第七章聚类

2021年10月13日 16:46

1. 聚类分析

   概念：

   - 聚类分析Cluster analysis，简称聚类Clustering，是一个把数据对象划分为子集的过程。
   - 簇Cluster：每一个子集是一个簇

   特点：

   - 无监督的
   - 一个好的聚类分析：高类内相似度、低类间相似度

2. 典型聚类应用
   - 图像像素聚类
   - 图像数据分析
     - 对相似区域进行聚类，产生主题地图
     - 图像数据集聚类
   - 商务数据分析
     - 帮市场分析人员发现不同的顾客群
     - 企业信用等级聚类
   - 万维网数据分析
     - 对WEB上的文档进行聚类
     - 对WEB日志的数据进行聚类，以发现相同的用户访问模式
   - 经济领域
     - 发现不同客户群
     - 住宅区聚类，确定ATM安放位置
     - 股票市场，找出最具活力的
   - 数据挖掘领域
     - 其他算法的预处理步骤，获得数据分布状况
     - 离群点检测、数据规约
3. K-MEANS聚类方法

   类别：划分聚类方法

   特点：

   ◘ k-means聚类算法是划分聚类方法中最常用、最流行的经典算法，许多其他的方法都是k-means聚类算法的变种。

   ◘ k-means聚类算法将各个聚类子集内的所有数据样本的均值作为该聚类的代表点，算法的主要思想是通过迭代过程把数据集划分为不同的类别，使得评价聚类性能的准则函数达到最优，从而使生成的每个聚类类内紧凑，类间独立。

   ◘ k-means聚类算法不适合处理离散型属性，但是对于连续型

属性具有较好的聚类效果。

具体实现步骤演示和推理见（来源网页博客）：K-Means聚类算法原理笔记

简洁实现步骤（课堂笔记）：

输入：数据集X={xj, |j=1,2,...,total}，其中的数据样本也只包含描述属性，不包含类别属性；聚类个数k。

输出：使误差平方和准则最小的k个聚类。

（1）从数据集X中随机地选择k个数据样本作为聚类的初始代表点，每一个代表点表示一个类别。

（2）对于X中的任一数据样本xj，(1≤j≤total)，计算它与k个初始代表点的距离，并且将它划分到距离最近的初始代表点所表示的类别中。

（3）完成数据样本的划分之后，对于每一个聚类，计算其中所有数据样本的均值，并且将其作为该聚类的新的代表点，由此得到k个均值代表点。

（4）对于X中的任一数据样本xj，(1≤j≤total)，计算它与k个均值代表点的距离，并且将它划分到距离最近的均值代表点所表示的类别中。

（5）重复步骤（3）和（4），直到各个聚类不再发生变化为止，即误差平方和准则函数的值达到最优。

聚类的结果使评价聚类性能的误差平方和准则 函数的值达到最优：

$$J_e = \sum_{i=1}^{c} \sum_{x \in D_i} \left\| x - m_i \right\|^2, \qquad m_i = \frac{1}{n_i} \sum_{x \in D_i} x$$

优点：

a. 与层次聚类相比，k均值可以得到更紧密的簇，尤 其是对于球状簇
b. 对大数据，是可伸缩和高效率的
c. 法尝试找出使平方误差函数值最小的k个划分。 当结果簇是密集的，而簇与簇之间区别明显的时候， 效果较好

缺点：

a. 聚类的个数是事先给定的，如何最优
b. 对初始值敏感
c. 对噪声点敏感

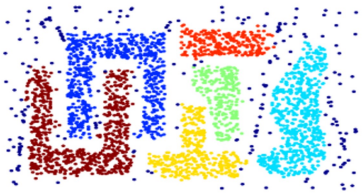
改进的k-medoids：

把k-means的均值mean改为中心点dedoid，即改为：从当前cluster 中选取这样一个点—— 它到其他所有（当前cluster中的）点的距离之和最小——作为中心点

4. 典型聚类方法

数据类型：

- 中心型：簇中心通常在平均值附近
- 密度型



- 连续型

聚类方法：

a. 划分聚类方法
   - 划分聚类方法
     - 选择合适的初始代表点将数据样本进行初始聚类，之后通过不断迭代的过程对聚类的结果进行不断的调整，直到满足聚类评价的准则为止。
     - 通常通过计算对象间距离进行划分
   - 典型的划分方法
     - k-means
     - k-中心点
     - 以上两种方法的变种

   划分聚类方法对数据集聚类时的三个要点
   - 选定某种距离作为数据样本间的相似性度量

     可以根据实际需要选择欧氏距离、曼哈顿距离或者明考斯基距离中的一种来作为算法的相似性度量，其中最常用的是欧氏距离。
   - 选择评价聚类性能的准则函数

     误差平方和准则函数表示数据集中的所有样本与相应聚类中心的方差之和，该准则的值达到最优时可以便各个聚类类内尽可能地紧凑，而各个聚类之间则尽可能地分开。
   - 选择某个初始分类，之后用迭代方法得到聚类结果，使得评价聚类的准则函数取得最优值

