聚类算法综述

Sunstone Zhang

聚类（Cluster）：相似文档的分组表达方式。在向量空间模型中，用户可以通过比较查询向量和聚类的中心进行检索，并在聚类中进一步检索以找到最相似的文档。

向量空间模型（Vector Space Model ）：文档和查询的一种表达方式，将它们转换为向量。向量的特征一般是对应文档或查询中处理过的单词（取过词根并且删除了 stopword）。这些向量被赋予权重以强调那些与语义相关的词目，这在检索中很有用。在检索中比较查询向量和文档向量，并将最接近的文档作为相关文档返回给用户。SMART 是使用向量空间模型的最有名的例子。

1. 分层次聚类法（最短距离法）

思路：寻找“ 距离 ”最近的两个样本结合

1.有 N 个样本的集合 $Z_s=\{Z_1, Z_2, ..., Z_N\}$

2.若想要聚成 K 个类 （事先给定 K）

[1] k=N, $C_i=\{Z_i\}$, i=1,2,...,N

[2] if k=K then END

[3] 找到 C_i 与 C_j 之间的距离 $d(C_i, C_j)$最小的一对

[4] C_i 和 C_j 合成一个类 C_i, 并计算新的 C_i 的中心

[5] 去除 C_j, k=k-1. goto [2]

类间距离 $d(C_i, C_j)$

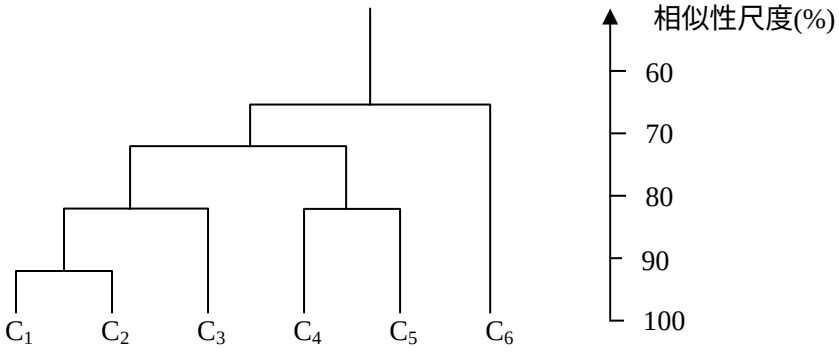
1. 类中心间距：$d_1=||M_i - M_j||$，其中 $M_i = \frac{1}{n_i}\sum_{Z\in C_i} Z$ ，n_i 是属于 C_i 的样本数。

2. 靠得最近的样本：$d_2 = \min_{Z_i \in C_i, Z_j \in C_j} \| Z_i - Z_j \|$

3. 离得最远的样本：$d_3 = \max_{Z_i \in C_i, Z_j \in C_j} \| Z_i - Z_j \|$

4. 类间平均距离：$d_4 = \frac{1}{n_i n_j} \sum_{Z_i \in C_i} \sum_{Z_j \in C_j} \| Z_i - Z_j \|$

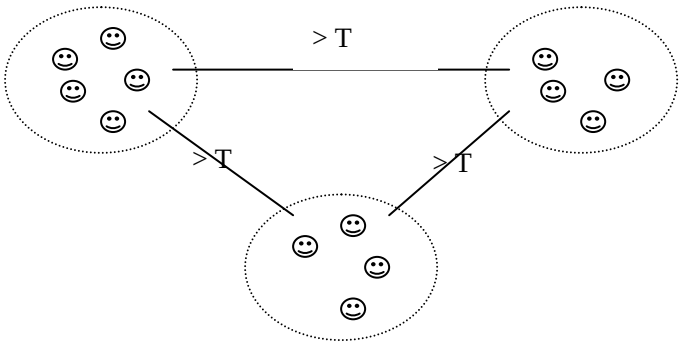
距离计算的次数：$C_N^2 = N(N-1)/2$。组合C_{N-1}^2，C_{N-2}^2，...



