第 5 章

CHAPTER 5

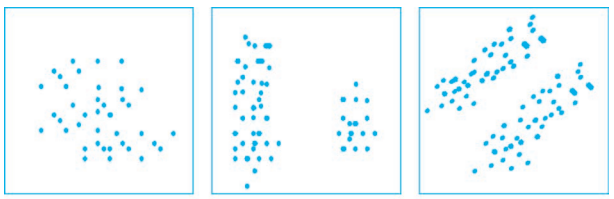
聚 类 分 析

俗话说“物以类聚，人以群分”，这句话实际上就反映了聚类分析的基本思想。一般来说，数据集根据其客观属性可分为若干个自然类，每个自然类中的数据的一些属性都具有较强的相似性。聚类分析是基于这种思想而建立的一种数据描述方法。在第 2 章中为了获取判别模型的参数，需要由带有类别标签的数据组成训练样本集，但在实际应用中，常常会遇到因条件限制无法得到训练样本集，只是要求对已获取的大量未知类别数据，根据这些数据中的特性进行分类。在模式识别系统中，我们称这种算法为聚类算法。聚类算法把彼此特征相似的数据归入同一类，而把特征不相似的数据分到不同的类中，而且在分类中不需要用训练样本进行学习，所以也称为无监督分类。

第 12 集
微课视频

5.1 模式相似性测度

在贝叶斯判决中，为了求得后验概率，需要已知先验概率和类条件概率。由于条件概率通常也未知，就需要用训练样本去对概率密度进行估计。在实际应用中，这一过程往往非常困难。聚类分析避免了估计类概率密度的困难，每个聚类中心都是局部密度极大值位置，越靠近聚合中心，密度越高；越远离聚合中心，密度越小。聚类算法把特征相似性的样本聚集为一个类别，在特征空间里占据着一个局部区域。每个局部区域都形成一个聚类中心，聚类中心代表相应类别。如图 5-1 所示为具有相同的平均值和协方差矩阵的数据集，无论采用参数估计，还是非参数估计，都无法取得合理的结果，而采用聚类分析，从图中可以直观看出图 5-1(a)具有一个类别，图 5-1(b)和图 5-1(c)各有两个类别。



(a) 一个类别　(b) 两个类别1　(c) 两个类别2

图 5-1　具有相同的平均值和协方差矩阵的数据集

特征选择是聚类分析的关键因素，选取不同的特征，聚类的结果可能不同。如图 5-2(a)所示为混合训练样本集；根据样本的面积大小可分成三类的情况，如图 5-2(b)所示；根据

外形特征分成三类的情况，如图 5-2(c)所示；根据线型分成两类的情况，如图 5-2(d)所示。可以想象，属于不同类别的样本，它们之间必然存在某些特征显著不同。如果在聚类分析中，未把不同类的样本区分开来，可能是因为特征选择不当，没有选取标志类别显著差别的特征，这时应当重新选择特征。特征选择不当不仅可能会使聚类性能下降，甚至会使聚类完全无效。特征较少可能会使特征向量包含的分类信息太少，特征太多又会使特征之间产生信息冗余，都会直接影响聚类的结果。因此，特征选择也成了聚类分析中最困难的环节之一。关于特征选择的方法，我们将在第 6 章详细介绍。

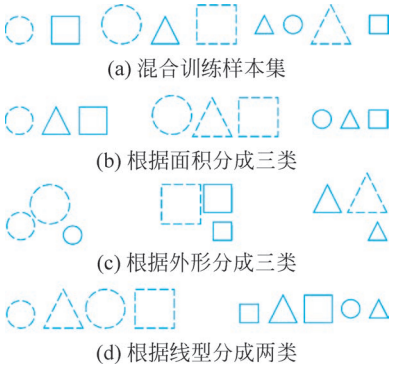


图 5-2　聚类分析的结果与特征的选取的关系示例

实际应用中，对已知样本集 $\boldsymbol{X}^{(N)}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$，是按某种相似性把 $\boldsymbol{X}$ 分类，衡量样本相似性的方法对聚类结果同样也有很大的影响。为了能区分样本的类别，首先需要定义模式相似性的测度。

5.1.1　距离测度

若一个样本模式被表示成特征向量，则对应于特征空间的一个点。当样本特征选择恰当，也即同类样本特征相似，不同类样本的特征显著不同时，同类样本就会聚集在一个区域，不同类样本相对远离。显然，样本点在特征空间的距离直接反映了相应样本所属类别，可以作为样本相似性度量。距离越近，相似性越大，属于同一类的可能性就越大；距离越远，相似性越小，属于同一类的可能性就越小。聚类分析中，最常用的就是距离相似性测度。实际应用中，有各种各样距离的定义，下面我们给出距离定义应满足的条件。

