第二章　聚类分析

2.1　聚类分析的概念

2.2　模式相似性测度

2.3　类的定义与类间距离

2.4　准则函数

2.5　聚类的算法

2.3.2 类间距离测度方法

(1) 最近距离法

(2) 最远距离法

(3) 中间距离法

(4) 重心距离法

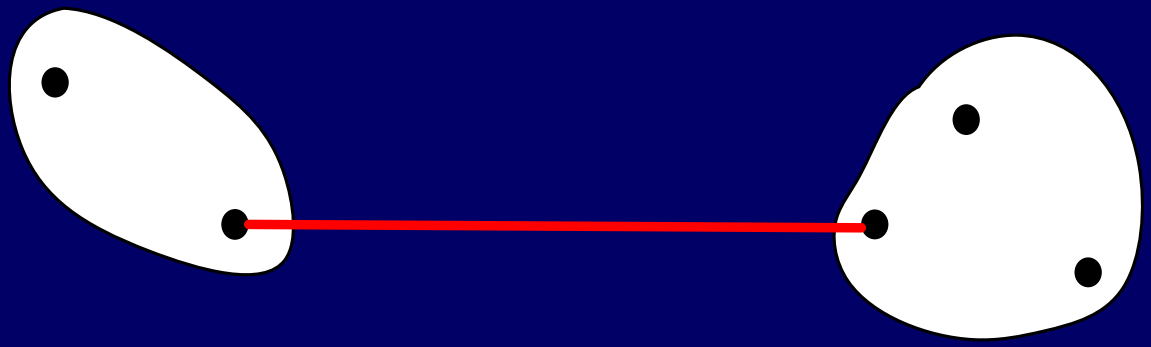
(5) 平均距离法

(6) 离差平方和法

$$D_{kl} = \min_{i,j} \left\lfloor d_{ij} \right\rfloor$$

式中 d_{ij} 表示 $\vec{x}_i \in \omega_k$ 和之间 $\vec{x}_j \in \omega_l$ 。

2.3.2 类间距离测度方法



最近距离法图示

2.3.2 类间距离测度方法

(1) 最近距离法

(2) 最远距离法

(3) 中间距离法

(4) 重心距离法

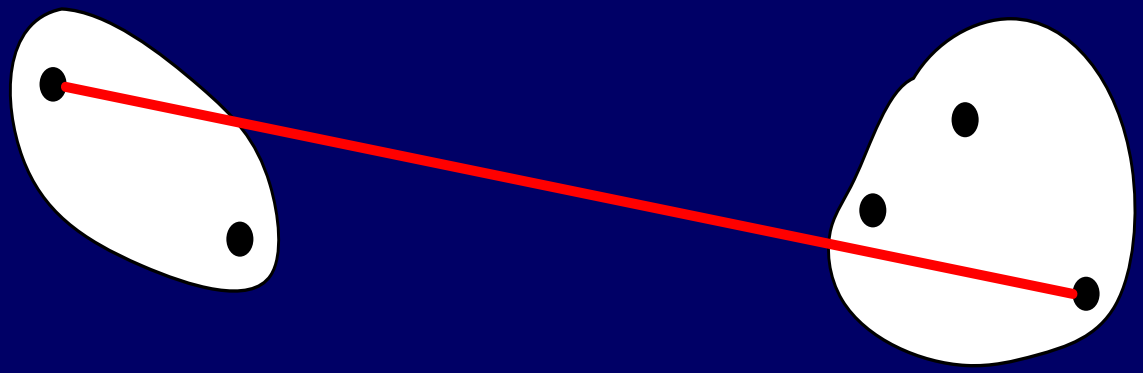
(5) 平均距离法

(6) 离差平方和法

$$D_{kl} = \max_{i,j} \left[d_{ij} \right]$$

式中 d_{ii} 表示 $\vec{x}_i \in \omega_k$ 和 $\vec{x}_j \in \omega_l$ 之间的距离。

2.3.2 类间距离测度方法



最远距离法图示

2.3.2 类间距离测度方法

(1) 最近距离法

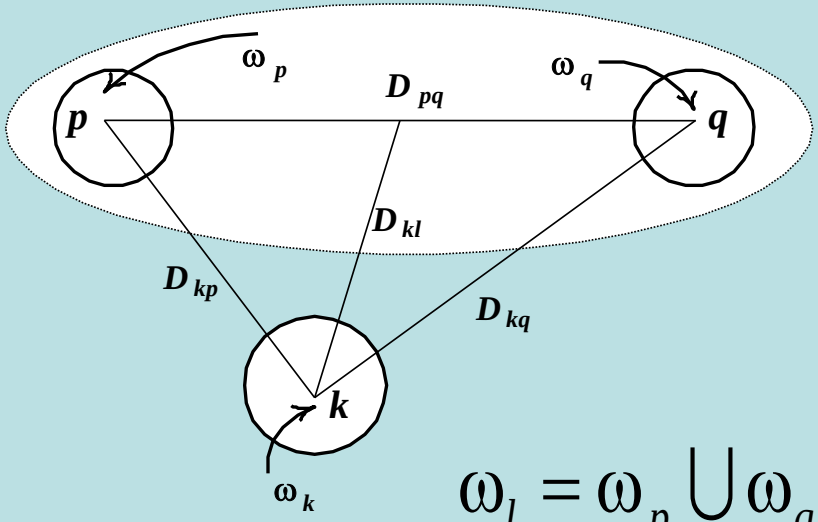
(2) 最远距离法

(3) 中间距离法

(4) 重心距离法

(5) 平均距离法

(6) 离差平方和法