2. 最简单的聚类方法

相似性尺度（距离）阈值，不需要事先给定 K。
有 N 个样本，$Z_s=\{Z_1, Z_2, ..., Z_N\}$
给定一个阈值 T。
任取一个样本，例如 Z_1，把 Z_1 作为第一个类的中心，$Z_1=Z_1$
然后依次取 Z_i（i=2,3,...,N），计算 Z_1 与 Z_i 的距离 D_{1i}
若 $D_{1i} \leq T$，则判定 Z_i 属于 Z_1 为中心的那个类；
若 $D_{1i} > T$，则把 Z_i 作为新的类中心 Z_2。
然后对剩下的样本 Z_i 分别计算与 Z_1，Z_2 的距离 D_{1i}，D_{2i}
若其中较小者≤T，则判定 Z_i 属于较小的那一类
否则，就把 Z_i 作为新的一个类的中心 Z_3
如此，继续...，直至对全体样本做完处理。

特点：不需要事先决定类数。适用于类内距离小，类间距离大的情况。否则结果与取样本的顺序有关，亦与 T 相关。



3. 最大距离样本

思路：取尽可能离得远的样本做中心。

有 N 个样本，$Z_s=\{Z_1, Z_2, ..., Z_N\}$

[1] 任取一个样本，例如 Z_1，把 Z_1 作为第一个类的中心，$Z_1=Z_1$

[2] 从集合 Z_s 中找出到 Z_1 距离最大的样本作为 Z_2

[3] 对 Z_s 中剩余样本 Z_i，分别计算到 Z_1，Z_2 的距离。令其中较小的那个为 D_{z_i}

[4] 计算 $\max_{Z_s}\{D_{Z_i}\}$。若其值大于某一计算值或给定阈值，则取此 Z_i 为新的类中心。计算值可取：大于等于 Z_1 和 Z_2 间距离的 n/m 倍（$\frac{1}{2} \le n/m < 1$）。

[5] 重复同样的处理，直到再也找不到符合条件的新的类中心。

[6] 把剩余样本分配到离它最近的那个中心所属的类

缺点：与首先选取哪个样本有关。

4. K 平均聚类法（距离平方和最小聚类法）

[1] 假设要聚成 K 个类。由人为决定 K 个类中心 $Z_1(1), Z_2(1), ..., Z_k(1)$。

[2] 在第 k 次叠代中，样本集{Z}用如下方法分类：

对所有 i=1,2,..,K，i≠j

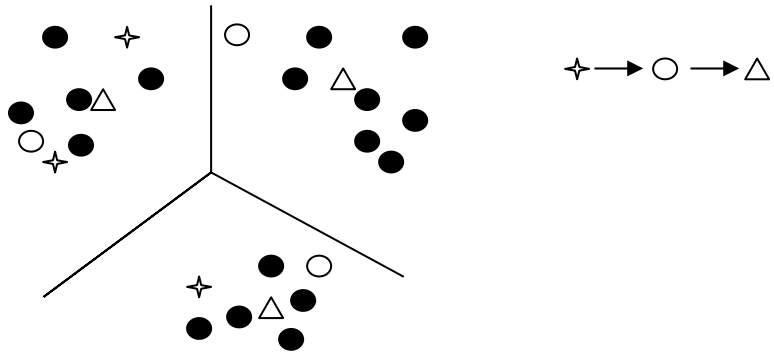
若 $\| Z - Z_j(k) \| < \| Z - Z_i(k) \|$，则 $Z \in S_j(k)$

[3] 令由[2]得到的 $S_j(k)$的新的类中心为 $Z_j(k+1)$

令 $J_j = \sum_{Z \in S_j(k)} \| Z - Z_j(k+1) \|^2$ 最小。 j=1,2,...,K

则 $Z_j(k+1) = \frac{1}{N_j} \sum_{Z \in S_j(k)} Z$，$N_j$: $S_j(k)$中的样本数。

[4] 对于所有的 j=1,2,...,K，若 $Z_j(k+1)=Z_j(k)$，则终止。否则 goto [2]。



开始时 K 的选择：最初选择哪些样本作为中心，将对叠代产生影响。多次叠代，多次修正。

5. 叠代自组织（ISODATA）聚类法

Iterative Self-Organizing Data Analysis Technology Algorithm

思路：给定一些大致参数（根据目的）。

原则：①样本数太少的类 - 取消；②类内离散太大的类 - 分裂；③距离近的类 - 合并。

1) 给一些参数

K：期望分类个数的大致范围

θ_K ：一个类内的最少样本数

θ_S ：关于类内分散程度的参数

θ_C ：关于类间距离（最小）的参数

L：每次叠代允许合并的类数

I：允许叠代的最大次数

2) 适当选取类中心$\{Z_1, Z_2, ..., Z_{Nc}\}$，$N_c$：类数

2)’分配样本。如果有$\{i=1,2,...,N_c\}$

$\| Z - Z_j \| \le \| Z - Z_i \|$，则 $Z \in S_j, j = 1,2,..., N_c$

3) 如果 S_j 类样本数 $N_j < \theta_k$，则取消 S_j 类。$N_c=N_c-1$, goto 2)’