设已知 3 个样本，它们分别为 $\boldsymbol{X}_i=(x_{i1},x_{i2},\cdots,x_{id})^{\mathrm{T}}$、$\boldsymbol{X}_j=(x_{j1},x_{j2},\cdots,x_{jd})^{\mathrm{T}}$ 和 $\boldsymbol{X}_k=(x_{k1},x_{k2},\cdots,x_{kd})^{\mathrm{T}}$。其中，$d$ 为特征空间的维数，向量 $\boldsymbol{X}_i$ 和 $\boldsymbol{X}_j$ 的距离以及 $\boldsymbol{X}_i$ 和 $\boldsymbol{X}_k$ 的距离分别记为 $D(\boldsymbol{X}_i,\boldsymbol{X}_j)$ 和 $D(\boldsymbol{X}_i,\boldsymbol{X}_k)$，对任意两向量的距离定义应满足下面的公理：

(1) $D(\boldsymbol{X}_i,\boldsymbol{X}_j)\geqslant 0$，当且仅当 $\boldsymbol{X}_i=\boldsymbol{X}_j$ 时，等号成立。

(2) $D(\boldsymbol{X}_i,\boldsymbol{X}_j)=D(\boldsymbol{X}_j,\boldsymbol{X}_i)$。

(3) $D(\boldsymbol{X}_i,\boldsymbol{X}_j)\leqslant D(\boldsymbol{X}_j,\boldsymbol{X}_k)+D(\boldsymbol{X}_i,\boldsymbol{X}_k)$。

需要指出，模式识别中定义的某些距离测度不满足第 3 个条件，只是在广义上称为距离。下面给出距离测度的几种具体算式。

1. 欧氏距离

$$D_e(\boldsymbol{X}_i,\boldsymbol{X}_j)=\|\boldsymbol{X}_i-\boldsymbol{X}_j\|=\sqrt{\sum_{k=1}^{d}|x_{ik}-x_{jk}|^2} \tag{5-1}$$

根据 $D_e(\boldsymbol{X}_1,\boldsymbol{X}_2)$的定义，通过选择合适的门限 d_s，可以判决 $\boldsymbol{X}_1$ 和 $\boldsymbol{X}_2$ 是否为同一类别。当 $D_e(\boldsymbol{X}_1,\boldsymbol{X}_2)$小于门限 d_s 时，表示 $\boldsymbol{X}_1$ 和 $\boldsymbol{X}_2$ 属于同一类别，反之，则属于不同类别。这里门限 d_s 的选取非常关键，若 d_s 选择过大，则全部样本被归为同一类别；若 d_s 选取过小，则可能造成每个样本都单独构成一个类别。必须正确选择门限值以保证正确分类。实际应用中还需注意以下两点。

(1) 模式特征向量的构成。一种物理量对应一种量纲，而一种量纲一般有不同的单位制式，每种单位制式下又有不同的单位，简单地说，就是一种物理量对应着一个具体的单位。对于各特征向量，对应的维度上应当是相同的物理量，并且要注意物理量的单位。

通常，特征向量中的每一维所表示的物理意义不尽相同，如 x_1 表示周长，x_2 表示面积等。如果某些维度上的物理量采用的单位发生变化，就可能会导致相同样本集出现不同的聚类结果。如图 5-3 所示，a、b、c 和 d 表示 4 个二维向量，向量的两个分量 x_1、x_2 均表示长度，当 x_1、x_2 的单位发生不同的变化时，会出现不同的聚类结果。如图 5-3(b)所示 a，b 为一类，c，d 为另一类，如图 5-3(c)所示 a，c 为一类，b，d 为另一类。由此可见，坐标轴的简单缩放就能引起样本点的重新聚类。

(2) 在实际应用中，可以采用特征数据标准化方法对原始特征进行预处理，使其与变量的单位无关。此时所描述的点是一种相对的位置关系，只要样本点间的相对位置关系不变，就不会影响聚类结果。例如，对图 5-3(b)和图 5-3(c)中的数据标准化后，4 个点的相对位置关系总是和图 5-3(a)相同。

需要指出的是，并不是所有的标准化都是合理的。如果数据散布恰恰是由于类别差异引起的，标准化反而会引起错误的聚类结果。因此，在聚类之前是否应进行标准化处理，建立在对数据各维度物理量充分研判的基础上。

