
第2章 聚类分析

第2章 聚类分析

2.1 距离聚类的概念

2.2 相似性测度和聚类准则

2.3 基于距离阈值的聚类算法

2.4 层次聚类法

2.5 动态聚类法

2.6 聚类结果的评价

2.1 距离聚类的概念

1. 概念：“物以类聚”

聚类分析：根据模式之间的相似性对模式进行分类，是一种非监督分类方法。

2. 相似性的含义

有n个特征则组成n维向量 $\boldsymbol{X}=[x_1, x_2, \ldots, x_n,]^{\mathrm{T}}$称为该样本的特征向量。它相当于特征空间中的一个点，以特征空间中，点间的距离函数作为模式相似性的测量，以“距离”作为模式分类的依据，距离越小，越“相似”。

注意：聚类分析是否有效，与模式特征向量的分布形式有很大关系。选取的特征向量是否合适非常关键。

2.2 相似性测度和聚类准则

2.2.1 相似性测度

相似性测度：衡量模式之间相似性的一种尺度。如：距离。

复习：已知向量 $\boldsymbol{Y}=\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$，则：

$$\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}=\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}\begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix}=\begin{bmatrix} y_1^2 & y_1y_2 & y_1y_3 \\ y_2y_1 & y_2^2 & y_2y_3 \\ y_3y_1 & y_3y_2 & y_3^2 \end{bmatrix}$$

$$\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}=\begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}=y_1^2+y_2^2+y_3^2=\|\boldsymbol{Y}\|^2$$

1. 欧氏距离（Euclid，欧几里德）——简称距离

设X_1、X_2为两个n维模式样本，

$$X_1=[x_{11},x_{12},\ldots,x_{1n}]^{\mathrm{T}} \qquad X_2=[x_{21},x_{22},\ldots,x_{2n}]^{\mathrm{T}}$$

欧氏距离定义为：

$$D(X_1,X_2)=\|X_1-X_2\|=\sqrt{(X_1-X_2)^{\mathrm{T}}(X_1-X_2)}$$

$$=\sqrt{(x_{11}-x_{21})^2+\cdots\cdots+(x_{1n}-x_{2n})^2}$$

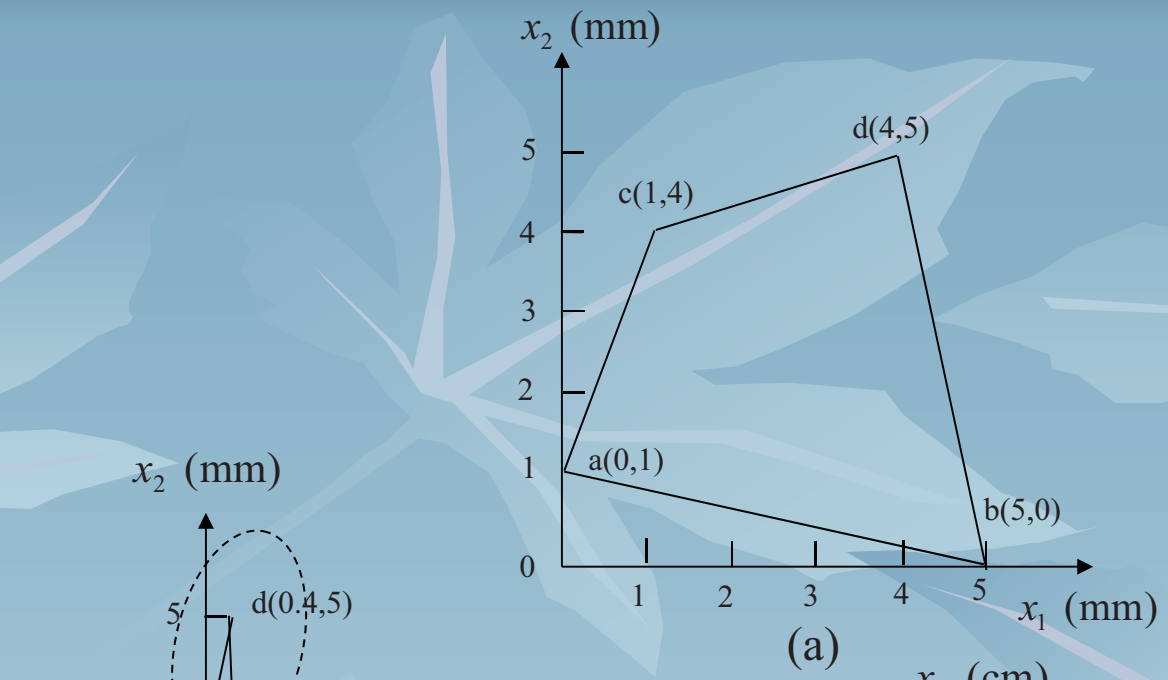
距离越小，越相似。 (D_Distance)

注意：

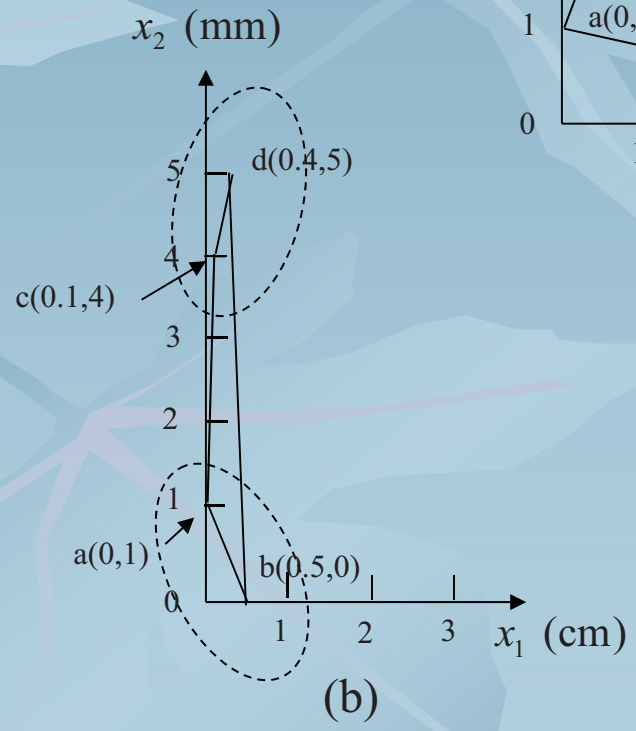
1）各特征向量对应的维上应当是相同的物理量；

注意物理量的单位。

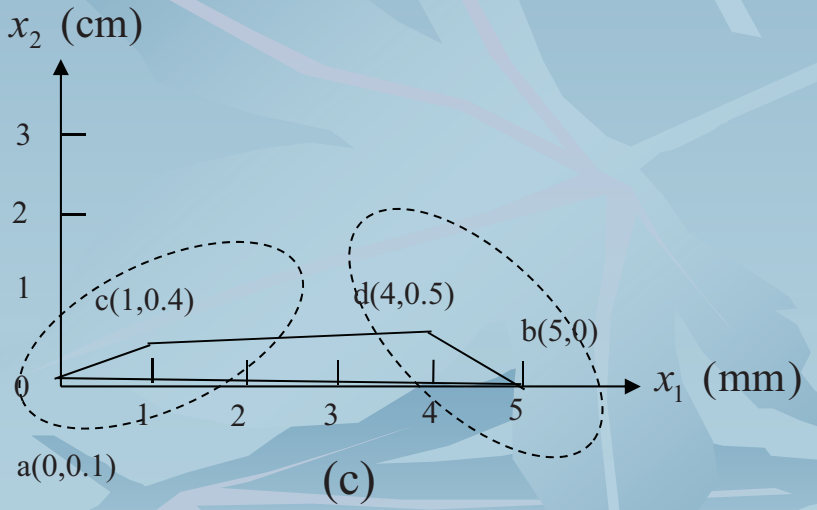
某些维上物理量采用的单位发生变化，会导致对同样的点集出现不同聚类结果的现象。



(a)



(b)



(c)

2）解决方法：使特征数据标准化，使其与变量的单位无关。

2. 马氏距离(Maharanobis)

平方表达式：$D^2 = (\boldsymbol{X}_1 - \boldsymbol{X}_2)^{\mathrm{T}} \boldsymbol{C}^{-1} (\boldsymbol{X}_1 - \boldsymbol{X}_2)$

令：

$\boldsymbol{M}$：均值向量；（M_Mean）

$\boldsymbol{C}$：该类模式总体的协方差矩阵。（C_covariance）

对n维向量：$\boldsymbol{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$，$\boldsymbol{M} = \begin{bmatrix} m_1 \\ \vdots \\ m_n \end{bmatrix}$

$$\boldsymbol{C} = E\left\{(\boldsymbol{X} - \boldsymbol{M})((\boldsymbol{X} - \boldsymbol{M}))^{\mathrm{T}}\right\}$$

$$= E\left\{\begin{bmatrix} (x_1 - m_1) \\ (x_2 - m_2) \\ \vdots \\ (x_n - m_n) \end{bmatrix} \begin{bmatrix} (x_1 - m_1) & (x_2 - m_2) & \cdots & (x_n - m_n) \end{bmatrix}\right\}$$

$$
=\begin{bmatrix}
E(x_1-m_1)(x_1-m_1) & E(x_1-m_1)(x_2-m_2) & \cdots & E(x_1-m_1)(x_n-m_n) \\
E(x_2-m_2)(x_1-m_1) & E(x_2-m_2)(x_2-m_2) & \cdots & \cdots \\
\vdots & \vdots & \vdots & \vdots \\
E(x_n-m_n)(x_1-m_1) & \cdots & \cdots & E(x_n-m_n)(x_n-m_n)
\end{bmatrix}
$$

$$
=\begin{bmatrix}
\sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1n}^2 \\
\sigma_{21}^2 & \ddots & \sigma_{jk}^2 & \vdots \\
\vdots & \vdots & \sigma_{kk}^2 & \vdots \\
\sigma_{n1}^2 & \cdots & \cdots & \sigma_{nn}^2
\end{bmatrix}
$$

表示的概念是各分量上模式样本到均值的距离，也就是在各维上模式的分散情况。σ_{jk}^2 越大，离均值越远。

当$\boldsymbol{C}=\boldsymbol{I}$时，马氏距离为欧氏距离。

马氏距离特点：

1）马氏距离的计算是建立在总体样本的基础上。同样的两个样本，放入两个不同的总体中，最后计算得出的两个样本间的马氏距离通常是不相同的；

2）不受量纲的影响，两点之间的马氏距离与原始数据的测量单位无关；

3) 马氏距离可以排除变量之间的相关性的干扰。

4) 夸大了变化微小的变量的作用。

3. 明氏距离(Minkowaki)

n维模式样本向量$\boldsymbol{X}_i$、$\boldsymbol{X}_j$间的明氏距离表示为：

$$D_m(\boldsymbol{X}_i,\boldsymbol{X}_j)=\left[\sum_{k=1}^{n}\left|x_{ik}-x_{jk}\right|^m\right]^{1/m}$$

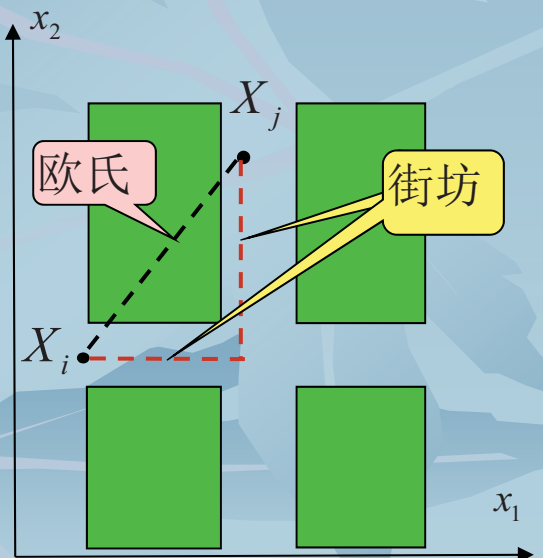
式中，x_{ik}、x_{jk}分别表示$\boldsymbol{X}_i$和$\boldsymbol{X}_j$的第k个分量。