4) 重新计算各类中心 $Z_j = \frac{1}{N_j} \sum_{Z \in S_j} Z, j = 1,2,..., N_c$

5) 计算类 S_j 内平均距离 $\overline{D_j} = \frac{1}{N_j} \sum_{Z \in S_j} \| Z - Z_j \|, j = 1,2,..., N_c$

6) 对全体样本求类内距离平均值 $\overline{D} = \frac{1}{N} \sum_{j=1}^{N_c} N_j \cdot \overline{D_j}, N = \sum_{j=1}^{N_c} N_j$

7) [a]如果叠代次数≥I，则转向 11)（合并）

[b]若 $N_c \le K/2$，则转向 8)（分裂）

[c]若偶数次叠代或 $N_c \ge 2K$，则转向 11)（合并）

8) 计算各类中各分量的标准差

$\sigma_{ij} = \sqrt{\frac{1}{N_j} \sum_{Z \in S_j} (x_{ik} - z_{ij})^2}$，i=1,2,...,n，j=1,2,...,$N_c$，k=1,2,...,$N_j$

x_{ik} 为 $Z \in S_j$ 的第 i 个分量，z_{ij} 为 Z_j 的第 i 个分量。

σ_{ij} 为第 j 类第 i 个分量标准差

9) 找到各类的标准差最大的分量

$$\sigma_{j\max} = \max\{\sigma_{1j}, \sigma_{2j}, ..., \sigma_{nj}\}, j = 1,2,...N_c$$

10) 分裂：条件 1. $\sigma_{j\max} > \theta_S$ 且 $\overline{D_j} > \overline{D}$ 且 $N_j > 2(\theta_k + 1)$

条件 2. $\sigma_{j\max} > \theta_S$ 且 $N_c \le K/2$

若满足两条件之一，则分裂 S_j

(a)建立 Z_j^+ 和 Z_j^-，2 个新的类中心，$N_c=N_c+1$

其中 Z_j^+ 和 Z_j^- 是沿着 $\sigma_{j\max}$ 轴，在原来的 Z_j 位置上，分别加上和减去一个数 $k\sigma_{j\max}(0 < k \le 1)$。k 是经验值。

(b) goto 2)'（分配样本）。

11) 计算所有各类中心的相互距离 $D_{ij} = \| Z_i - Z_j \|, i = 1,2,..., N_{c-1}, j = i+1,..., N_c$

12) 对于比 θ_c 小的 D_{ij} 从小到大排队。假定为

$$D_{i_1 j_1} \le D_{i_2 j_2} \le ... \le D_{i_L j_L}$$

13) 按 l=1,2,...,L 的顺序，把 $D_{i_l j_l}$ 对应的 Z_{i_l} 和 Z_{j_l} 合并

$$Z_l^* = \frac{1}{N_{i_l} + N_{j_l}} \left[N_{i_l} Z_{i_l} + N_{j_l} Z_{j_l} \right], N_c = N_c - 1$$

计算 $D_{i_l j_l}$ 时的 Z_i，Z_j，若至少其中一个是在本次叠代中合并取得类中心，则越过此项。

14) 若叠代次数≥I，或参数无改变，则终止。
否则 goto 2)'，需要时可返回 1)修改参数。

6. ISODATA 法的改进

聚类好：满足客观需要，客观标准
客观性：类内近可能相似（类内距小），类间相似性尽可能小（类间距大）

定义一个类间相似性：

定义：D_{ii} 是类 i 的类内离散程度，例如 $D_{ii}=\left[\frac{1}{N_i}\sum_{j=1}^{N_i}\| Z_j - Z_i \|^2\right]^{1/2}$，其中 N_i 为 i 类中样本数，Z_i 为 i 类的中心。