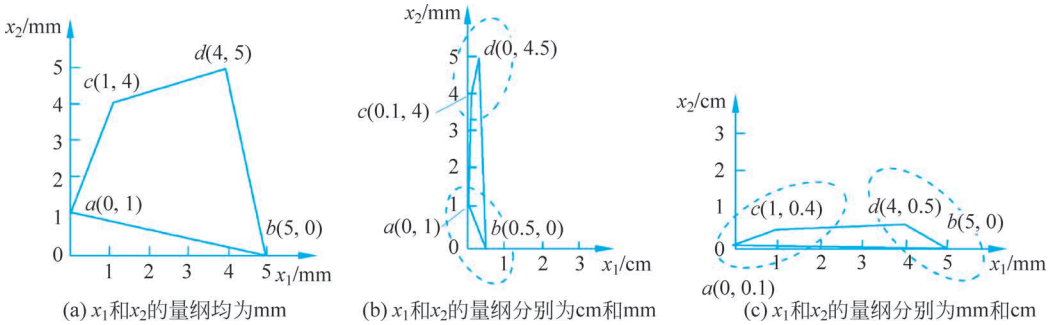


图 5-3 特征量纲对聚类结果的影响

2. 绝对值距离(街坊距离或曼哈顿距离)

$$D(\boldsymbol{X}_i,\boldsymbol{X}_j)=\sum_{k=1}^{n}|x_{ik}-x_{jk}|,\quad k=1,2,\cdots,d \tag{5-2}$$

3. 切比雪夫(Chebyshev)距离

$$D(\boldsymbol{X}_i,\boldsymbol{X}_j)=\max|x_{ik}-x_{jk}|,\quad k=1,2,\cdots,d \tag{5-3}$$

4. 闵可夫斯基(Minkowski)距离

$$D_\lambda(\boldsymbol{X}_i,\boldsymbol{X}_j)=\left[\sum_{k=1}^{d}|x_{ik}-x_{jk}|^\lambda\right]^{\frac{1}{\lambda}},\quad \lambda>0\quad k=1,2,\cdots,d \tag{5-4}$$

它是若干距离函数的通式：当 $\lambda=2$ 时，等于欧氏距离；当 $\lambda=1$ 时，称为“街坊”(City Block)距离。

5. 马哈拉诺比斯(Mahalanobis)距离

设 n 维向量 $\boldsymbol{X}_i$ 是向量集 $\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$ 中的一个向量，其马哈拉诺比斯距离的平方定义为

$$D^2(\boldsymbol{X}_i,\boldsymbol{\mu})=(X_i-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i-\boldsymbol{\mu}) \tag{5-5}$$

式中，$\boldsymbol{\Sigma}=\dfrac{1}{N-1}\sum_{i=1}^{N}(\boldsymbol{X}_i-\boldsymbol{\mu})(\boldsymbol{X}_i-\boldsymbol{\mu})^{\mathrm{T}}$，$\boldsymbol{\mu}=\dfrac{1}{N}\sum_{i=1}^{N}X_i$。

容易证明，马哈拉诺比斯距离对一切非奇异线性变换都是不变的，即具有坐标系比例、旋转、平移不变性，并且从统计意义上尽量去掉了分量间的相关性。这说明它不受特征量纲选择的影响。另外，由于$\boldsymbol{\Sigma}$ 的含义是这个向量集的协方差阵的统计量，所以马哈拉诺比斯距离对特征的相关性也做了考虑。当$\boldsymbol{\Sigma}$ 为单位矩阵时，马哈拉诺比斯距离和欧氏距离是等价的。

【例 5.1】 已知二维正态母体 G 的分布为

$$N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1 & 0.9\\0.9 & 1\end{bmatrix}\right)$$

求点 $\boldsymbol{A}=[1,1]^{\mathrm{T}}$ 和 $\boldsymbol{B}=[1,-1]^{\mathrm{T}}$ 至均值点$\boldsymbol{\mu}=[0,0]^{\mathrm{T}}$ 的马哈拉诺比斯距离和欧氏距离。

【解】 由题设，可得

$$\boldsymbol{\Sigma}=\begin{bmatrix}1 & 0.9\\0.9 & 1\end{bmatrix},\quad \boldsymbol{\Sigma}^{-1}=\frac{1}{0.19}\begin{bmatrix}1 & -0.9\\-0.9 & 1\end{bmatrix}$$

从而马哈拉诺比斯距离

$$D^2(A,\boldsymbol{\mu})=[1,1]\boldsymbol{\Sigma}^{-1}\begin{bmatrix}1\\1\end{bmatrix}=\frac{0.2}{0.19}$$

$$D^2(B,\boldsymbol{\mu})=[1,-1]\boldsymbol{\Sigma}^{-1}\begin{bmatrix}1\\-1\end{bmatrix}=\frac{3.8}{0.19}$$

点 B 是点 A 的$\sqrt{19}$倍，若用欧氏距离，算得的距离值相同，均为$\sqrt{2}$。由分布函数知，A 和 B 两点的概率密度分别为 $p(1,1)=0.2157$ 和 $p(1,-1)=0.00001658$。