当m=2时，明氏距离为欧氏距离。

当m=1时：

$$D_1(\boldsymbol{X}_i,\boldsymbol{X}_j)=\sum_{k=1}^{n}\left|x_{ik}-x_{jk}\right|$$

称为“街坊”距离 (“City block”distance)。



当k=2时：图示

$$D_1(\boldsymbol{X}_i,\boldsymbol{X}_j)=\left|x_{i1}-x_{j1}\right|+\left|x_{i2}-x_{j2}\right|$$

4．汉明(Hamming)距离

设$\boldsymbol{X}_i$、$\boldsymbol{X}_j$为n维二值（1或－1）模式样本向量，则

汉明距离：$$D_h(\boldsymbol{X}_i,\boldsymbol{X}_j)=\frac{1}{2}\left(n-\sum_{k=1}^{n}x_{ik}\cdot x_{jk}\right)$$

式中，x_{ik}、x_{jk}分别表示$\boldsymbol{X}_i$和$\boldsymbol{X}_j$的第k个分量。

两个模式向量的各分量取值均不同：$D_h(\boldsymbol{X}_i,\boldsymbol{X}_j)=n$；

全相同：$D_h(\boldsymbol{X}_i,\boldsymbol{X}_j)=0$

5．角度相似性函数

$$S(\boldsymbol{X}_i,\boldsymbol{X}_j)=\frac{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_j}{\|\boldsymbol{X}_i\|\cdot\|\boldsymbol{X}_j\|}$$

是模式向量$\boldsymbol{X}_i$，$\boldsymbol{X}_j$之间夹角的余弦。

6. Tanimoto测度

用于0，1二值特征的情况，

$$S(\boldsymbol{X}_i,\boldsymbol{X}_j)=\frac{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_j}{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i+\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j-\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_j}$$

$$=\frac{\boldsymbol{X}_i,\boldsymbol{X}_j\text{中共有的特征数目}}{\boldsymbol{X}_i\text{和}\boldsymbol{X}_j\text{中占有的特征数目的总数}}$$

相似性测度函数的共同点都涉及到把两个相比较的向量$\boldsymbol{X}_i$，$\boldsymbol{X}_j$的分量值组合起来，但怎样组合并无普遍有效的方法，对具体的模式分类，需视情况作适当选择。

2.2.2 聚类准则

聚类准则：根据相似性测度确定的，衡量模式之间是否相似的标准。即把不同模式聚为一类还是归为不同类的准则。

确定聚类准则的两种方式：

1. 阈值准则：根据规定的距离阈值进行分类的准则。
2. 函数准则：利用聚类准则函数进行分类的准则。

聚类准则函数：在聚类分析中，表示模式类间相似或差异性的函数。

它应是模式样本集$\{X\}$和模式类别$\{S_j, \mathrm{j}=1,2,\cdots\cdots,c\}$的函数。可使聚类分析转化为寻找准则函数极值的最优化问题。一种常用的指标是误差平方之和。

聚类准则函数：$$J=\sum_{j=1}^{c}\sum_{X\in S_j}\left\|X-M_j\right\|^2$$

式中：c为聚类类别的数目，

$M_j=\frac{1}{N_j}\sum_{X\in S_j}X$为属于 S_j 集的样本的均值向量，

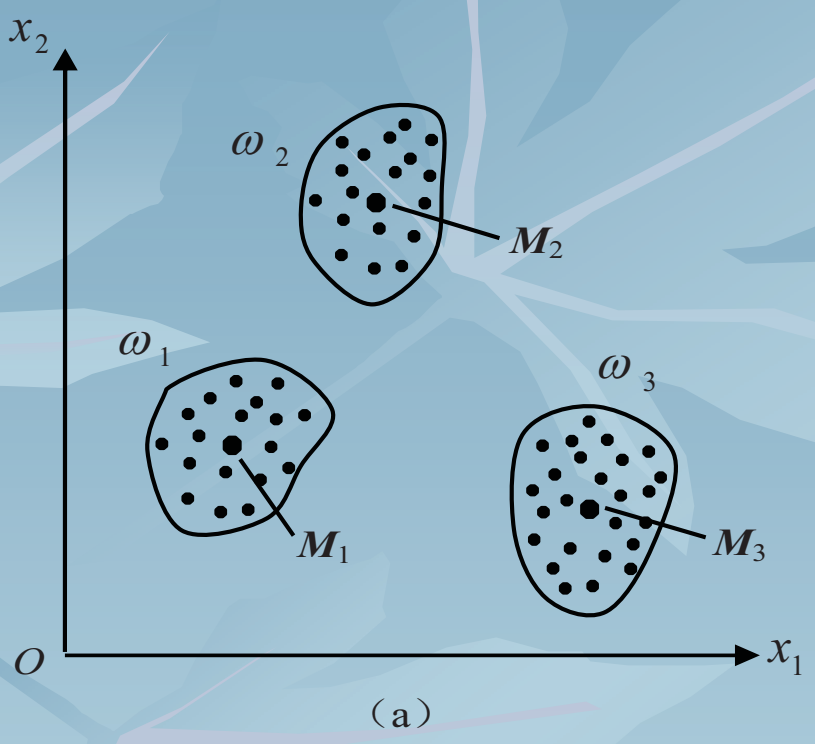
N_j 为 S_j中样本数目。

J代表了分属于c个聚类类别的全部模式样本与其相应类别模式均值之间的误差平方和。

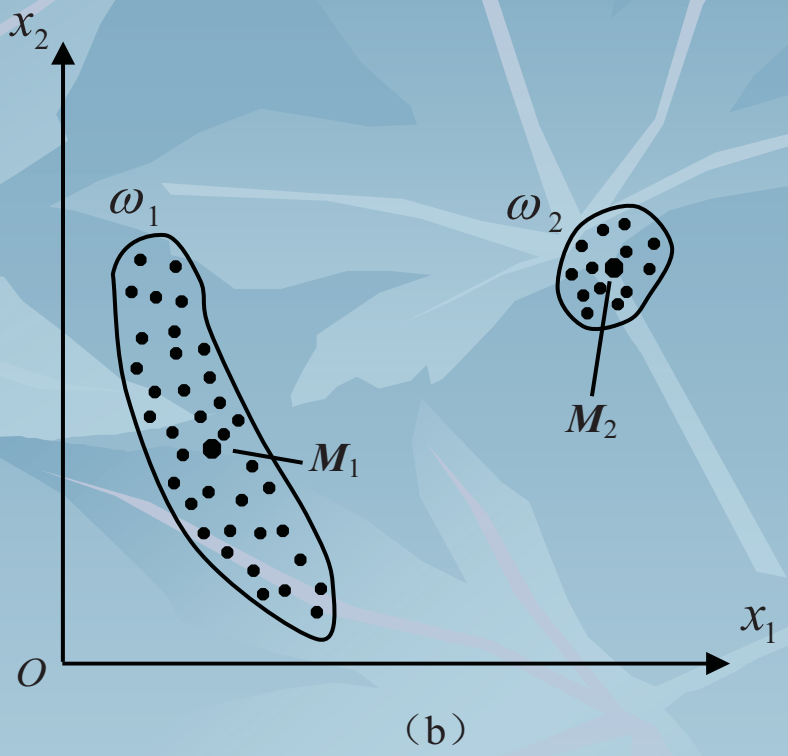
适用范围：

适用于各类样本密集且数目相差不多，而不同类间的样本又明显分开的情况。

例1：



（a）



（b）

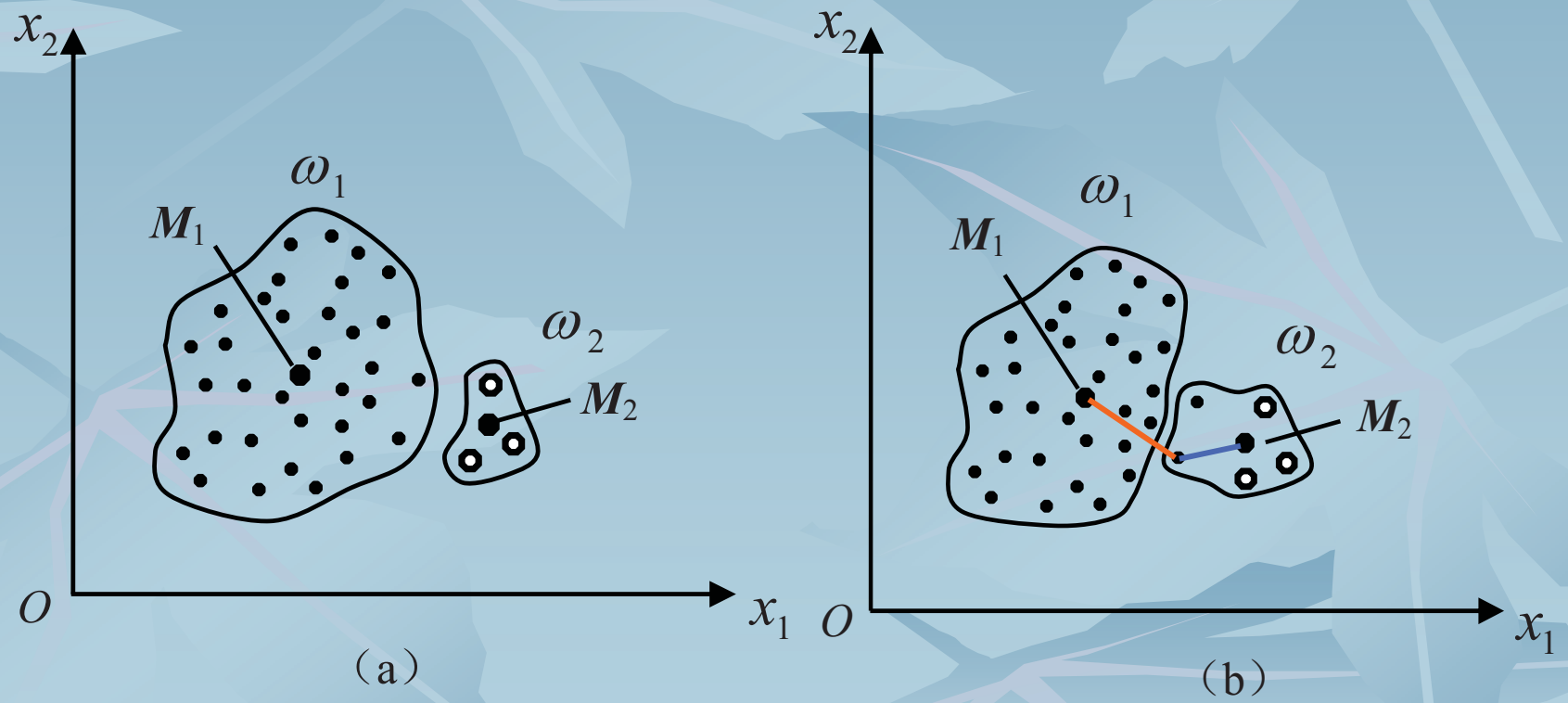
类内误差平方和很小，类间距离很远。可得到最好的结果。

ω_1类长轴两端距离中心很远，J值较大，结果不易令人满意。

例2：另一种情况

有时可能把样本数目多的一类分拆为二，造成错误聚类。

原因：这样分开，J值会更小。



2.3 基于距离阈值的聚类算法

2.3.1 近邻聚类法

1. 问题： 有N个待分类的模式 $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_N\}$，要求按距离阈值$T$分类到以 $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \cdots$ 为聚类中心的模式类中。

(T_threshold)

2. 算法描述

① 任取样本$\boldsymbol{X}_i$作为第一个聚类中心的初始值，如令$\boldsymbol{Z}_1 = \boldsymbol{X}_1$。

② 计算样本$\boldsymbol{X}_2$ 到$\boldsymbol{Z}_1$ 的欧氏距离 $D_{21} = \|\boldsymbol{X}_2 - \boldsymbol{Z}_1\|$，