$$\omega_l = \omega_p \bigcup \omega_q$$

$$D^2_{kl} = \frac{1}{2}D^2_{kp} + \frac{1}{2}D^2_{kq} - \frac{1}{4}D^2_{pq}$$

2.3.2 类间距离测度方法

$$D_{kl}^2 = \frac{n_p}{n_p + n_q} D_{kp}^2 + \frac{n_q}{n_p + n_q} D_{kq}^2 - \frac{n_p n_q}{(n_p + n_q)^2} D_{pq}^2$$

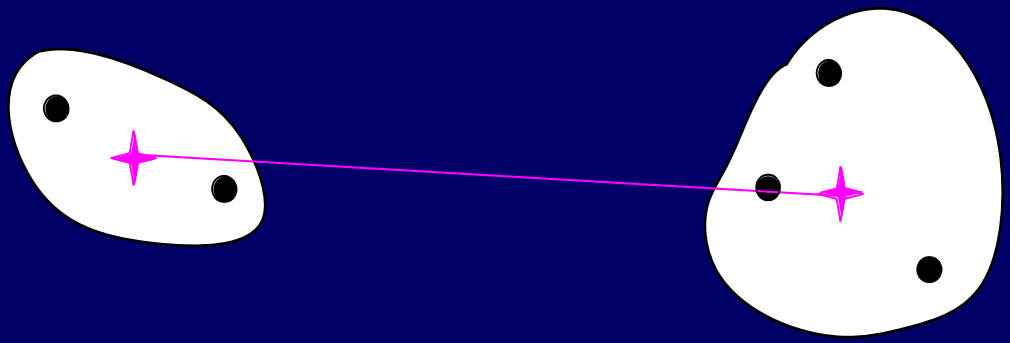
(4) 重心距离法

(5) 平均距离法

(6) 离差平方和法

n_p, n_q分别为类ω_p和ω_q的样本个数

2.3.2 类间距离测度方法



重心距离法图示

$$D_{kl}^2 = \frac{n_p}{n_p + n_q} D_{kp}^2 + \frac{n_q}{n_p + n_q} D_{kq}^2 - \frac{n_p n_q}{(n_p + n_q)^2} D_{pq}^2$$

2.3.2 类间距离测度方法

(1) 最近距离法

(2) 最远距离法

(3) 中间距离法

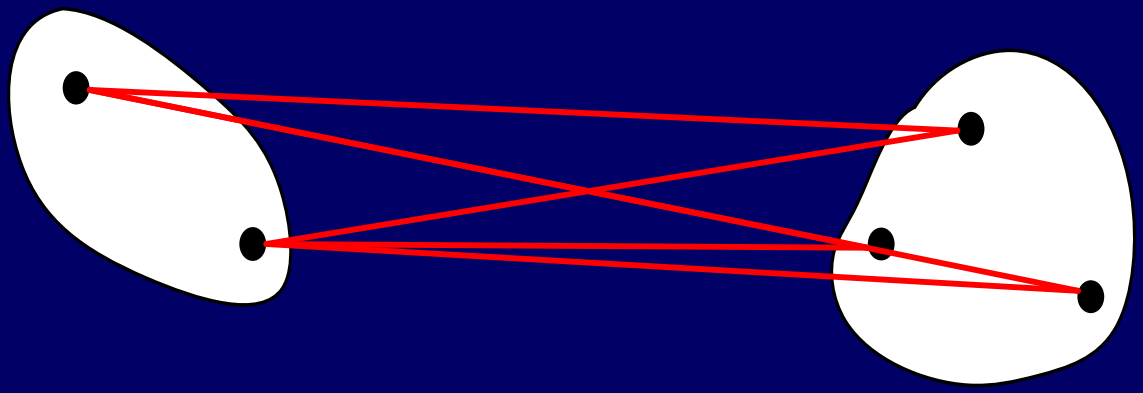
(4) 重心距离法

(5) 平均距离法

(6) 离差平方和法

$$D_{pq}^2 = \frac{1}{n_p n_q} \sum_{\substack{\vec{x}_i \in \omega_p \\ \vec{x}_j \in \omega_q}} d_{ij}^2$$