6. 堪培拉(Canberra)距离(兰氏距离)

$$D(\boldsymbol{X}_i,\boldsymbol{X}_j)=\sum_{k=1}^{n}\frac{|x_{ik}-x_{jk}|}{|x_{ik}+x_{jk}|},\quad (x_{ik},x_{jk}\geqslant 0,x_{ik}+x_{jk}\neq 0) \tag{5-6}$$

该距离能克服量纲引起的问题，但不能克服分量间的相关性。

5.1.2　相似测度

与距离测度不同，相似测度考虑两向量的方向是否相近，向量长度并不重要。如果两样本点在特征空间的方向越接近，则两样点划归为同一类别的可能性越大。下面给出相似测度的几种定义。

1. 角度相似系数(夹角余弦)

样本 $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_j$ 之间的角度相似性度量定义为它们之间夹角的余弦,也是单位向量之间的点积(内积)即

$$S(\boldsymbol{X}_i,\boldsymbol{X}_j)=\cos\theta=\frac{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_j}{\|\boldsymbol{X}_i\|\cdot\|\boldsymbol{X}_j\|} \tag{5-7}$$

$|S(\boldsymbol{X}_i,\boldsymbol{X}_j)|\leqslant 1$,$S(\boldsymbol{X}_i,\boldsymbol{X}_j)$越大,$\boldsymbol{X}_i$ 与 $\boldsymbol{X}_j$ 越相似;当 $\boldsymbol{X}_i=\boldsymbol{X}_j$ 时,$S(\boldsymbol{X}_i,\boldsymbol{X}_j)$达到最大值。因向量长度已规格化,$S(\boldsymbol{X}_i,\boldsymbol{X}_j)$对于坐标系的旋转及放大、缩小是不变的,但对位移和一般性的线性变换不具有不变性。当 $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_j$ 的各特征为(0,1)二元取值时,$S(\boldsymbol{X}_i,\boldsymbol{X}_j)$的意义如下:①若模式样本的第 i 维特征取值为 1,则该样本占有第 i 维特征;②若模式样本的第 i 维特征取值为 0,则该样本无此维特征。此时,$\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_j$ 表示 $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_j$ 两个样本中共有的特征数目。$S(\boldsymbol{X}_i,\boldsymbol{X}_j)$反映 $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_j$ 共有的特征数目的相似性度量。$S(\boldsymbol{X}_i,\boldsymbol{X}_j)$越大,共有特征数目越多,相似性越高。

2. 相关系数

相关系数定义为数据中心化后的向量夹角余弦,即

$$R(\boldsymbol{X},\boldsymbol{Y})=\frac{(\boldsymbol{X}-\boldsymbol{\mu}_X)^{\mathrm{T}}(\boldsymbol{Y}-\boldsymbol{\mu}_Y)}{[(\boldsymbol{X}-\boldsymbol{\mu}_X)^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{\mu}_X)(\boldsymbol{Y}-\boldsymbol{\mu}_Y)^{\mathrm{T}}(\boldsymbol{Y}-\boldsymbol{\mu}_Y)]^{1/2}} \tag{5-8}$$

式中,$\boldsymbol{X}=(x_1,x_2,\cdots,x_d)$,$\boldsymbol{Y}=(y_1,y_2,\cdots,y_d)$分别为两个数据集的样本,$\boldsymbol{\mu}_X$ 和 $\boldsymbol{\mu}_Y$ 分别为这两个数据集的平均向量。

相关系数对于坐标系的平移、旋转和尺度缩放具有不变性。

3. 指数相似系数

已知样本 $\boldsymbol{X}_i=(x_{i1},x_{i2},\cdots,x_{id})$、$\boldsymbol{X}_j=(x_{j1},x_{j2},\cdots,x_{jd})$,其指数相似系数定义为

$$E(\boldsymbol{X}_i,\boldsymbol{X}_j)=\frac{1}{d}\sum_{k=1}^{d}\exp\left[-\frac{3(x_{ik}-x_{jk})^2}{4\sigma_k}\right] \tag{5-9}$$

式中,σ_k^2 为相应分量的协方差,d 为向量维数。

4. 其他相似测度

当样本 $\boldsymbol{X}_i=(x_{i1},x_{i2},\cdots,x_{id})$、$\boldsymbol{X}_j=(x_{j1},x_{j2},\cdots,x_{jd})$各特征值非负时,还可定义下列相似系数:

$$S_1(\boldsymbol{X}_i,\boldsymbol{X}_j)=\frac{\sum_{k=1}\min(x_{ik},x_{jk})}{\sum_{k=1}\max(x_{ik},x_{jk})} \tag{5-10}$$

$$S_2(\boldsymbol{X}_i,\boldsymbol{X}_j)=\frac{\sum_{k=1}\min(x_{ik},x_{jk})}{\frac{1}{2}\sum_{k=1}(x_{ik}+x_{jk})} \tag{5-11}$$

$$S_3(\boldsymbol{X}_i,\boldsymbol{X}_j)=\frac{\sum_{k=1}\min(x_{ik},x_{jk})}{\sum_{k=1}\sqrt{x_{ik}x_{jk}}} \tag{5-12}$$