若$D_{21} > T$，定义一新的聚类中心$\boldsymbol{Z}_2 = \boldsymbol{X}_2$；

否则 $\boldsymbol{X}_2 \in$ 以$\boldsymbol{Z}_1$为中心的聚类。

③ 假设已有聚类中心$\boldsymbol{Z}_1$、$\boldsymbol{Z}_2$，计算$D_{31}=\|\boldsymbol{X}_3-\boldsymbol{Z}_1\|$和$D_{32}=\|\boldsymbol{X}_3-\boldsymbol{Z}_2\|$，

若$D_{31}>T$且$D_{32}>T$，则建立第三个聚类中心$\boldsymbol{Z}_3=\boldsymbol{X}_3$；

否则$\boldsymbol{X}_3\in$离$\boldsymbol{Z}_1$和$\boldsymbol{Z}_2$中最近者（最近邻的聚类中心）。

……依此类推，直到将所有的N个样本都进行分类。

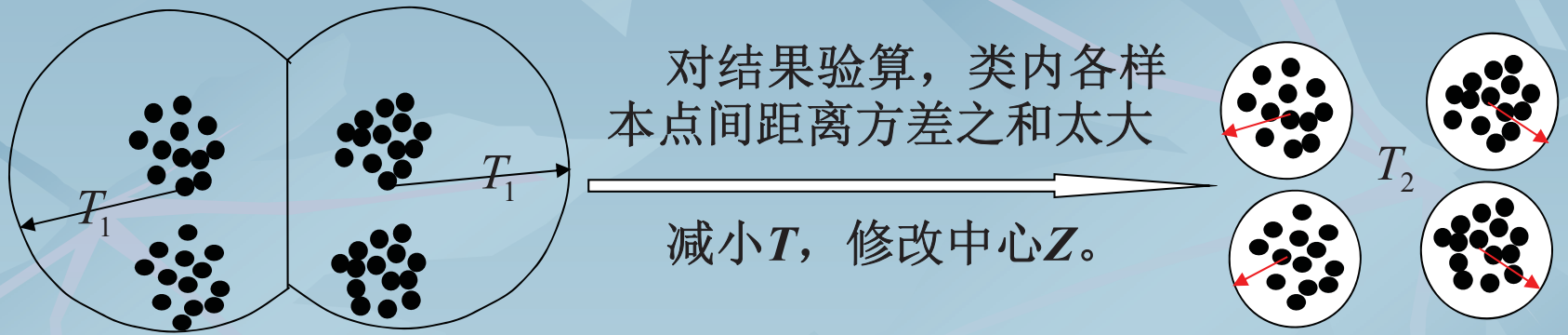
3. 算法特点

1）局限性：很大程度上依赖于第一个聚类中心的位置选择、待分类模式样本的排列次序、距离阈值T的大小以及样本分布的几何性质等。

2）优点：计算简单。（一种虽粗糙但快速的方法）

4．算法讨论

用先验知识指导阈值T和起始点Z_1的选择，可获得合理的聚类结果。否则只能选择不同的初值重复试探，并对聚类结果进行验算，根据一定的评价标准，得出合理的聚类结果。



2.3.2 最大最小距离算法（小中取大距离算法）

1. 问题： 已知N个待分类的模式$\{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_N\}$，

分类到聚类中心$\boldsymbol{Z}_1, \boldsymbol{Z}_2, \cdots$对应的类别中 。

2. 算法描述

① 选任意一模式样本做为第一聚类中心$\boldsymbol{Z}_1$。

② 选择离$\boldsymbol{Z}_1$距离最远的样本作为第二聚类中心$\boldsymbol{Z}_2$。

③ 逐个计算各模式样本与已确定的所有聚类中心之间的距离，并选出其中的最小距离。例当聚类中心数k=2时，计算

$$D_{i1} = \|\boldsymbol{X}_i - \boldsymbol{Z}_1\| \qquad D_{i2} = \|\boldsymbol{X}_i - \boldsymbol{Z}_2\|$$

$\min(D_{i1}, D_{i2})$，i=1，…，N（N个最小距离）

④ 在所有最小距离中选出最大距离，如该最大值达到$\|\boldsymbol{Z}_1-\boldsymbol{Z}_2\|$的一定分数比值(阈值$T$) 以上，则相应的样本点取为新的聚类中心，返回③；否则，寻找聚类中心的工作结束。

例k=2时

若$\max\{\min(D_{i1},D_{i2}),i=1,2,...,N\}>\theta\|\boldsymbol{Z}_1-\boldsymbol{Z}_2\|,\ 0<\theta<1$

则$\boldsymbol{Z}_3$存在。（θ：用试探法取为一固定分数，如1/2。）

⑤ 重复步骤③④，直到没有新的聚类中心出现为止。

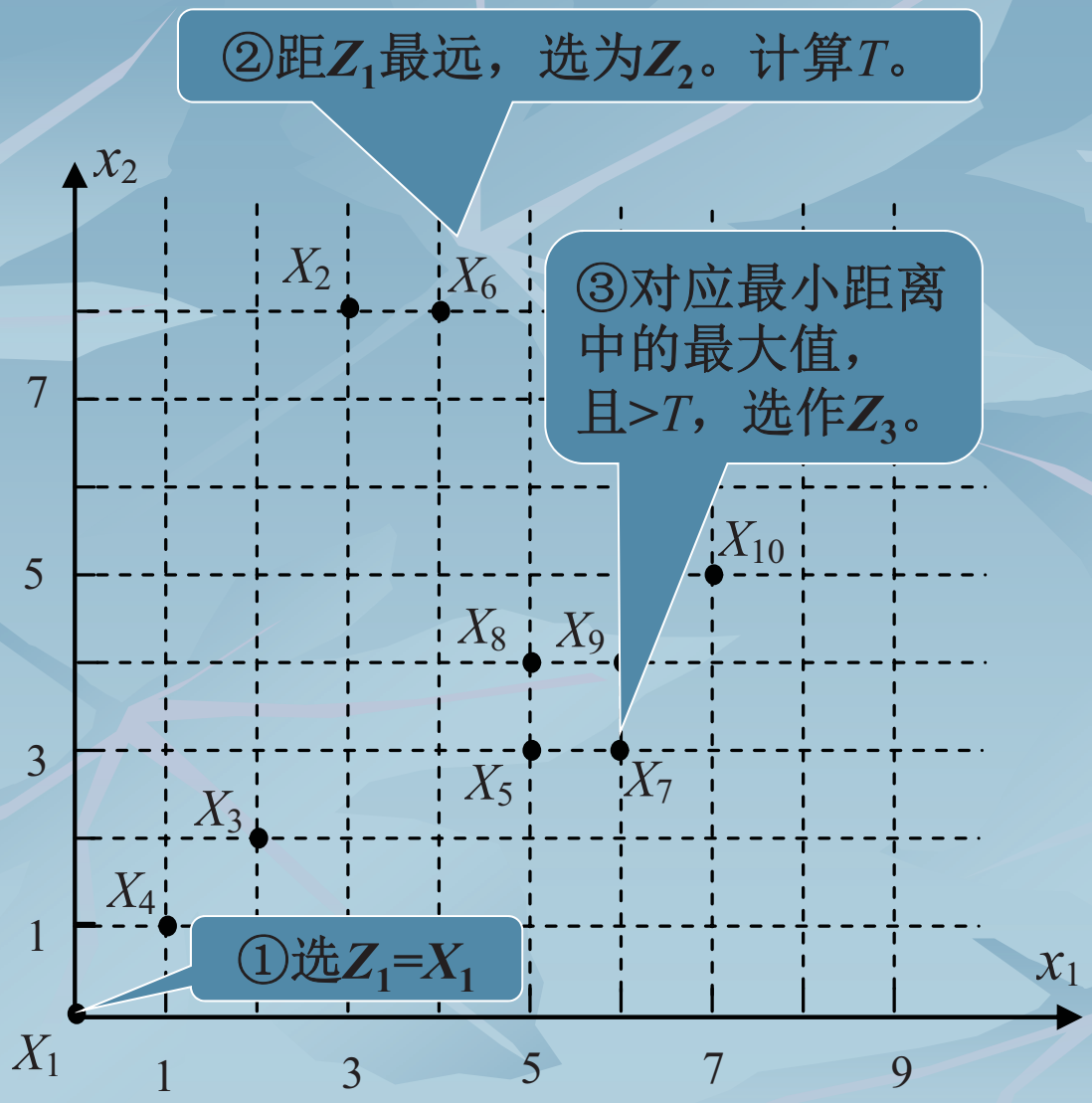
⑥ 将样本$\{\boldsymbol{X}_i,i=1,2,\cdots,N\}$按最近距离划分到相应聚类中心对应的类别中。

思路总结：

先找中心后分类；关键：怎样开新类，聚类中心如何定。

为使聚类中心更有代表性，可取各类的样本均值作为聚类中心。

例2.1 对图示模式样本用最大最小距离算法进行聚类分析。



③$T=\frac{1}{2}\|\boldsymbol{Z}_1-\boldsymbol{Z}_2\|=\frac{1}{2}\sqrt{80}$

10个最小距离中，$\boldsymbol{X}_7$对应的距离>T,

$$\therefore \boldsymbol{Z}_3=\boldsymbol{X}_7$$

④ 用全体模式对三个聚类中心计算最小距离中的最大值，无>T 情况，停止寻找中心。

结果：$\boldsymbol{Z}_1=\boldsymbol{X}_1$；$\boldsymbol{Z}_2=\boldsymbol{X}_6$；

$\boldsymbol{Z}_3=\boldsymbol{X}_7$。

⑤ 聚类

2.4 层次聚类法

（Hierarchical Clustering Method）

（系统聚类法、分级聚类法）

思路：每个样本先自成一类，

然后按距离准则逐步合并，减少类数。

1. 算法描述

1）N个初始模式样本自成一类，即建立N类：

$$G_1(0), G_2(0), \cdots, G_N(0)$$

(G_Group)

计算各类之间（即各样本间）的距离，得一$N \times N$维距离矩阵$\boldsymbol{D}(0)$。“0”表示初始状态。

2）假设已求得距离矩阵$\boldsymbol{D}(n)$（n为逐次聚类合并的次数），找出$\boldsymbol{D}(n)$中的最小元素，将其对应的两类合并为一类。由此建立新的分类：

$$G_1(n+1), G_2(n+1), \cdots$$

3）计算合并后新类别之间的距离，得$\boldsymbol{D}(n+1)$。

4）跳至第2步，重复计算及合并。

结束条件：

1）取距离阈值T，当$\boldsymbol{D}(n)$的最小分量超过给定值 T 时，算法停止。所得即为聚类结果。

2）或不设阈值T，一直将全部样本聚成一类为止，输出聚类的分级树。

2. 问题讨论：类间距离计算准则

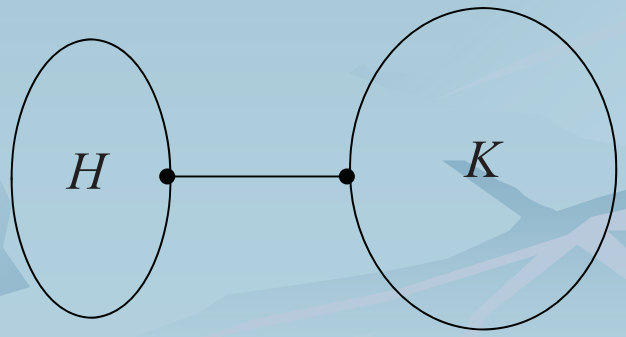
1）最短距离法

如H、K是两个聚类，则两类间的最短距离定义为：

$$D_{HK}=\min\{D(\boldsymbol{X}_H,\boldsymbol{X}_K)\}\quad \boldsymbol{X}_H\in H,\boldsymbol{X}_K\in K$$

$D(\boldsymbol{X}_H,\boldsymbol{X}_K)$：$H$类中的某个样本$\boldsymbol{X}_H$和$K$类中的某个样本$\boldsymbol{X}_K$之间的欧氏距离。

