4.8 吸引力传播
（Affinity Propagation，AP）

CSDN学院

2017年11月

常用聚类算法

- 基于距离、相似度的聚类算法
  - *K*-means（*K*均值）及其变种（ *K*-centers 、 Mini Batch *K*-Means）
  - Mean shift
  - 吸引力传播（Affinity Propagation ，AP）
  - 层次聚类
    - 聚合聚类（Agglomerative Clustering）
- 基于密度的聚类算法
  - DBSCAN、DensityPeak（密度最大值聚类）
- 基于连接的聚类算法
  - 谱聚类

Affinity Propagation

- 吸引子传播算法（Affinity Propagation 算法，AP算法）2007年发表在Science
  - Clustering by Passing Messages Between Data Points. Brendan J. Frey and Delbert Dueck, University of Toronto，*Science* **315**, 972–976, February 2007
- AP算法相对于*K*-means的优势：
  - 无需要指定聚类数量
  - 对初始值的不敏感
- AP算法的缺点：计算复杂度高$O(N^2T)$，其中N为样本数目，T为迭代次数
  - 适合中小规模的数据

Affinity Propagation

- AP 算法用样本间的相似矩阵作为输入
- 相似矩阵对角线上的元素叫preference： 描述每个样本作为聚类中心的适合程度
  - 通常取相似矩阵的中值，表示所有样本作为中心的可能性相等
  - preference 会影响聚类数量的多少， preference越小，聚类数就会相对较少

Affinity Propagation

- 吸引子传播：算法主要考虑样本间的消息传递：
  - $r(i,k)$： responsibility，描述数据对象k适合作为数据对象i的聚类中心的程度。这是节点i传递向节点k的信息，即节点k对节点i的吸引度
  - $a(i,k)$ ： availability，数据对象i选择数据对象k作为其据聚类中心的适合程度
  - 聚类中心：与大多数样本足够相似（$r(i,k)$大）、且被很多样本选为代表样本（ $a(i,k)$大）

Affinity Propagation

- 吸引度（ responsibility ） $r(i,k)$：

样本i不太可能取其他点k'作为代表点
样本i与其他点k'的相似度小

$$r(i,k) \leftarrow s(i,k) - max[a(i,k') + s(i,k') \forall k' \neq k]$$

  - 其中s为相似度矩阵， $s(i,k)$表示样本i和样本k之间的相似度

- 归属度（ availability ） $a(i,k)$ ：

$$a(i,k) \leftarrow min[0, r(k,k) + \sum_{i' \ s.t. \ i' \notin \{i,k\}} r(i',k)]$$

如果节点k作为其他节点i'的聚类中心的合适度很大，那么节点k作为节点i的聚类中心的合适度也可能会较大

- 迭代：不断更新每一个点的吸引度和归属度值
  - 初始值：r和a都置为0

Affinity Propagation

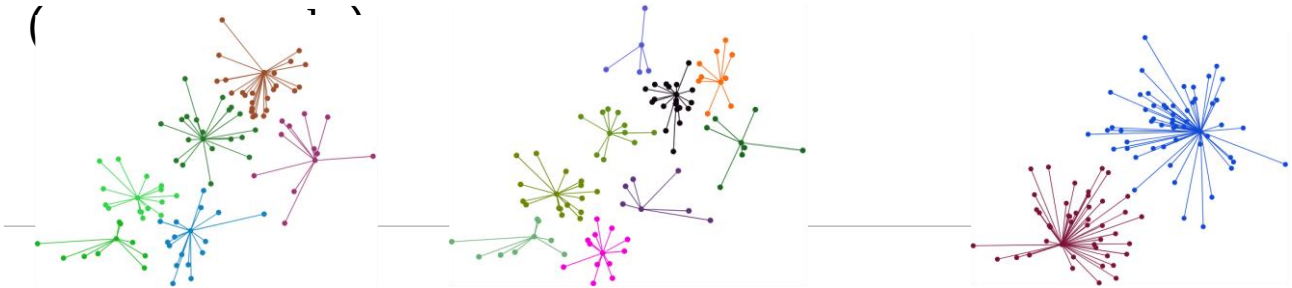
- 阻尼因子λ：为了避免消息传递中的数值震荡，引入阻尼因子（damping factor）：

$$r_{t+1}(i,k) = \lambda \cdot r_t(i,k) + (1-\lambda) \cdot r_{t+1}(i,k)$$

$$a_{t+1}(i,k) = \lambda \cdot a_t(i,k) + (1-\lambda) \cdot a_{t+1}(i,k)$$

Scikit learn中AP算法的实现

- *class* sklearn.cluster.AffinityPropagation(*damping=0.5, max_iter=200, convergence_iter=15, copy=True, preference=None, affinity='euclidean', verbose=False*)
  - damping=0.5：阻尼系数
  - *preference=None*：描述每个样本作为聚类中心的适合程度，缺省值为相似矩阵的中值，表示所有样本作为中心的可能性相等，preference 会影响聚类数量的多少，越小聚类数就会相对较少
  - *affinity='euclidean'*：相似度形式，缺省为欧式距离，也可以预计算



AP聚类算法的优缺点

- 无需指定聚类“数量”参数
  - 不过*preference*参数起到类似的作用
- 对距离矩阵的对称性没要求
  - AP通过输入相似度矩阵来启动算法，因此允许数据呈非对称，适用范围宽
- 初始值不敏感
  - 多次执行AP聚类算法，得到的结果是完全一样的，即不需要进行随机选取初值步骤（对比*K*-Means的随机初始值）
- 算法复杂度较高，为$O(N^2\log N)$，而K-Means只是$O(N*K)$
  - 不适合N比较大时(N>3000)的场合
- 若以误差平方和来衡量算法间的优劣，AP聚类比其他方法的误差低

CSDN
不 止 于 代 码

THANK YOU

AI100