上述相似性系数,均可作为样本相似测度。当两个样本 $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_j$ 越相似时,$S(\boldsymbol{X}_i,\boldsymbol{X}_j)$

的值越大；当 $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_j$ 相等时，其值为 1。

5.1.3　匹配测度

当 $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_j$ 的各特征为(0,1)二元取值时，我们称之为二值特征。对于给定的二值特征向量 $\boldsymbol{X}_i=(x_{i1},x_{i2},\cdots,x_{id})$ 和 $\boldsymbol{X}_j=(x_{j1},x_{j2},\cdots,x_{jd})$，根据它们两个相应分量 x_{ik} 与 x_{jk} 的取值，可定义如下四种匹配关系：若 $x_{ik}=1$ 和 $x_{jk}=1$，则称 x_{ik} 与 x_{jk} 是(1-1)匹配；若 $x_{ik}=1$ 和 $x_{jk}=0$，则称 x_{ik} 与 x_{jk} 是(1-0)匹配；若 $x_{ik}=1$ 和 $x_{jk}=0$，则称 x_{ik} 与 x_{jk} 是(0-1)匹配；若 $x_{ik}=0$ 和 $x_{jk}=0$，则称 x_{ik} 与 x_{jk} 是(0-0)匹配。令

$$a=\sum_{i=1}x_iy_i\quad b=\sum_{i=1}y_i(1-x_i)\quad c=\sum_{i=1}x_i(1-y_i)\quad e=\sum_{i=1}(1-x_i)(1-y_i)$$

则 a、b、c、e 分别表示 $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_j$ 的(1-1)、(0-1)、(1-0)和(0-0)的匹配特征数目。对于二值 d 维特征向量可定义如下相似性测度。

(1) 谷本(Tanimoto)测度：

$$S_t(\boldsymbol{X}_i,\boldsymbol{X}_j)=\frac{a}{a+b+c}=\frac{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_j}{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i+\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j-\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_j}\tag{5-13}$$

可以看出，$S_t(\boldsymbol{X},\boldsymbol{Y})$等于 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 共同具有的特征数目与 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 分别具有的特征种类总数之比。这里只考虑(1-1)匹配而不考虑(0-0)匹配。

(2) Rao 测度：

$$S_r(\boldsymbol{X}_i,\boldsymbol{X}_j)=\frac{a}{a+b+c+e}=\frac{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_j}{d}\tag{5-14}$$

上式等于(1-1)匹配特征数目和所选用的特征数目之比。

(3) 简单匹配系数：

$$M(\boldsymbol{X}_i,\boldsymbol{X}_j)=\frac{a+e}{d}\tag{5-15}$$

这时，匹配系数分子为(1-1)匹配特征数目与(0-0)匹配特征数目之和，分母为所考虑的特征数目。

(4) Dice 系数：

$$M_D(\boldsymbol{X}_i,\boldsymbol{X}_j)=\frac{a}{2a+b+c}=\frac{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_j}{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i+\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j}\tag{5-16}$$

(5) Kulzinsky 系数：

$$M_K(\boldsymbol{X}_i,\boldsymbol{X}_j)=\frac{a}{b+c}=\frac{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_j}{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i+\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j-2\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_j}\tag{5-17}$$

上式分子为(1-1)匹配特征数目，分母为(1-0)和(0-1)匹配特征数目之和，即不匹配特征数目之和。

【例 5.2】 已知两样本 $\boldsymbol{X}_1=(010110)^{\mathrm{T}}$，$\boldsymbol{X}_2=(001110)^{\mathrm{T}}$，求其 Tanimoto 测度。

【解】

$$\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{X}_2=2,\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{X}_1=3,\boldsymbol{X}_2^{\mathrm{T}}\boldsymbol{X}_2=3$$

$$S_t(\boldsymbol{X}_1,\boldsymbol{X}_2)=\frac{2}{3+3-2}=\frac{1}{2}=0.5$$

上面从不同角度给出了许多样本相似性测度的定义，各种相似性测度有其特点和适用

的条件，在实际使用时应根据具体问题进行选择。建立了模式相似性测度之后，两个样本的相似程度就可量化了，据此便可以进行聚类分析。

5.2 类间距离测度方法

在有些聚类算法中要用到类间距离，下面给出一些类间距离定义方式。

5.2.1 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

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

式中，$D(\boldsymbol{X}_H,\boldsymbol{X}_K)$表示 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} 表示 H 类中所有样本与 K 类中所有样本之间的最小距离，如图 5-4(a)所示。如果 K 类由 I 和 J 两类合并而成，如图 5-4(b)所示，则得到递推公式：