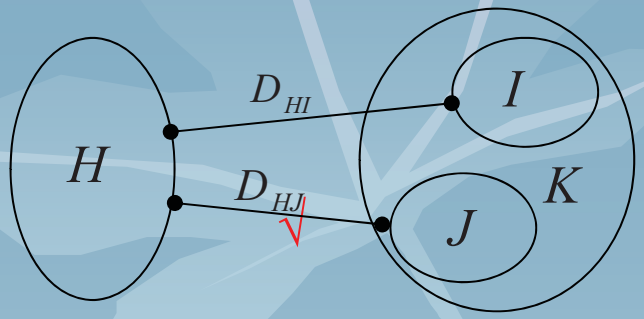
D_{HK}：H类中所有样本与K类中所有样本之间的最小距离。



如果K类由I和J两类合并而成，则

$$D_{HI}=\min\{D(X_H,X_I)\}\quad X_H\in H,X_I\in I$$

$$D_{HJ}=\min\{D(X_H,X_J)\}\quad X_H\in H,X_J\in J$$



得到递推公式：　$D_{HK}=\min\{D_{HI},D_{HJ}\}$

2）最长距离法

$$D_{HK}=\max\{D(X_H,X_K)\}\quad X_H\in H,X_K\in K$$

若K类由I、J两类合并而成，则

$$D_{HI}=\max\{D(X_H,X_I)\}\quad X_H\in H,X_I\in I$$

$$D_{HJ}=\max\{D(X_H,X_J)\}\quad X_H\in H,X_J\in J$$

有：　$D_{HK}=\max\{D_{HI},D_{HJ}\}$

3）中间距离法

介于最长与最短的距离之间。如果K类由I类和J类合并而成，则H和K类之间的距离为

$$D_{HK}=\sqrt{\frac{1}{2}D_{HI}^2+\frac{1}{2}D_{HJ}^2-\frac{1}{4}D_{IJ}^2}$$

4）重心法

将每类中包含的样本数考虑进去。若I类中有n_I个样本，J类中有n_J个样本，则类与类之间的距离递推式为

$$D_{HK}=\sqrt{\frac{n_I}{n_I+n_J}D_{HI}^2+\frac{n_J}{n_I+n_J}D_{HJ}^2-\frac{n_In_J}{(n_I+n_J)^2}D_{IJ}^2}$$

5）类平均距离法

$$D_{HK}=\sqrt{\frac{1}{n_H n_K}\sum_{\substack{i\in H\\ j\in K}} d_{ij}^2}$$

d_{ij}^2：H类任一样本$\boldsymbol{X}_i$和K类任一样本$\boldsymbol{X}_j$之间的欧氏距离平方。

若K类由I类和J类合并产生，则递推式为

$$D_{HK}=\sqrt{\frac{n_I}{n_I+n_J}D_{HI}^2+\frac{n_J}{n_I+n_J}D_{HJ}^2}$$

定义类间距离的方法不同，分类结果会不太一致。实际问题中常用几种不同的方法，比较分类结果，从而选择一个比较切合实际的分类。

例：给出6个五维模式样本如下，按最短距离准则进行系统聚类分类。

$$\boldsymbol{X}_1=[0,3,1,2,0]^{\mathrm{T}} \quad \boldsymbol{X}_2=[1,3,0,1,0]^{\mathrm{T}} \quad \boldsymbol{X}_3=[3,3,0,0,1]^{\mathrm{T}}$$

$$\boldsymbol{X}_4=[1,1,0,2,0]^{\mathrm{T}} \quad \boldsymbol{X}_5=[3,2,1,2,1]^{\mathrm{T}} \quad \boldsymbol{X}_6=[4,1,1,1,0]^{\mathrm{T}}$$

解：（1）将每一样本看作单独一类，得：

$$G_1(0)=\{\boldsymbol{X}_1\} \quad G_2(0)=\{\boldsymbol{X}_2\} \quad G_3(0)=\{\boldsymbol{X}_3\}$$

$$G_4(0)=\{\boldsymbol{X}_4\} \quad G_5(0)=\{\boldsymbol{X}_5\} \quad G_6(0)=\{\boldsymbol{X}_6\}$$

计算各类间欧氏距离：

$$D_{12}(0)=\|\boldsymbol{X}_1-\boldsymbol{X}_2\|=\left[(x_{11}-x_{21})^2+(x_{12}-x_{22})^2+(x_{13}-x_{23})^2+(x_{14}-x_{24})^2+(x_{15}-x_{25})^2\right]^{1/2}$$

$$=[1+0+1+1+0]^{1/2}=\sqrt{3}$$

$$D_{13}(0)=\left[3^2+0+1+2^2+1\right]^{1/2}=\sqrt{15}\ ,\ D_{14}(0)\ ,\ D_{15}(0)\ ,\ D_{16}(0)\ ;$$

$$D_{23}(0)\ D_{24}(0)\ D_{25}(0)\ D_{26}(0)\ ;\quad D_{34}(0)\ D_{35}(0)\ D_{36}(0)\ \cdots$$

得距离矩阵$\boldsymbol{D}(0)$:

$\boldsymbol{D}(0)$	$G_1(0)$	$G_2(0)$	$G_3(0)$	$G_4(0)$	$G_5(0)$	$G_6(0)$
$G_1(0)$	0					
$G_2(0)$	$*\sqrt{3}$	0				
$G_3(0)$	$\sqrt{15}$	$\sqrt{6}$	0			
$G_4(0)$	$\sqrt{6}$	$\sqrt{5}$	$\sqrt{13}$	0		
$G_5(0)$	$\sqrt{11}$	$\sqrt{8}$	$\sqrt{6}$	$\sqrt{7}$	0	
$G_6(0)$	$\sqrt{21}$	$\sqrt{14}$	$\sqrt{8}$	$\sqrt{11}$	$\sqrt{4}$	0

（2）将最小距离$\sqrt{3}$对应的类$G_1(0)$ 和$G_2(0)$ 合并为1类，得新的分类。

$$G_{12}(1)=\{G_1(0),G_2(0)\}$$

$$G_3(1)=\{G_3(0)\} \qquad G_4(1)=\{G_4(0)\}$$

$$G_5(1)=\{G_5(0)\} \qquad G_6(1)=\{G_6(0)\}$$

计算聚类后的距离矩阵$\boldsymbol{D}(1)$:

由$\boldsymbol{D}(0)$ 递推出$\boldsymbol{D}(1)$ 。

$\boldsymbol{D}(0)$	$G_1(0)$	$G_2(0)$	$G_3(0)$	$G_4(0)$	$G_5(0)$	$G_6(0)$
$G_1(0)$	0					
$G_2(0)$	*$\sqrt{3}$	0				
$G_3(0)$	$\sqrt{15}$	$\sqrt{6}$	0			
$G_4(0)$	$\sqrt{6}$	$\sqrt{5}$	$\sqrt{13}$	0		
$G_5(0)$	$\sqrt{11}$	$\sqrt{8}$	$\sqrt{6}$	$\sqrt{7}$	0	
$G_6(0)$	$\sqrt{21}$	$\sqrt{14}$	$\sqrt{8}$	$\sqrt{11}$	$\sqrt{4}$	0

$\boldsymbol{D}(1)$	$G_{12}(1)$	$G_3(1)$	$G_4(1)$	$G_5(1)$	$G_6(1)$
$G_{12}(1)$	0				
$G_3(1)$	$\sqrt{6}$	0			
$G_4(1)$	$\sqrt{5}$	$\sqrt{13}$	0		
$G_5(1)$	$\sqrt{8}$	$\sqrt{6}$	$\sqrt{7}$	0	
$G_6(1)$	$\sqrt{14}$	$\sqrt{8}$	$\sqrt{11}$	*$\sqrt{4}$	0

（3）将$\boldsymbol{D}(1)$中最小值$\sqrt{4}$对应的类合为一类，得$\boldsymbol{D}(2)$。

$\boldsymbol{D}(2)$	$G_{12}(2)$	$G_3(2)$	$G_4(2)$	$G_{56}(2)$
$G_{12}(2)$	0			
$G_3(2)$	$\sqrt{6}$	0		
$G_4(2)$	$\sqrt{5}$	$\sqrt{13}$	0	
$G_{56}(2)$	$\sqrt{8}$	$\sqrt{6}$	$\sqrt{7}$	0

（4）将$\boldsymbol{D}(2)$中最小值$\sqrt{5}$对应的类合为一类，得$\boldsymbol{D}(3)$。

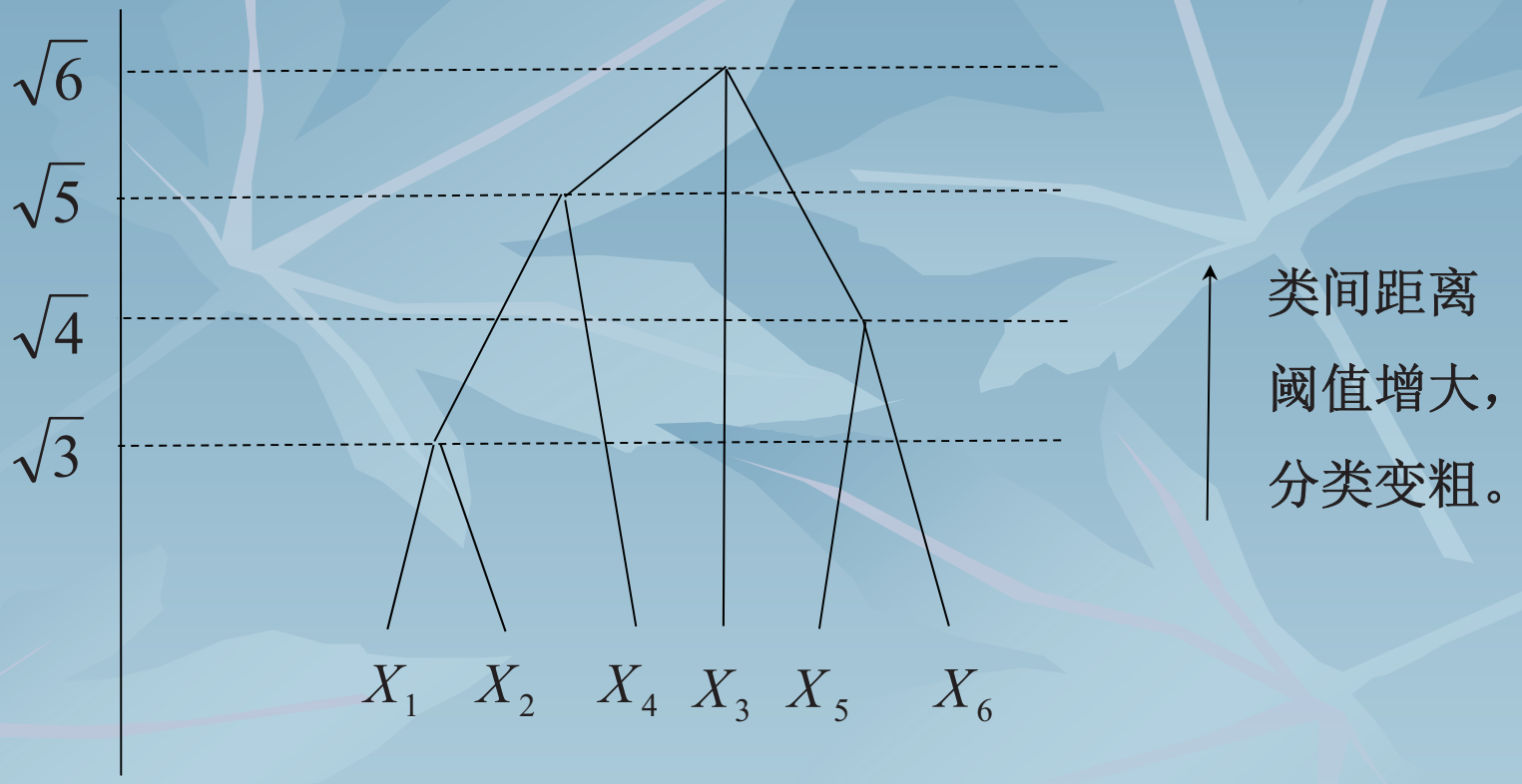
$\boldsymbol{D}(2)$	$G_{12}(2)$	$G_3(2)$	$G_4(2)$	$G_{56}(2)$
$G_{12}(2)$	0			
$G_3(2)$	$\sqrt{6}$	0	$\sqrt{13}$	
$G_4(2)$	$*\sqrt{5}$	$\sqrt{13}$	0	
$G_{56}(2)$	$\sqrt{8}$	$\sqrt{6}$	$\sqrt{7}$	0