2.3.2 类间距离测度方法



平均距离法图示

2.3 类

2.3.2 类间距离

(1) 最近距离法

(2) 最远距离法

(3) 中间距离法

(4) 重心距离法

(5) 平均距离法

(6) 离差平方和法

类内离差平方和

$$s_t = \sum_{\vec{x}_i \in \omega_t} (\vec{x}_i - \vec{x}_t)'(\vec{x}_i - \vec{x}_t)$$

$$\omega_l = \omega_p \cup \omega_q \qquad D_{pq}^2 = s_l - s_p - s_q$$

$$D_{pq}^2 = \frac{n_p n_q}{n_p + n_q}(\vec{x}_p - \vec{x}_q)'(\vec{x}_p - \vec{x}_q)$$

$\vec{x}_t \ \vec{x}_p \vec{x}_q$ 分别为对应类的重心

递推公式为:

$$D_{kl}^2 = \frac{n_k + n_p}{n_k + n_l} D_{kp}^2 + \frac{n_k + n_q}{n_k + n_l} D_{kq}^2 - \frac{n_k}{n_k + n_l} D_{pq}^2$$

第二章　聚类分析

2.1　聚类分析的概念

2.2　模式相似性测度

2.3　类的定义与类间距离

2.4　准则函数

2.5　聚类的算法

2·4　准则函数

2.4.1 点与集合间的距离

第一类：对集合的分布没有先验知识时，可采用前面介绍的类间距离计算方法进行。

第二类：当知道集合的中点分布的先验知识时，可用相应的模型进行计算。

点模型、超平面模型、超球面模型

2.4.2 聚类的准则函数

判别分类结果好坏的一般标准：

类内距离小，类间距离大。

某些算法需要一个能对分类过程或分类结果的优劣进行评估的准则函数。如果聚类准则函数选择得好，聚类质量就会高。聚类准则往往是和类的定义有关的，是类的定义的某种体现。

2.4.2 聚类的准则函数

1、类内距离准则

设有待分类的模式集 $\{\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_N\}$ 在某种相似性测度基础上被划分为 c 类，$\{\vec{x}_i^{(j)};\ j=1,2,\cdots,c;\ i=1,2,\cdots,n_j\}$ 类内距离准则函数 J_W 定义为:($\vec{m}_j$ 表示 ω_j 类的模式均值矢量。)

$$J_W = \sum_{j=1}^{c} \sum_{i=1}^{n_j} \left\| \vec{x}_i^{(j)} - \vec{m}_j \right\|^2$$

2·4 准则函数

1、类内距离准则

$$J_W = \sum_{j=1}^{c} \sum_{i=1}^{n_j} \left\| \vec{x}_i^{(j)} - \vec{m}_j \right\|^2$$

我们的目标是使 J_W 取最小，即 $J_W \Rightarrow \min$ ，这种方法也称为误差平方和准则。

显然，J_W 是模式 $\vec{x}_i$ 和类心 $\vec{m}_j$ 的函数，在样本集 $\{\vec{x}_i\}$ 给定条件下，J_W 的值取决于类心的选取。

2·4 准则函数

加权类内距离准则 J_{WW}：

$$J_{WW}=\sum_{j=1}^{c}\frac{n_j}{N}\bar{d}_j^{\,2} \qquad \bar{d}_j^{\,2}=\frac{2}{n_j(n_j-1)}\sum_{\substack{\vec{x}_k^{(j)}\in\omega_j\\ \vec{x}_i^{(j)}\in\omega_j}}\left\|\vec{x}_i^{(j)}-\vec{x}_k^{(j)}\right\|^2$$

式中，$\sum\left\|\vec{x}_i^{(j)}-\vec{x}_k^{(j)}\right\|^2$表示$\omega_j$类内任两个模式距离平方和，共有 $\frac{n_j(n_j-1)}{2}$个组合数，所以$\bar{d}_j^{\,2}$表示类内两模式间的均方距离。*N*为待分类模式总数，$\frac{n_j}{N}$表示ω_j类先验概率的估计——频率。