$$D_{HK}=\min\{D_{HI},D_{HJ}\} \tag{5-18}$$

(a) H类所有样本与K类所有样本间的最小距离示意图

(b) K类由I和J两类合并而成的情况下最小距离示意图

图 5-4 最短距离法示意图

5.2.2 最长距离法

与最短距离法类似，两个聚类 H 和 K 之间的最长距离定义为

$$D_{HK}=\max\{D(\boldsymbol{X}_H,\boldsymbol{X}_K)\} \tag{5-19}$$

其中，$\boldsymbol{X}_H\in H,\boldsymbol{X}_K\in K$。

若 K 类由 I 和 J 两类合并而成，则得到递推公式：

$$D_{HK}=\max\{D_{HI},D_{HJ}\} \tag{5-20}$$

5.2.3 中间距离法

中间距离法介于最长与最短的距离之间。若 K 类由 I 和 J 两类合并而成，则 H 和 K 类之间的距离为

$$D_{HK}=\sqrt{\frac{1}{2}D_{HI}^2+\frac{1}{2}D_{HJ}^2-\frac{1}{4}D_{IJ}^2} \tag{5-21}$$

5.2.4 重心法

以上定义的类间距离中并未考虑每一类所包含的样本数目，重心法在这一方面有所改进。从物理的观点看，一个类的空间位置若要用一个点表示，那么用它的重心代表较合理。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

$$D_{HK}=\sqrt{\frac{N_I}{N_I+N_J}D_{HI}^2+\frac{N_J}{N_I+N_J}D_{HJ}^2-\frac{N_IN_J}{(N_I+N_J)^2}D_{IJ}^2} \tag{5-22}$$

5.2.5　平均距离法(类平均距离法)

设 H、K 是两个聚类,则 H 类和 K 类间的距离定义为

$$D_{HK}=\sqrt{\frac{1}{N_HN_K}\sum_{\substack{i\in H\\ j\in K}}D_{ij}^2} \tag{5-23}$$

式中,D_{ij}^2 是 H 类任一样本 $\boldsymbol{X}_H$ 和 K 类任一样本 $\boldsymbol{X}_K$ 之间的欧氏距离的平方,N_H 和 N_K 分别表示 H 和 K 类中的样本数目。

如果 K 类由 I 类和 J 类合并产生,则可以得到 H 和 K 类之间距离的递推式为

$$D_{HK}=\sqrt{\frac{N_I}{N_I+N_J}D_{HI}^2+\frac{N_J}{N_I+N_J}D_{HJ}^2} \tag{5-24}$$

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

【例 5.3】 已知 6 个五维模式样本 $\boldsymbol{X}_1$、$\boldsymbol{X}_2$、$\boldsymbol{X}_3$、$\boldsymbol{X}_4$、$\boldsymbol{X}_5$ 和 $\boldsymbol{X}_6$,试按最短距离法进行聚类分类。

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

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

【解】 对每一样本可表示为

$$\boldsymbol{X}_1=[x_{11},x_{12},x_{13},x_{14},x_{15}]^{\mathrm{T}},\boldsymbol{X}_2=[x_{21},x_{22},x_{23},x_{24},x_{25}]^{\mathrm{T}},\cdots,$$

$$\boldsymbol{X}_6=[x_{61},x_{62},x_{63},x_{64},x_{65}]^{\mathrm{T}}$$

(1) 将每一样本看成单独一类,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:

$$\begin{aligned}D_{12}(0)&=\|\boldsymbol{X}_1-\boldsymbol{X}_2\|=[(x_{11}-x_{21})^2+(x_{12}-x_{22})^2+(x_{13}-x_{23})^2\\&\quad+(x_{14}-x_{24})^2+(x_{15}-x_{25})^2]^{\frac{1}{2}}\\&=[1+0+1+1+0]^{\frac{1}{2}}=\sqrt{3}\end{aligned}$$

同理可求得

$D_{13}(0),D_{14}(0),D_{15}(0),D_{16}(0);D_{21}(0),D_{22}(0),D_{23}(0),D_{24}(0),D_{25}(0),D_{26}(0),\cdots$

距离矩阵 $\boldsymbol{D}(0)$可由表格表示,见表 5-1。

表 5-1　距离矩阵 $\boldsymbol{D}(0)$的表格形式

$\boldsymbol{D}(0)$	$G_1(0)$	$G_2(0)$	$G_3(0)$	$G_4(0)$	$G_5(0)$	$G_6(0)$
$G_1(0)$	0					
$G_2(0)$	$\sqrt{3}$ *	0				