$\boldsymbol{D}(3)$	$G_{124}(3)$	$G_3(3)$	$G_{56}(3)$
$G_{124}(3)$	0		
$G_3(3)$	$\sqrt{6}$	0	
$G_{56}(3)$	$\sqrt{7}$	$\sqrt{6}$	0

若给定的阈值为$T=\sqrt{5}$，$\boldsymbol{D}(3)$中的最小元素$\sqrt{6}>T$，聚类结束。

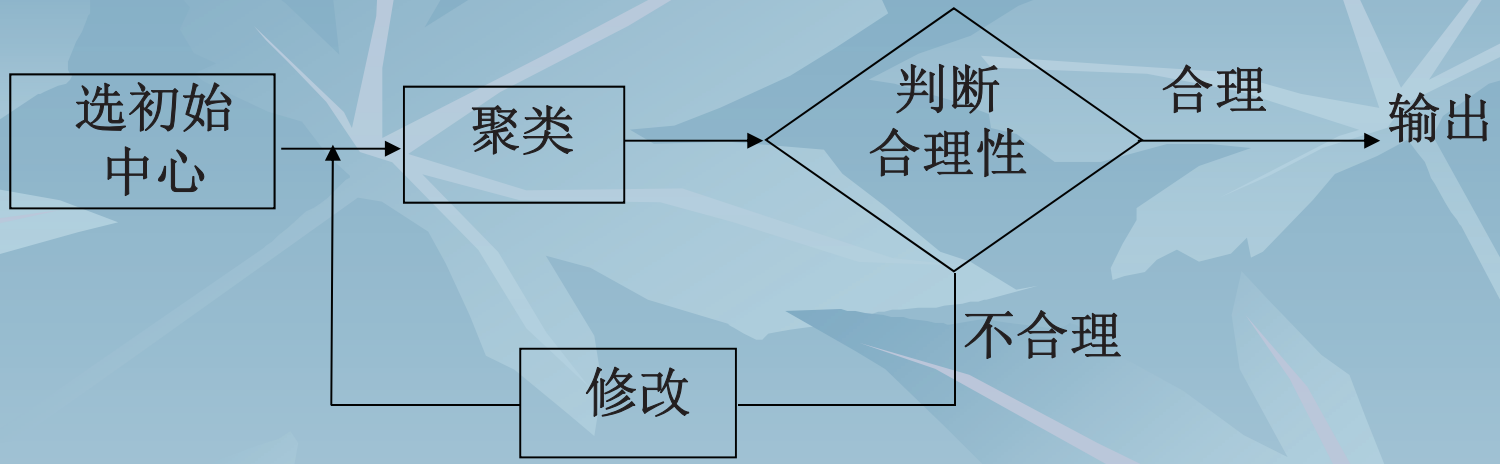
$$G_1=\{X_1,X_2,X_4\}\qquad G_2=\{X_3\}\qquad G_3=\{X_5,X_6\}$$

若无阈值，继续分下去，最终全部样本归为一类。可给出聚类过程的树状表示图。



层次聚类法的树状表示

2.5 动态聚类法



两种常用算法：

* K-均值算法(或C-均值算法)
* 迭代自组织的数据分析算法(ISODATA, iterative self-organizing data analysis techniques algorithm)

2.5.1 K-均值算法

基于使聚类准则函数最小化，

准则函数：聚类集中每一样本点到该类中心的距离平方和。

对于第j个聚类集，准则函数定义为

$$J_j = \sum_{i=1}^{N_j} \| \boldsymbol{X}_i - \boldsymbol{Z}_j \|^2, \quad \boldsymbol{X}_i \in S_j$$

S_j：第j个聚类集（域），聚类中心为$\boldsymbol{Z}_j$；

N_j：第j个聚类集S_j中所包含的样本个数。

对所有K个模式类有

$$J = \sum_{j=1}^{K} \sum_{i=1}^{N_j} \| \boldsymbol{X}_i - \boldsymbol{Z}_j \|^2, \quad \boldsymbol{X}_i \in S_j$$

K-均值算法的聚类准则：聚类中心的选择应使准则函数J极小，即使J_j的值极小。

应有 $\dfrac{\partial J_j}{\partial \boldsymbol{Z}_j} = 0$

即 $$\frac{\partial}{\partial \boldsymbol{Z}_j}\sum_{i=1}^{N_j}\| \boldsymbol{X}_i - \boldsymbol{Z}_j \|^2 = \frac{\partial}{\partial Z_j}\sum_{i=1}^{N_j}(\boldsymbol{X}_i - \boldsymbol{Z}_j)^{\mathrm{T}}(\boldsymbol{X}_i - \boldsymbol{Z}_j) = 0$$

可解得 $$\boldsymbol{Z}_j = \frac{1}{N_j}\sum_{i=1}^{N_j}\boldsymbol{X}_i, \quad \boldsymbol{X}_i \in S_j$$

上式表明，S_j类的聚类中心应选为该类样本的均值。

1. 算法描述

（1）任选K个初始聚类中心：$\boldsymbol{Z}_1(1)$，$\boldsymbol{Z}_2(1)$，…，$\boldsymbol{Z}_K(1)$

括号内序号：迭代运算的次序号。

（2）按最小距离原则将其余样品分配到*K*个聚类中心中的某一个，即：

若 $\min\left\{\left\|\boldsymbol{X}-\boldsymbol{Z}_i(k)\right\|, i=1,2,\cdots,K\right\}=\left\|\boldsymbol{X}-\boldsymbol{Z}_j(k)\right\|=D_j(k)$，则 $X\in S_j(k)$

注意：*k*——迭代运算次序号；*K*——聚类中心的个数 。

（3）计算各个聚类中心的新向量值：$\boldsymbol{Z}_j(k+1)\quad j=1,2,\cdots,K$

$$\boldsymbol{Z}_j(k+1)=\frac{1}{N_j}\sum_{X\in S_j(k)}\boldsymbol{X}\quad j=1,2,\cdots,K$$

N_j：第*j*类的样本数。

这里：分别计算*K*个聚类中的样本均值向量，故称K-均值算法。

（4）如果 $\boldsymbol{Z}_j(k+1)\neq\boldsymbol{Z}_j(k)\quad j=1,2,\cdots,K$，则回到（2），将模式样本逐个重新分类，重复迭代计算。

如果 $\boldsymbol{Z}_j(k+1)=\boldsymbol{Z}_j(k)\quad j=1,2,\cdots,K$，算法收敛，计算完毕。

“动态”聚类法

？

聚类过程中，
聚类中心位置或个数发生变化。

2. 算法讨论

结果受到所选聚类中心的个数和其初始位置，以及模式样本的几何性质及读入次序等的影响。实际应用中需要试探不同的K值和选择不同的聚类中心起始值。

例2.3：已知20个模式样本如下，试用K-均值算法分类。

$$\boldsymbol{X}_1=[0,0]^{\mathrm{T}}\quad \boldsymbol{X}_2=[1,0]^{\mathrm{T}}\quad \boldsymbol{X}_3=[0,1]^{\mathrm{T}}\quad \boldsymbol{X}_4=[1,1]^{\mathrm{T}}$$

$$\boldsymbol{X}_5=[2,1]^{\mathrm{T}}\quad \boldsymbol{X}_6=[1,2]^{\mathrm{T}}\quad \boldsymbol{X}_7=[2,2]^{\mathrm{T}}\quad \boldsymbol{X}_8=[3,2]^{\mathrm{T}}$$

$$\boldsymbol{X}_9=[6,6]^{\mathrm{T}}\quad \boldsymbol{X}_{10}=[7,6]^{\mathrm{T}}\quad \boldsymbol{X}_{11}=[8,6]^{\mathrm{T}}\quad \boldsymbol{X}_{12}=[6,7]^{\mathrm{T}}$$

$$\boldsymbol{X}_{13}=[7,7]^{\mathrm{T}}\quad \boldsymbol{X}_{14}=[8,7]^{\mathrm{T}}\quad \boldsymbol{X}_{15}=[9,7]^{\mathrm{T}}\quad \boldsymbol{X}_{16}=[7,8]^{\mathrm{T}}$$

$$\boldsymbol{X}_{17}=[8,8]^{\mathrm{T}}\quad \boldsymbol{X}_{18}=[9,8]^{\mathrm{T}}\quad \boldsymbol{X}_{19}=[8,9]^{\mathrm{T}}\quad \boldsymbol{X}_{20}=[9,9]^{\mathrm{T}}$$

解：① 取K=2，并选：$\boldsymbol{Z}_1(1)=\boldsymbol{X}_1=[0,0]^{\mathrm{T}}\quad \boldsymbol{Z}_2(1)=\boldsymbol{X}_2=[1,0]^{\mathrm{T}}$

② 计算距离，聚类：

$$\boldsymbol{X}_1:\ \left.\begin{aligned} D_1&=\|\boldsymbol{X}_1-\boldsymbol{Z}_1(1)\|=0\\ D_2&=\|\boldsymbol{X}_1-\boldsymbol{Z}_2(1)\|=\sqrt{(0-1)^2+(0-0)^2}=\sqrt{1}\end{aligned}\right\}\Rightarrow D_1<D_2\Rightarrow \boldsymbol{X}_1\in S_1(1)$$

$$\boldsymbol{X}_2:\ \left.\begin{aligned} D_1&=\|\boldsymbol{X}_2-\boldsymbol{Z}_1(1)\|=\sqrt{1}\\ D_2&=\|\boldsymbol{X}_2-\boldsymbol{Z}_2(1)\|=0\end{aligned}\right\}\Rightarrow D_2<D_1\Rightarrow \boldsymbol{X}_2\in S_2(1)$$

X_3:
$$\left.\begin{aligned} D_1 &= \| X_3 - Z_1(1) \| = \sqrt{(0-0)^2 + (1-0)^2} = \sqrt{1} \\ D_2 &= \| X_3 - Z_2(1) \| = \sqrt{(0-1)^2 + (1-0)^2} = \sqrt{2} \end{aligned}\right\} \Rightarrow D_1 < D_2 \Rightarrow X_3 \in S_1(1)$$

X_4:
$$\left.\begin{aligned} D_1 &= \| X_4 - Z_1(1) \| = \sqrt{(1-0)^2 + (1-0)^2} = \sqrt{2} \\ D_2 &= \| X_4 - Z_2(1) \| = \sqrt{(1-1)^2 + (1-0)^2} = \sqrt{1} \end{aligned}\right\} \Rightarrow D_2 < D_1 \Rightarrow X_4 \in S_2(1)$$

……，可得到：$S_1(1) = \{X_1, X_3\}$ $N_1 = 2$

$S_2(1) = \{X_2, X_4, X_5, \ldots, X_{20}\}$ $N_2 = 18$

③ 计算新的聚类中：

$$Z_1(2) = \frac{1}{N_1} \sum_{X \in S_1(1)} X = \frac{1}{2}(X_1 + X_3) = \frac{1}{2}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix}\right) = \begin{bmatrix} 0 \\ 0.5 \end{bmatrix}$$

$$Z_2(2) = \frac{1}{N_2} \sum_{X \in S_2(1)} X = \frac{1}{18}(X_2 + X_4 + \cdots + X_{20}) == \begin{bmatrix} 5.67 \\ 5.33 \end{bmatrix}$$

④ 判断：∵ $Z_j(2) \neq Z_j(1)$ $j = 1,2$，故返回第②步。

② 从新的聚类中心得：

$$\boldsymbol{X}_1:\quad \left.\begin{aligned} D_1 &= \| \boldsymbol{X}_1 - \boldsymbol{Z}_1(2) \| = \cdots \\ D_2 &= \| \boldsymbol{X}_1 - \boldsymbol{Z}_2(2) \| = \cdots \end{aligned}\right\} \Rightarrow \boldsymbol{X}_1 \in S_1(2)$$

$$\boldsymbol{X}_{20}:\quad \left.\begin{aligned} D_1 &= \| \boldsymbol{X}_{20} - \boldsymbol{Z}_1(2) \| = \cdots \\ D_2 &= \| \boldsymbol{X}_{20} - \boldsymbol{Z}_2(2) \| = \cdots \end{aligned}\right\} \Rightarrow \boldsymbol{X}_{20} \in S_2(2)$$