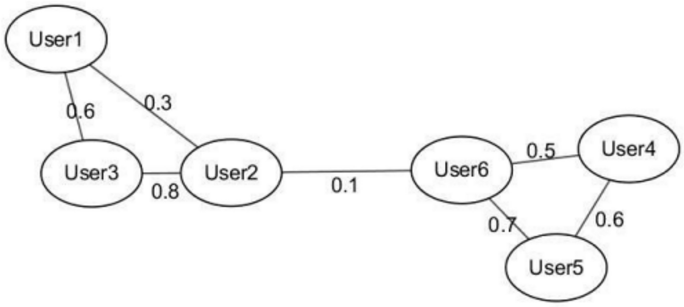
     为了得到最优的聚类结果，首先要对给定数据集进行初始划分，通常的做法是事先从数据集中选择各个聚类的代表点，之后把其余的数据样本按照某种方式归类到相应的聚类中去。

b. 密度聚类方法
   - 主要思想：只要临近区域的密度（样本的数目）超过某个阈值则继续聚类。即对于给定簇中的每个样本，在一个给定范围的区域中（例如半径ε）必须至少包含某个数目的样本
   - 主要特征：
     - 可以发现任意形状的簇
     - 可以去除噪声

- 无需预先设定簇的数量
- 需要参数作为终止条件

- 主要方法：
  - DBSCAN: Ester, et al. (KDD'96)
  - OPTICS: Ankerst, et al (SIGMOD'99)
  - DENCLUE: Hinneburg & D. Keim (KDD'98)
  - CLIQUE: Agrawal, et al. (SIGMOD'98) (more grid-based)
- 优点：
  - 克服基于距离算法只能发现“类圆形‘的聚类缺点，可以发现任意形状的聚类
  - 可以去除噪声
  - 对数据输入顺序不敏感
- 缺点：
  - 输入参数敏感：确定参数ε，MinPts困难，若选取不当，将造成聚类质量下降。由于在DBSCAN算法中，变量ε，MinPts是全局惟一的， 当空间聚类的密度不均匀、聚类间距离相差很大时，聚类质量较差
  - 计算密度单元的计算复杂度大：需要建立空间索引来降低计算量，且对数据维数的伸缩性较差。这类方法需要扫描整个数据库，每个数据对象都可能引起一次查询，因此当数据量大时会造成频繁的I/O操作

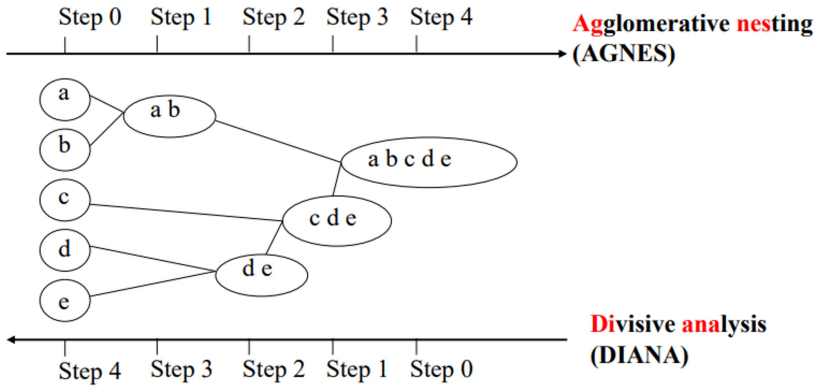
c. 图聚类方法



- 谱聚类
  - 优点：
    - 谱聚类只需要数据之间的相似度矩阵，因此对于处理稀疏数据的聚类很有效
    - 建立在谱图理论基础上，能在任意形状的样本空间上聚类，适合高维数据的聚类
  - 缺点：
    - 复杂度高
    - 谱聚类对相似度图的改变和聚类参数的选择非常的敏感

d. 层次聚类方法

- 方法：对给定数据集分层进行划分，形成一个以各个聚类为结点的树形结构
  - 分为凝聚型、分裂型：



- 改进的层次聚类方法：
  - BIRCH利用层次方法的平衡迭代规约和聚类（Balanced Iterative Reducing and Clustering Using Hierarchies）
  - Chameleon（变色龙）多阶段层次聚类
- 优点
  - 简单、易实现、容易理解
- 缺点
  - 不能纠正错误的合并或者分裂
  - 不具有很好的可伸缩性，因为合并或分裂的决定需要检查和估算大量的对象或簇

5. 评估

大分类：

- 外在方法
  - 当有基准可用时，可以将它与聚类进行比较，评估 聚类。
- 内在方法
  - 没有基准可用时，通过考察簇的分离情况和紧凑情 况评估聚类。

评估指标：

- 纯度（purity）：正确聚类的样本数占总样本数的比例
- 最小误差
  - 误差平方和准则表示数据集中的所有样本与相应聚类中心的方差之和，该准则的值达到最优时可以便各个聚类类内尽可能地紧凑，而各个聚类之间则尽可能地分开。
    - Cohesion（凝聚度） is measured by the within cluster sum of squares error(SSE)

$$WSS = \sum_{i} \sum_{x \in C_i} (x - m_i)^2$$

    - Separation（分散度） is measured by the between cluster sum of squares

$$BSS = \sum_{i} |C_i| (m - m_i)^2$$

      Where |Ci | is the size of cluster i , m is the center of all samples.
- NMI(Normalized Mutual Information)归一化互信息

- RI(Rand Index)将聚类看成一系列的决策过程
- ARI(Adjusted Rand Index)