续表

$\boldsymbol{D}(0)$	$G_1(0)$	$G_2(0)$	$G_3(0)$	$G_4(0)$	$G_5(0)$	$G_6(0)$
$G_3(0)$	$\sqrt{15}$	$\sqrt{6}$	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$G_2(0)$合并为一类，得到新的分类

$$G_{12}(1)=\{G_1(0),G_2(0)\}\quad G_3(1)=\{G_3(0)\}$$

$$G_4(1)=\{G_4(0)\}\quad G_5(1)=\{G_5(0)\}\quad G_6(1)=\{G_6(0)\}$$

按最小距离准则计算类间距离，由$\boldsymbol{D}(0)$矩阵递推得到聚类后的距离矩阵$\boldsymbol{D}(1)$也可由表格表示，见表 5-2。

表 5-2　第 1 次合并后的距离矩阵 $D(1)$的表格形式

$\boldsymbol{D}(1)$	$G_{12}(1)$	$G_3(1)$	$G_4(1)$	$G_5(1)$	$G_6(1)$
$G_{12}(1)$	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$\sqrt{4}$对应的类合并为一类，得$\boldsymbol{D}(2)$，其可由表格表示，见表 5-3。

表 5-3　第 2 次合并后的距离矩阵 $D(2)$的表格形式

$\boldsymbol{D}(2)$	$G_{12}(2)$	$G_3(2)$	$G_4(2)$	$G_{56}(2)$
$G_{12}(2)$	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$\sqrt{5}$对应的类合并为一类，得$\boldsymbol{D}(3)$，其可由表格表示，见表 5-4。

表 5-4　第 3 次合并后的距离矩阵 $D(3)$的表格形式

$\boldsymbol{D}(3)$	$G_{124}(3)$	$G_3(3)$	$G_{56}(2)$
$G_{124}(3)$	0		
$G_3(3)$	$\sqrt{6}$	0	
$G_{56}(2)$	$\sqrt{7}$	$\sqrt{6}$	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，结果为

$$G_1=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\boldsymbol{X}_4\}\quad G_2=\{\boldsymbol{X}_2\}\quad G_3=\{\boldsymbol{X}_5,\boldsymbol{X}_6\}$$

若无阈值，继续聚类下去，最终全部样本归为一类，这时给出聚类过程的树状表示，如

图 5-5 所示。类间距离阈值增大，分类变粗。

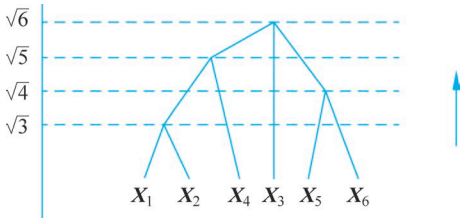


图 5-5　分级聚类法的树状表示

5.3　聚类准则函数

样本相似性度量是聚类分析的基础，针对具体问题，选择适当的相似性度量是保证聚类效果的基础。但有了相似性度量还不够，还必须有适当的聚类准则函数，才能把真正属于同一类的样本聚合成一个类别的子集，把不同类的样本分离开来。因此，聚类准则函数对聚类质量也有重要影响。相似性度量用于解决集合与集合的相似性问题；相似性准则用于来评价分类效果的好坏。如果聚类准则函数选得好，聚类质量就会高。同时，聚类准则函数还可以用来评价一种聚类结果的质量，如果聚类质量不满足要求，就要重复执行聚类过程，以优化结果。在重复优化中，可以改变相似性度量的方法，也可以选用新的聚类准则。

5.3.1　误差平方和准则

给定样本集$\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$依据某种相似性测度划分为$c$类$\{\omega_1,\omega_2,\cdots,\omega_c\}$，定义误差平方和准则函数为

$$J=\sum_{i=1}^{c}\sum_{\boldsymbol{X}_i\in\omega_i}\|\boldsymbol{X}_i-\boldsymbol{M}_i\|^2$$

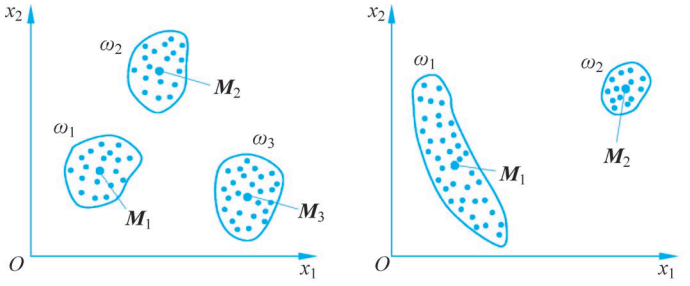
式中，$\boldsymbol{M}_i=\dfrac{1}{N_i}\sum_{\boldsymbol{X}_i\in\omega_i}\boldsymbol{X}_i$为属于$\omega_i$集的样本的均值向量，$N_i$为$\omega_i$中样本数目。