有：

$$S_1(2) = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_8\} \quad N_1 = 8$$

$$S_2(2) = \{\boldsymbol{X}_9, \boldsymbol{X}_{10}, ..., \boldsymbol{X}_{20}\} \quad N_2 = 12$$

③ 计算聚类中心：

$$\boldsymbol{Z}_1(3) = \frac{1}{N_1} \sum_{X \in S_1(2)} \boldsymbol{X} = \frac{1}{8}(\boldsymbol{X}_1 + \boldsymbol{X}_2 + \cdots + \boldsymbol{X}_8) = \begin{bmatrix} 1.25 \\ 1.13 \end{bmatrix}$$

$$\boldsymbol{Z}_2(3) = \frac{1}{N_2} \sum_{X \in S_2(2)} \boldsymbol{X} = \frac{1}{12}(\boldsymbol{X}_9 + \boldsymbol{X}_{10} + \cdots + \boldsymbol{X}_{20}) == \begin{bmatrix} 7.67 \\ 7.33 \end{bmatrix}$$

④ $\because \boldsymbol{Z}_j(3) \neq \boldsymbol{Z}_j(2) \quad j=1,2$

返回第②步，以$\boldsymbol{Z}_1(3)$，$\boldsymbol{Z}_2(3)$为中心进行聚类。

② 以新的聚类中心分类，求得的分类结果与前一次迭代结果相同：$S_1(3)=S_1(2) \quad S_2(3)=S_2(2)$

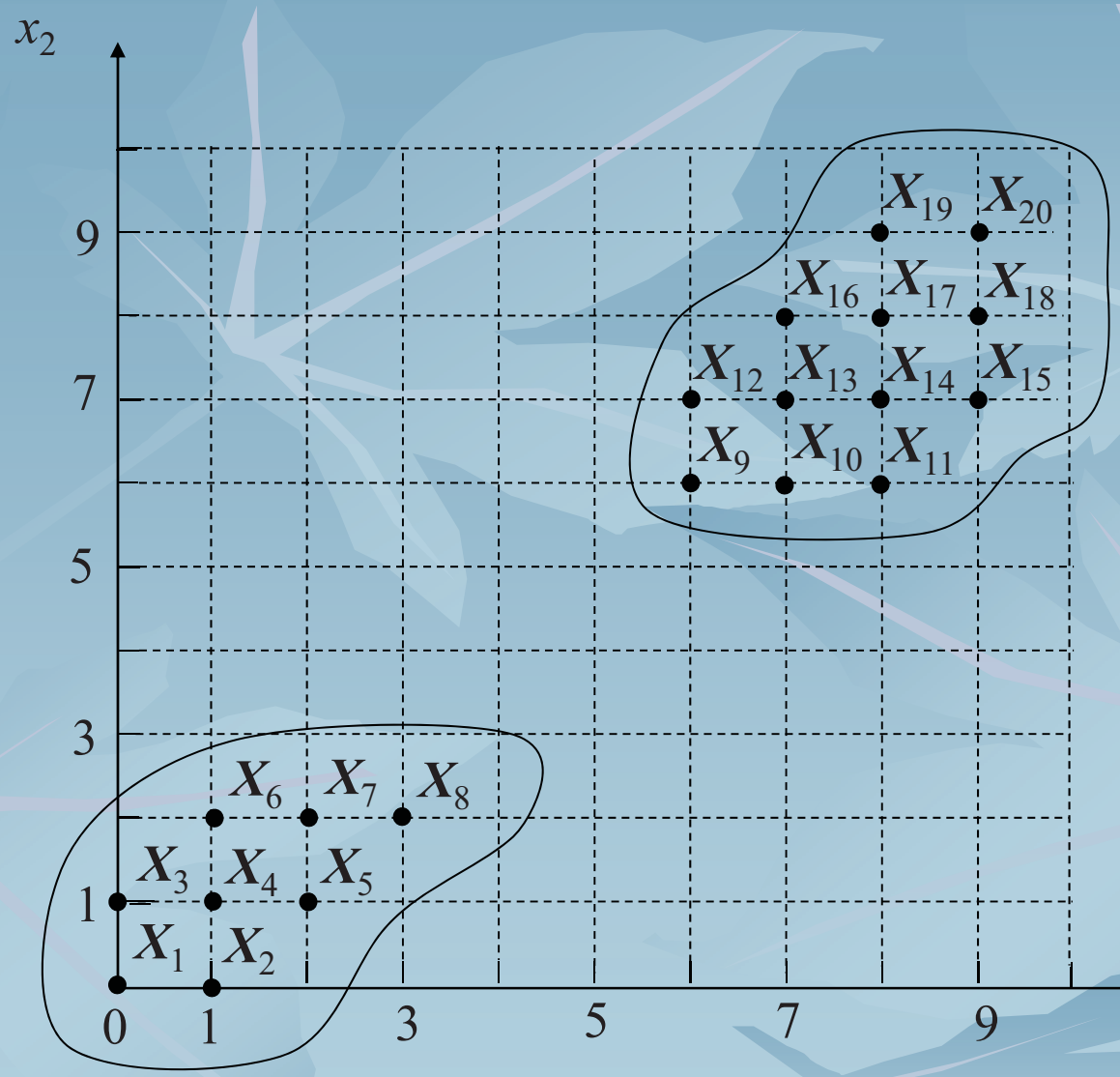
③ 计算新聚类中心向量值，聚类中心与前一次结果相同，即：

$$\boldsymbol{Z}_j(4)=\boldsymbol{Z}_j(3), \quad j=1,2$$

④ $\because \boldsymbol{Z}_j(4)=\boldsymbol{Z}_j(3)$ 故算法收敛，得聚类中心为

$$\boldsymbol{Z}_1=\begin{bmatrix}1.25\\1.13\end{bmatrix}, \quad \boldsymbol{Z}_2=\begin{bmatrix}7.67\\7.33\end{bmatrix}$$

结果图示：

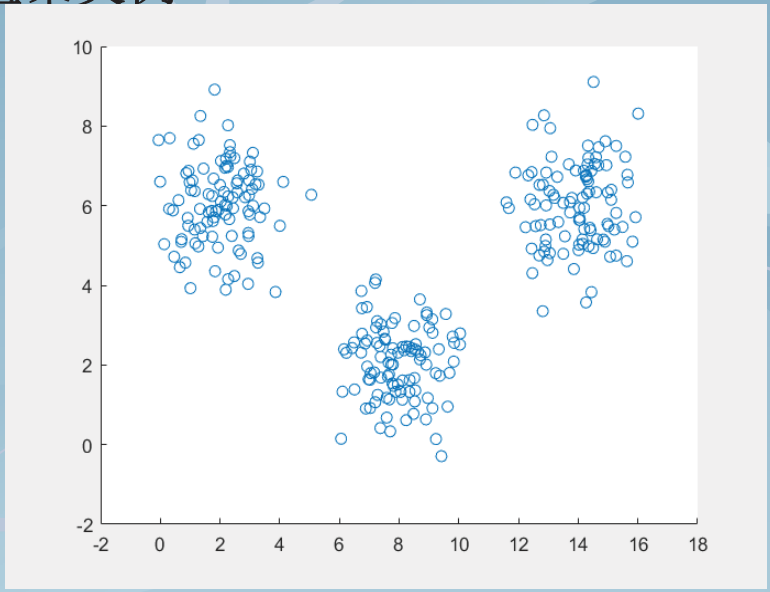


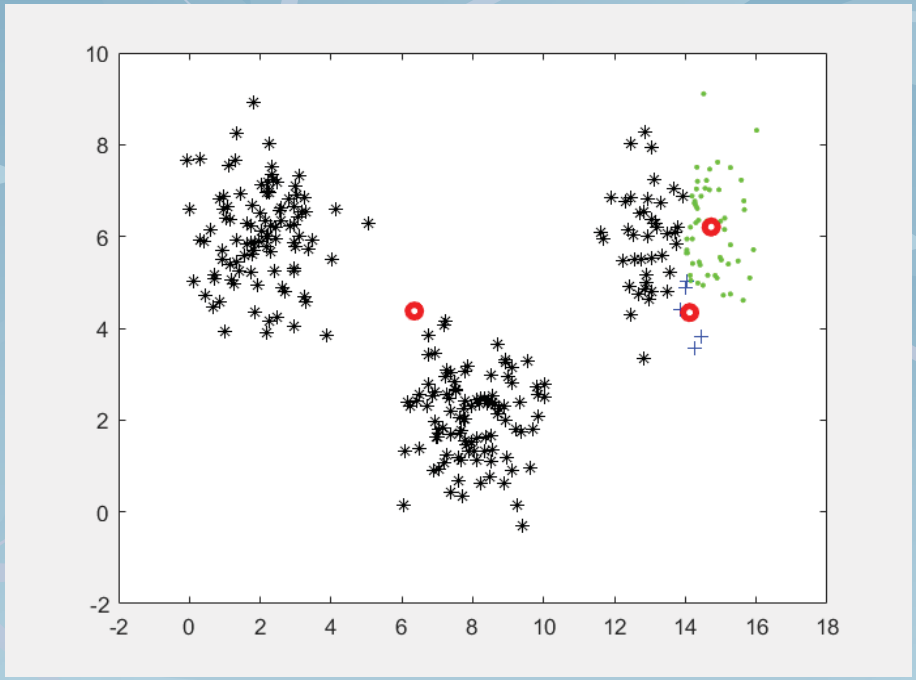
$$Z_1 = \begin{bmatrix} 1.25 \\ 1.13 \end{bmatrix}$$