2、类间距离准则

$$J_B = \sum_{j=1}^{c} (\vec{m}_j - \vec{m})'(\vec{m}_j - \vec{m}) \Rightarrow \max$$

这里，$\vec{m}_j$ 为 ω_j 类的模式平均矢量，$\vec{m}$ 为总的模式平均矢量。设 n_j 为 ω_j 类所含模式个数，

$$\vec{m}_j = \frac{1}{n_j} \sum_{\vec{x}_i \in \omega_j} \vec{x}_i \qquad \vec{m} = \frac{1}{N} \sum_{i=1}^{N} \vec{x}_i$$

2·4 准则函数

加权类间距离准则：

$$J_{WB}=\sum_{j=1}^{c}\frac{n_j}{N}(\vec{m}_j-\vec{m})'(\vec{m}_j-\vec{m})\Rightarrow \max$$

对于两类问题，类间距离有时取

$$J_{B2}=(\vec{m}_1-\vec{m}_2)'(\vec{m}_1-\vec{m}_2)$$

J_{B2}和 J_{WB}的关系是 $J_{WB}=\frac{n_1}{N}\frac{n_2}{N}J_{B2}$

3、基于类内距离类间距离的准则函数

我们希望聚类结果使类内距离越小越好，类间距离越大越好。为此构造能同时反映出类内距离和类间距离的准则函数。

设待分类模式集 $\{\vec{x}_i, i=1,2,\cdots,N\}$，将它们分成 c 类，ω_j 类含 n_j 个模式，分类后各模式记为

$$\left\{\vec{x}_i^{(j)}, j=1,2,\cdots,c; i=1,2,\cdots,n_j\right\}$$

3、基于类内距离类间距离的准则函数

ω_j 的类内离差阵定义为

$$S_W^{(j)} = \frac{1}{n_j}\sum_{i=1}^{n_j}(\vec{x}_i^{(j)} - \vec{m}_j)(\vec{x}_i^{(j)} - \vec{m}_j)' \qquad (j = 1,2,\cdots,c)$$

式中 $\vec{m}_j$ 为类 ω_j 的模式均值矢量

$$\vec{m}_j = \frac{1}{n_j}\sum_{i=1}^{n_j}\vec{x}_i^{(j)} \qquad (j = 1,2,\cdots,c)$$

2·4 准则函数

总的类内离差阵定义为：$S_W = \sum_{j=1}^{c} \frac{n_j}{N} S_W^{(j)}$

类间离差阵定义为：$S_B = \sum_{j=1}^{c} \frac{n_j}{N} (\vec{m}_j - \vec{m})(\vec{m}_j - \vec{m})'$

式中，$\vec{m}$ 为所有待分类模式的均值矢量：$\vec{m} = \frac{1}{N} \sum_{i=1}^{N} \vec{x}_i$

总的离差阵 S_T ，定义为：$S_T = \frac{1}{N} \sum_{i=1}^{N} (\vec{x}_i - \vec{m})(\vec{x}_i - \vec{m})'$

于是有下面关系 $S_T = S_W + S_B$

3、基于类内距离类间距离的准则函数

聚类的基本目的是使 $\mathrm{Tr}[S_B] \Rightarrow \max$ 或 $\mathrm{Tr}[S_W] \Rightarrow \min$ 。利用线形代数有关矩阵的迹和行列式的性质，可以定义如下4个聚类的准则函数：

$$J_1 = \mathrm{Tr}[S_W^{-1} S_B] \qquad J_2 = \left| S_W^{-1} S_B \right|$$

$$J_3 = \mathrm{Tr}[S_W^{-1} S_T] \qquad J_4 = \left| S_W^{-1} S_T \right|$$

3、基于类内距离类间距离的准则函数

$$J_1 = \mathrm{Tr}[S_W^{-1} S_B] \qquad J_2 = \left| S_W^{-1} S_B \right|$$

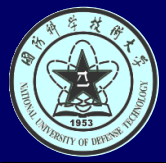
$$J_3 = \mathrm{Tr}[S_W^{-1} S_T] \qquad J_4 = \left| S_W^{-1} S_T \right|$$

由它们的构造可以看出，为得到好的聚类结果，应该使它们尽量的大。

作业

1、设有样本集S={$x_1,x_2,\ldots,x_n$},证明类心z到S中各样本点距离平方和 $\sum_{i=1}^{n}(\vec{x}_i-\vec{z})^T(\vec{x}_i-\vec{z})$ 为最小时，有 $\vec{z}=\frac{1}{n}\sum_{i=1}^{n}\vec{x}_i$

2、P64: 2.9



谢 谢！