类间距：$D_{ij}=\left[\sum_{k=1}^{n}(z_{ki}-z_{kj})^2\right]^{1/2}$，其中 z_{ki} 为 Z_i 中的第 k 个分量。

情况⑴ 若 $D_{jj}=D_{kk}$，且 $D_{ij}<D_{ik}$，则 $R_{ij}(D_{ii},D_{ij},D_{jj})>R_{ik}(D_{ii},D_{ik},D_{kk})$，其中 R_{ij}, R_{ik} 为相似度。

情况⑵ 若 $D_{ij}=D_{ik}$，且 $D_{jj}<D_{kk}$，则 $R_{ij}(D_{ii},D_{ij},D_{jj})>R_{ik}(D_{ii},D_{ik},D_{kk})$。

定义：类间相似性（测度）

$R_{ij}=\frac{D_{ii}+D_{jj}}{D_{ij}}$，其中 D_{ii}, D_{jj}, D_{ij} 有各种定义。

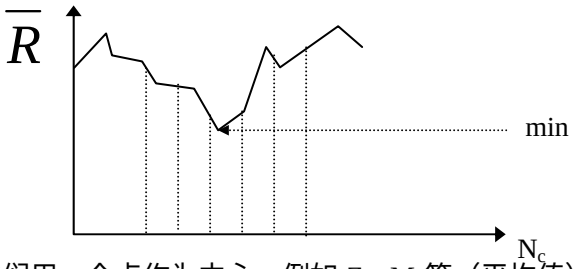
在 ISODATA 每次分配样本后，评估某一类 i，与其它所有各类的相似性计算，i=1,2,...,N_c，N_c 为当前类数。

当 $i\neq j$，令 $R_i=\max_{j,\,j\neq i}\{R_{ij}\}$

计算 $\overline{R}=\frac{1}{N_c}\sum_{i=1}^{N_c}R_i$ ：各类相似性最大值的平均。

当 $\overline{R}$ 最小时，可以认为聚类最优。

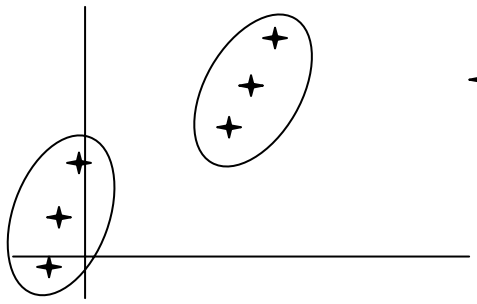
本方法作为一种评估标准，用于调整聚类情况。
而 ISODATA 法是一种手段，一种过程
手段和评估可以分离，即该评估标准不一定要用 ISODATA 法。
例：某 225 个样本



一般，每类我们用一个点作为中心，例如 Z_i，M_i 等（平均值）。如果分布对于中心非完美对称，则结果有时不能令人满意。

7. 基于“核”的评估聚类方法

例：



类内距离（离散程度）

属于该类样本 $Z \in S$，Z 到 S_j 类“ 核 ”的距离

*假定全部样本已聚成 N_c 类 $S = \{S_1, S_2, ..., S_{Nc}\}$

当 $i \neq j$ 时，$S_i \cap S_j = \phi$

每个类都有自己的“ 核 ”，$E = \{E_1, E_2, ..., E_{Nc}\}$

每个核内样本数，$M_1, M_2, ..., M_{Nc}$

每个类内样本数，$N_1, N_2, ..., N_{Nc}$

定义：样本 Z 到核内一点 Z 距离为 d(X, Z)，例如欧氏距离

样本 Z 到 i 类“ 核 ”的距离 $D(Z, E_i) = \sum_{Z \in E_i} d(X, Z)$

定义每一个类内距离和：

$$D(E_i, S_i) = \sum_{X \in S_i} \sum_{Z \in E_i} d(X, Z)$$

X 分配到各类的原则：若 $D(X, E_i) \leq D(X, E_j)$，则判定 $X \in S_i$。

如何确定 E_i？

目前为计算方便，各类核的点数一般相同。对属于 S_i 类的 N_i 个样本，每次取 M 个样本，计算类内距离 $C_{N_i}^{M_i}$ 哪一组 E_i 类内距离小，但计算量大。