$$Z_2 = \begin{bmatrix} 7.67 \\ 7.33 \end{bmatrix}$$

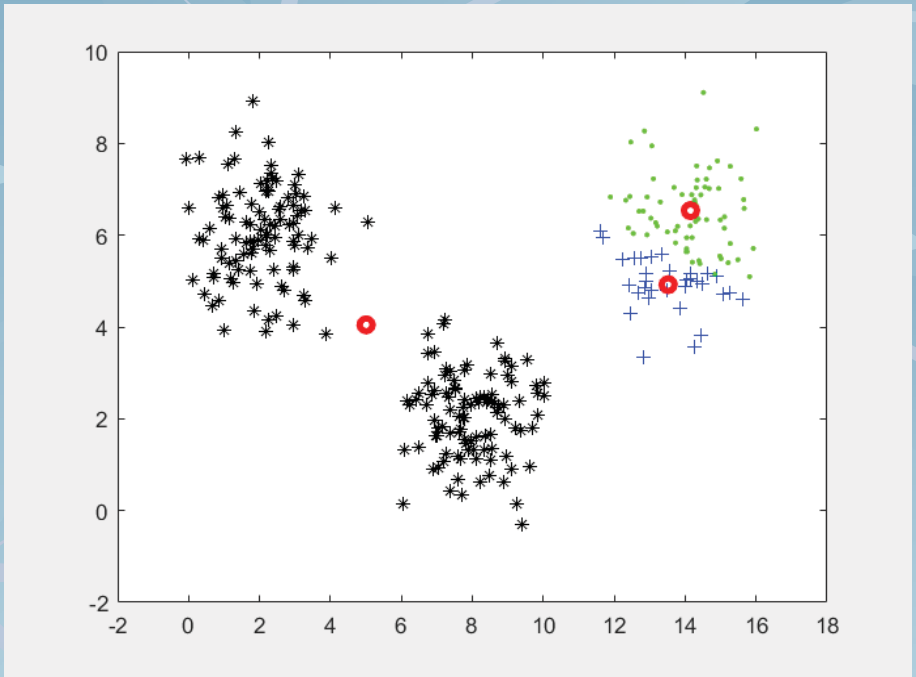
图2.10 K-均值算法聚类结果

K-均值聚类例

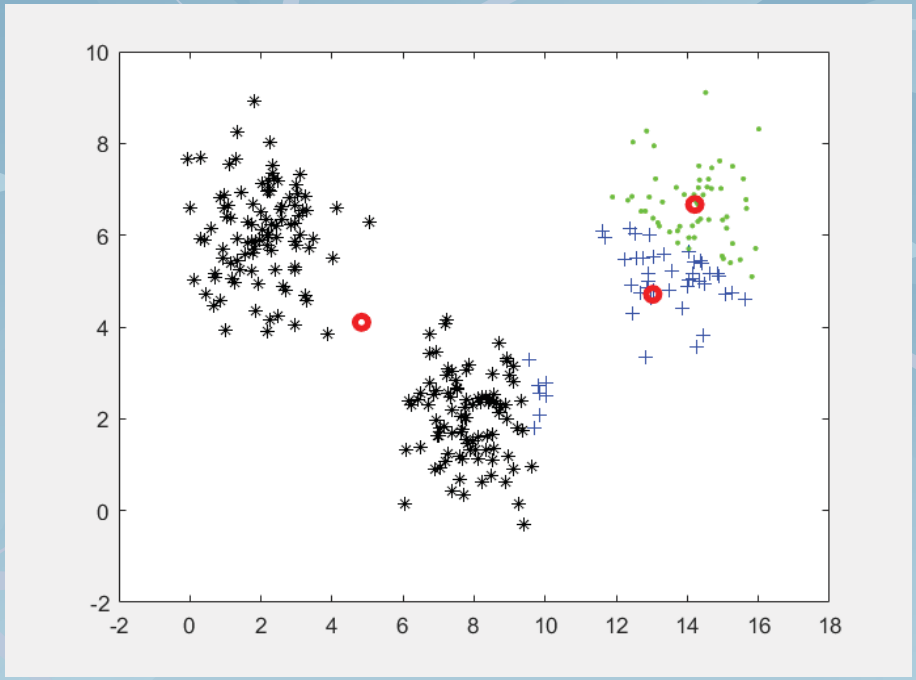




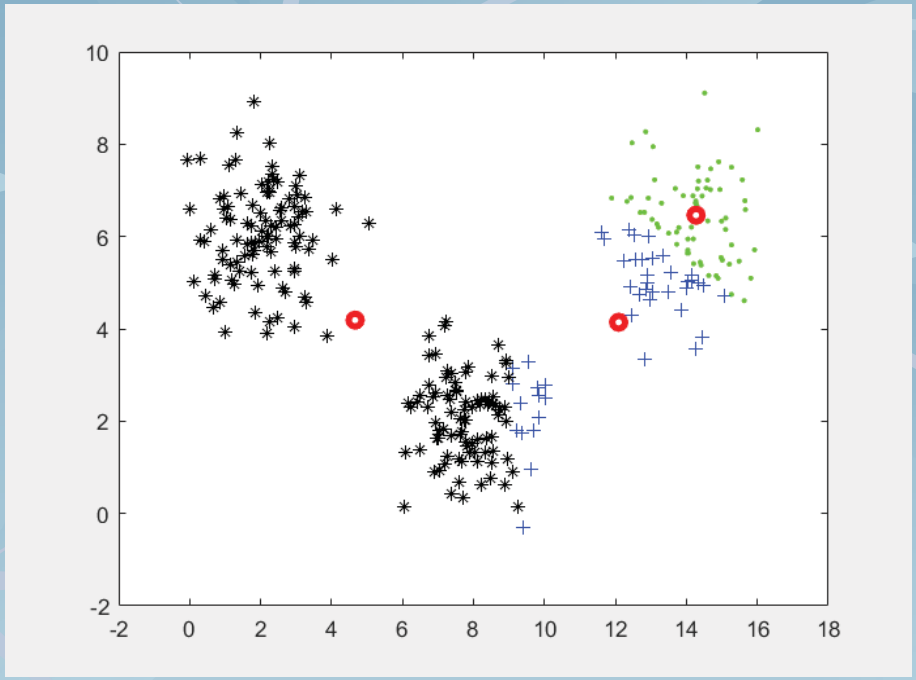
第一次迭代



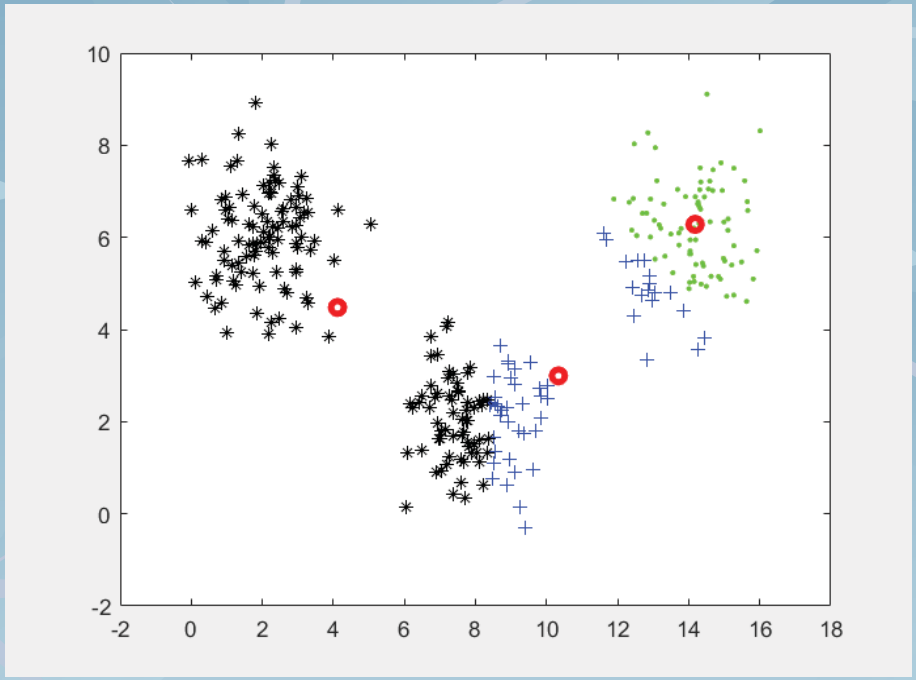
第二次迭代



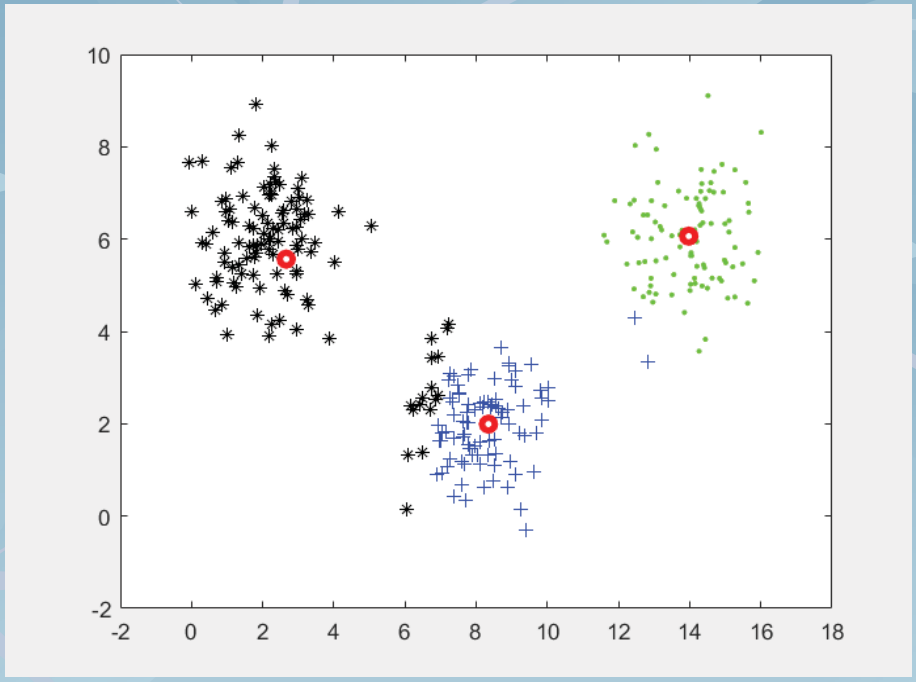
第三次迭代



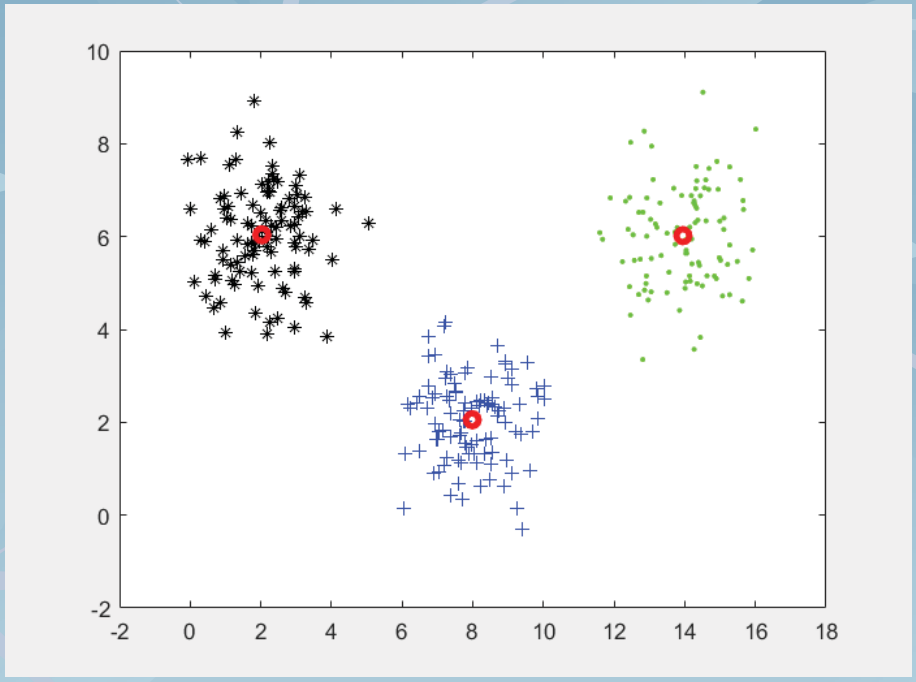
第四次迭代



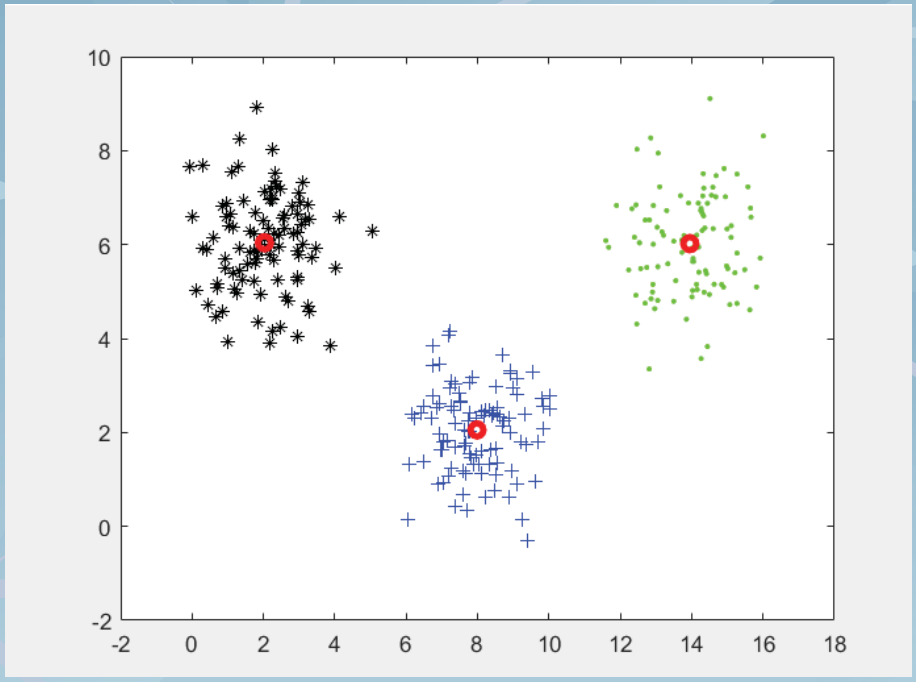
第五次迭代



第六次迭代



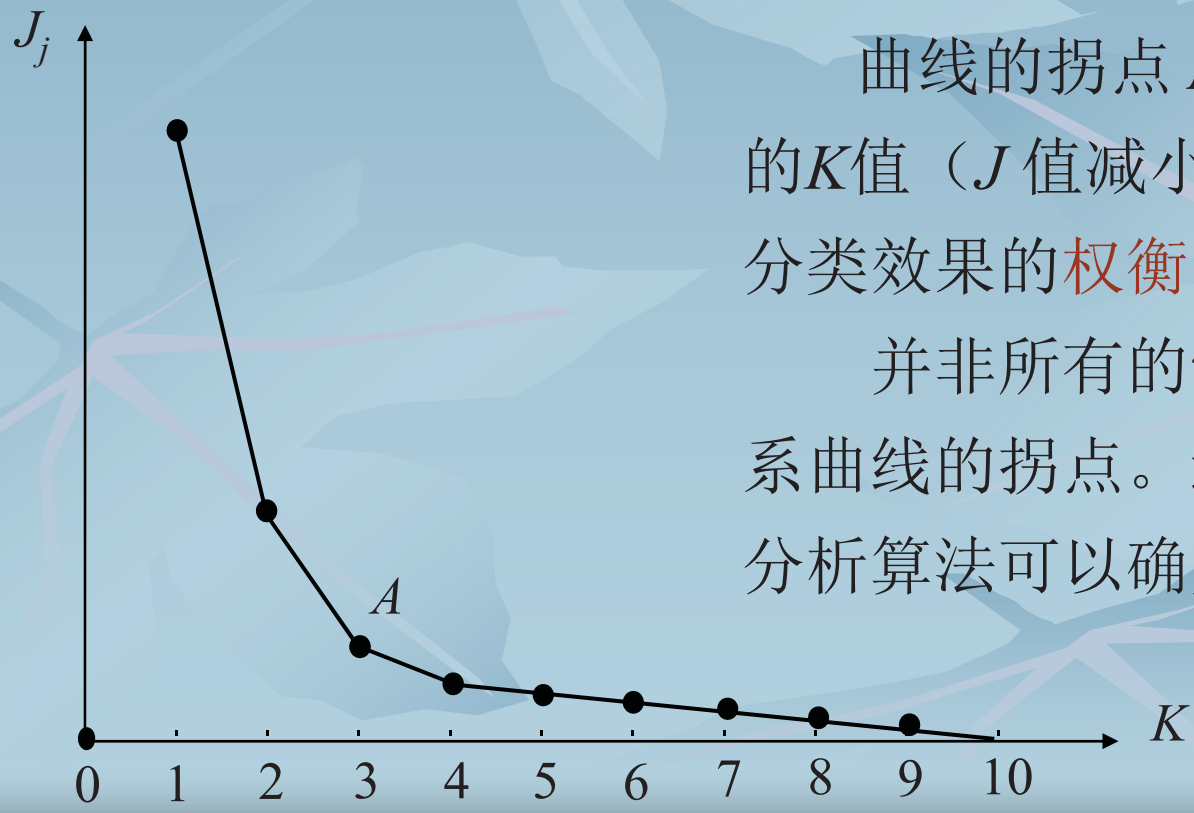
第七次迭代



第八次迭代

3、聚类准则函数J_j与K的关系曲线

上述K-均值算法，其类型数目假定已知为K个。当K未知时，可以令K逐渐增加， 此时J_j会单调减少。最初减小速度快，但当K增加到一定数值时，减小速度会减慢，直到K=总样本数N时，$J_j=0$。J_j-K关系曲线如下图：



曲线的拐点A对应着接近最优的K值（J值减小量、计算量以及分类效果的权衡）。

并非所有的情况都容易找到关系曲线的拐点。迭代自组织的数据分析算法可以确定模式类的个数K。

2.5.2 迭代自组织的数据分析算法

(iterative self-organizing data analysis techniques algorithm,ISODATA)

算法特点

加入了试探性步骤，组成人机交互的结构；

可以通过类的自动合并与分裂得到较合理的类别数。

与K-均值算法比较：

相似：聚类中心的位置均通过样本均值的迭代运算决定。

相异：K-均值算法的聚类中心个数不变；

ISODATA的聚类中心个数变化。

1．算法简介

基本思路：

（1）选择初始值——包括若干聚类中心及一些指标。可在迭代运算过程中人为修改，据此将N个模式样本分配到各个聚类中心去。

（2）按最近邻规则进行分类。

（3）聚类后的处理：计算各类中的距离函数等指标，按照给定的要求，将前次获得的聚类集进行分裂或合并处理，以获得新的聚类中心，即调整聚类中心的个数。

（4）判断结果是否符合要求：

符合，结束；

否则，回到（2）。

算法共分十四步：

第一～六步：预选参数，进行初始分类。

为合并和分裂准备必要的数据。

第七步：决定下一步是进行合并还是进行分裂。

第八～十步：分裂算法。

第十一～十三步：合并算法。

第十四步：决定算法是否结束。

2．算法描述

设有N个模式样本$\boldsymbol{X}_1$，$\boldsymbol{X}_2$，…，$\boldsymbol{X}_N$。

预选参数，进行初始分类。

第一步：预选N_C个聚类中心$\{\boldsymbol{Z}_1, \boldsymbol{Z}_2, \cdots, \boldsymbol{Z}_{N_C}\}$，$N_C$也是聚类过程中实际的聚类中心个数。预选指标：

K：希望的聚类中心的数目。

θ_N：每个聚类中应具有的最少样本数。若样本少于θ_N，则该类不能作为一个独立的聚类，应删去。

θ_S：一个聚类域中样本距离分布的标准差阈值。标准差向量的每一分量反映样本在特征空间的相应维上，与聚类中心的位置偏差（分散程度）。要求每一聚类内，其所有分量中的最大分量应小于θ_S，否则该类将被分裂为两类。

θ_C：两聚类中心之间的最小距离。若两类中心之间距离小于θ_C，则合并为一类。

L：在一次迭代中允许合并的聚类中心的最大对数。

I：允许迭代的次数。

第二步：把N个样本按最近邻规则分配到N_C个聚类中。

若 $$\|\boldsymbol{X}-\boldsymbol{Z}_j\|=\min\{\|\boldsymbol{X}-\boldsymbol{Z}_i\|, i=1,2,\cdots,N_C\}$$

则 $$\boldsymbol{X}\in S_j$$

第三步：若S_j中的样本数$N_j<\theta_N$，则取消该类，并且N_C减去1。

第四步：修正各聚类中心值。

$$\boldsymbol{Z}_j=\frac{1}{N_j}\sum_{\boldsymbol{X}\in S_j}\boldsymbol{X} \qquad j=1,2,\cdots,N_C$$

θ_N：每类应具有的最少样本数。

第五步：计算S_j类的类内平均距离$\overline{D_j}$。

$$\overline{D_j}=\frac{1}{N_j}\sum_{\boldsymbol{X}\in S_j}\|\boldsymbol{X}-\boldsymbol{Z}_j\| \qquad j=1,2,\cdots,N_C$$

第六步：计算总体平均距离$\overline{D}$，即全部样本到各自聚类中心距离的平均距离。

