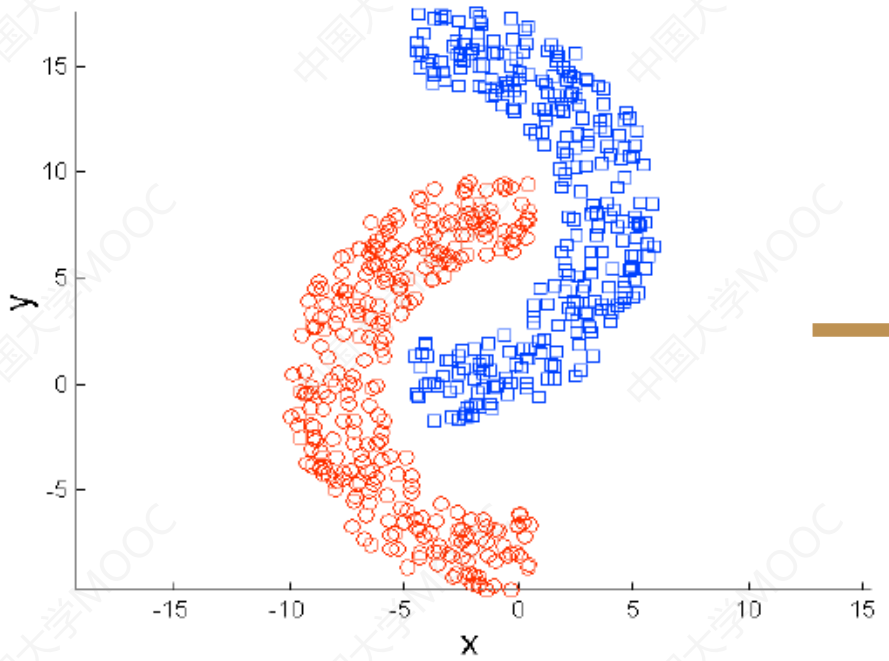


K-means (3 Clusters)

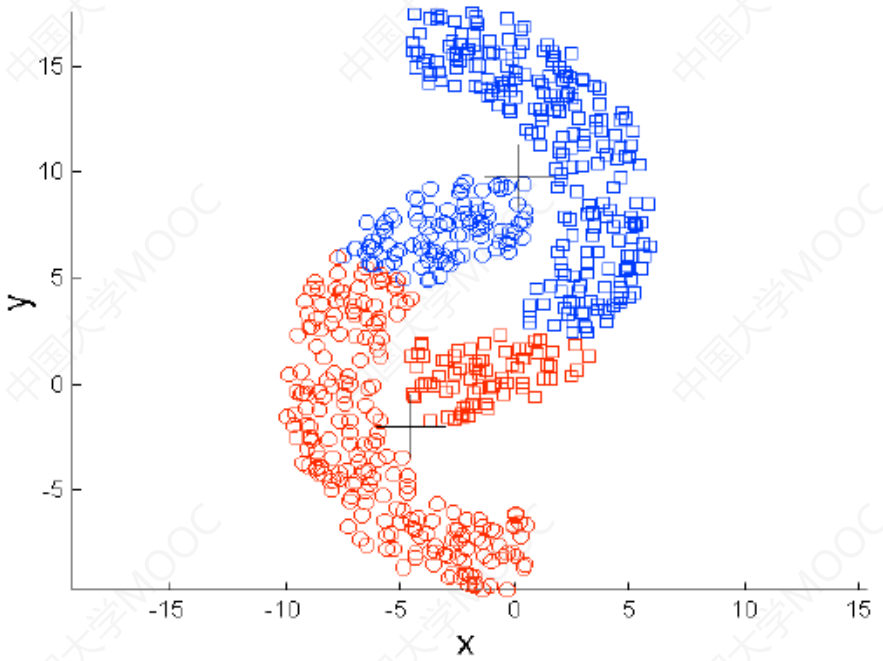
受到数据聚集的密度影响

K-Means聚类算法

K-Means Clustering Algorithm



Original Points

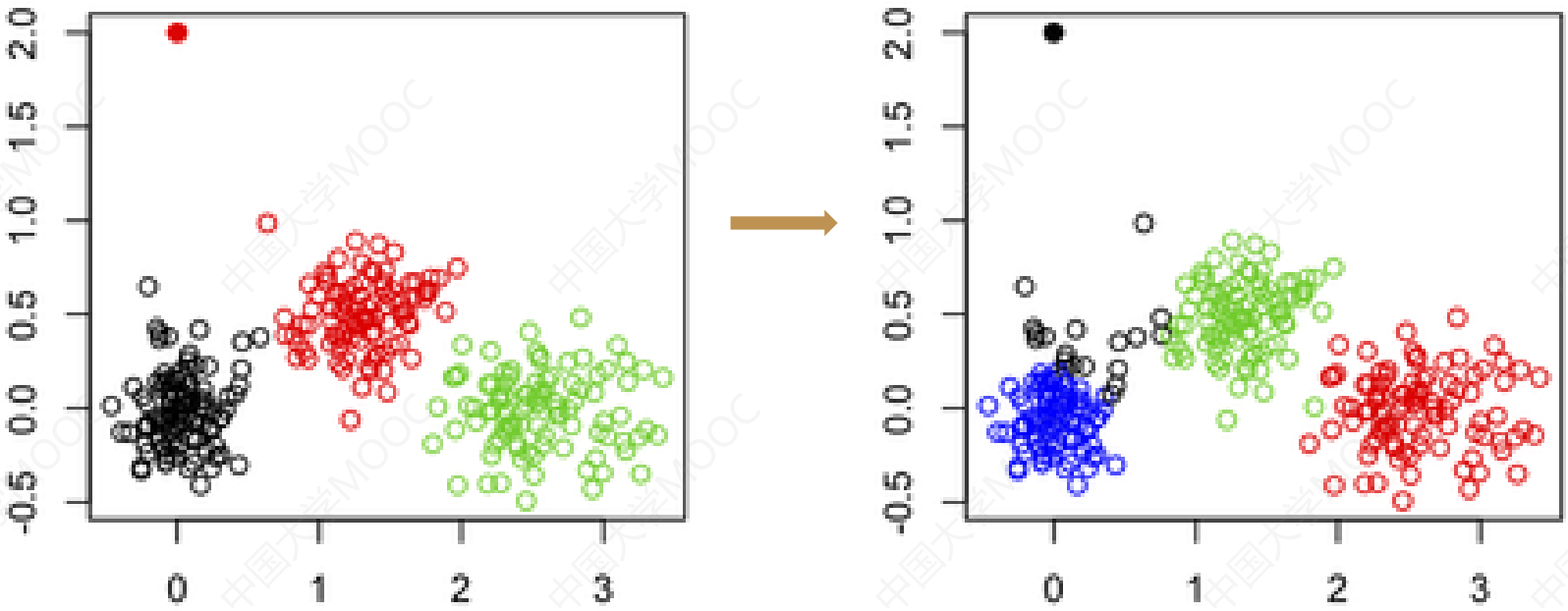


K-means (2 Clusters)

不规则形状

K-Means聚类算法

K-Means Clustering Algorithm



有离群点的情况

K-Means聚类算法

K-Means Clustering Algorithm

“Talk is cheap, show me the code”

—— Linus Torvalds（林纳斯·托瓦兹）

谢 谢 观 看