$$\overline{D}=\frac{1}{N}\sum_{j=1}^{N_C}\sum_{\boldsymbol{X}\in S_j}\|\boldsymbol{X}-\boldsymbol{Z}_j\|=\frac{1}{N}\sum_{j=1}^{N_C}N_j\overline{D_j}$$

判断分裂还是合并。

第七步：判决是进行分裂还是进行合并，决定迭代步骤等。

1) 若$N_C \le K/2$，即聚类中心小于或等于希望数的一半，进入第八步(分裂) 。

2) 如果迭代的次数是偶数，或$N_C \ge 2K$，即聚类中心数目大于或等于希望数的两倍，则跳到第十一步(合并)。否则进入第八步(分裂)。

I：允许迭代的次数。

θ_C：两聚类中心之间的最小距离。

N_C：预选的聚类中心数。

K：希望的聚类中心的数目。

分裂处理。

第八步：计算每个聚类中样本距离的标准差向量。对第S_j类有

$$\boldsymbol{\sigma}_j = [\sigma_{j1}, \sigma_{j2}, \cdots, \sigma_{jn}]^{\mathrm{T}}$$

$$\text{分量：} \sigma_{ji} = \sqrt{\frac{1}{N_j} \sum_{X \in S_j} (x_{ji} - z_{ji})^2} = \sqrt{\text{方差}}$$

$j = 1,2,\cdots,N_C$ 是聚类数；

$i = 1,2,\cdots,n$ 是维数（特征个数）。

第九步：求每个标准差向量的最大分量。$\boldsymbol{\sigma}_j$的最大分量记为

$\sigma_{j\max}$，j=1,2, …,N_C。

第十步：在最大分量集$\{\sigma_{j\max}, j=1,2,\cdots,N_C\}$中，如有$\sigma_{j\max}>\theta_S$，说明$S_j$类样本在对应方向上的标准差大于允许的值。此时，又满足以下两个条件之一：

1) $\overline{D_j}>\overline{D}$和$N_j>2(\theta_N+1)$，即类内平均距离大于总体平均距离，并且$S_j$类中样本数很大。

2) $N_C\le K/2$，即聚类数小于或等于希望数目的一半。

则将Z_j分裂成两个新的聚类中心Z_j^+和Z_j^-，并且N_C加1。其中

$Z_j^+=\sigma_{j\max}$对应的分量$+k\sigma_{j\max}$　　$0<k\le 1$：分裂系数

$Z_j^-=\sigma_{j\max}$对应的分量$-k\sigma_{j\max}$

按邻近规则聚类

若完成了分裂运算，迭代次数加1，跳回第二步；否则，继续。

θ_S：聚类域中样本距离分布的标准差阈值。

θ_N：每个聚类中应具有的最少样本数。

合并处理。

第十一步：计算所有聚类中心之间的距离。S_i类和S_j类中心间的距离为

$$D_{ij} = \left\| \boldsymbol{Z}_i - \boldsymbol{Z}_j \right\| \qquad i = 1,2,\cdots,N_C - 1 \qquad j = i+1,\cdots,N_C$$

第十二步：比较所有D_{ij}与θ_C的值，将小于θ_C的D_{ij}按升序排列

$$\left\{ D_{i_1 j_1}, D_{i_2 j_2}, \cdots, D_{i_L j_L} \right\}$$

第十三步：如果将距离为$D_{i_l j_l}$的两类合并，得到新的聚类中心为

$$\boldsymbol{Z}_l^* = \frac{1}{N_{i_l} + N_{j_l}} \left(N_{i_l} \boldsymbol{Z}_{i_l} + N_{j_l} \boldsymbol{Z}_{j_l} \right) \qquad l = 1,2,\cdots,L$$

每合并一对，N_C减1。

θ_C：两聚类中心之间的最小距离。

判断结束。

第十四步：若是最后一次运算(迭代次数为I)，算法结束。

否则，有两种情况：

1) 需要由操作者修改输入参数时(试探性步骤)，跳到第一步；

2) 输入参数不需改变时，跳到第二步。

按邻近规则聚类

此时，选择两者之一，迭代次数加1，然后继续进行运算。

2.6 聚类结果的评价

1、评价的重要性

1) 对高维特征向量样本，不能直观看清聚类效果时。

2) 人机交互系统中，需要迅速地判断中间结果，及时指导输入参数的改变，较快地获得较好的聚类结果。

2、常用的几个指标

各指标综合考虑。

1) 聚类中心之间的距离。

2) 诸聚类域中样本数目。

3) 诸聚类域内样本的标准差向量。

例：

$$\boldsymbol{\sigma}_1 = (1.2, 0.9, 0.7, 1.0)^{\mathrm{T}}$$

聚类域内样本分布近似为超球体。

$$\boldsymbol{\sigma}_2 = (4.2, 5.4, 18.3, 3.3)^{\mathrm{T}}$$

沿第三轴形成长条的（四维）超椭球体分布。