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。在此准则函数下，聚类的目标转化为使J取最小值，即聚类的结果应使全部样本与其相应模式均值之间的误差平方和最小。该准则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；当类别样本数相差较大，且类间距离又不足够大时，并不适宜采用该准则函数。因为可能会由于样本数较多造成类中的边缘处样本距离另一类的类心更近，从而产生错误的划分。

如图 5-6(a)所示，类内误差平方和很小，类间距离很远，可得到较好的结果。图 5-6(b)中类长轴两端距离中心很远，J值较大，结果不易令人满意。在该准则下，有时可能把样本数目多的一类分拆为二，造成错误聚类。如图 5-7(a)和(b)的正确分类与错误分类情况所示，ω_1中的某些样本被错分到ω_2中，因为这样分类，J值会更小。

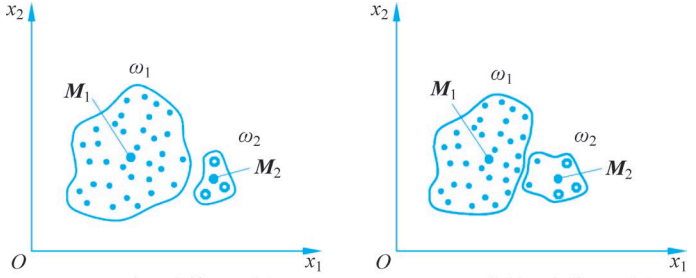
5.3.2　加权平均平方距离和准则

误差平方和准则只是考虑了各样本到判定类心的距离，并没有考虑样本周围空间其他



(a) 样本分类良好　　(b) 样本分类较差

图 5-6　样本分布



(a) 正确分类情况示例　　(b) 错误分类情况示例

图 5-7　正确分类与错误分类示例

样本对聚类的影响，当综合考虑这些因素时，误差平方和准则可改进为加权平均平方距离和准则。加权平均平方距离和准则函数定义为

$$J=\sum_{i=1}^{c}\frac{N_i}{N}\overline{D}_i^2 \tag{5-25}$$

式中，

$$\overline{D}_i^2=\frac{2}{N_i(N_i-1)}\sum_{\substack{x_{ik}\in\omega_i\\x_{ij}\in\omega_i}}\|x_{ik}-x_{ij}\|^2 \tag{5-26}$$

$\sum\limits_{\substack{x_{ik}\in\omega_i\\x_{ij}\in\omega_i}}\|x_{ik}-x_{ij}\|^2$ 表示 ω_i 类中任意两个不同样本距离平方和，由于 ω_i 中包含样本的个数为 N_i，因此共有 $\dfrac{N_i(N_i-1)}{2}$ 个组合，由此可见，$\overline{D}_i^2$ 的含义是类内任意两样本的平均平方距离。N 为样本总数，$\dfrac{N_i}{N}$ 为 ω_i 类的先验概率，因此 J 被称为加权平均平方距离和准则。

5.3.3　类间距离和准则

给定待分样本 $\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$，将它们分成 c 类 $\{\omega_1,\omega_2,\cdots,\omega_c\}$，定义 ω_i 类的样本均值向量 $\boldsymbol{M}_i$ 和总体样本均值向量 $\boldsymbol{M}$ 为

$$\boldsymbol{M}_i=\frac{1}{N_i}\sum_{\boldsymbol{X}_i\in\omega_i}\boldsymbol{X}_i \tag{5-27}$$

$$\boldsymbol{M}=\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{M}_i \tag{5-28}$$

则类间距离和准则定义为

$$J=\sum_{i=1}^{c}(\boldsymbol{M}_i-\boldsymbol{M})^{\mathrm{T}}(\boldsymbol{M}_i-\boldsymbol{M}) \tag{5-29}$$

聚类的目标是最大化式(5-29)，J 越大表示各类之间的可分离性越好，聚类效果越好。

对于二分类问题，类间距离常用式(5-30)表示类间距离。

$$J=(\boldsymbol{M}_1-\boldsymbol{M}_2)^{\mathrm{T}}(\boldsymbol{M}_1-\boldsymbol{M}_2) \tag{5-30}$$

5.3.4　离散度矩阵

给定待分样本$\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$，将它们分成 c 类$\{\omega_1,\omega_2,\cdots,\omega_c\}$。定义 ω_i 类的离散度矩阵为

$$\boldsymbol{S}_i=\sum_{\boldsymbol{X}_i\in\omega_i}(\boldsymbol{X}_i-\boldsymbol{M}_i)(\boldsymbol{X}_i-\boldsymbol{M}_i)^{\mathrm{T}} \tag{5-31}$$

定义类内离散度矩阵为

$$\boldsymbol{S}_{\mathrm{w}}=\sum_{i=1}^{c}\boldsymbol{S}_i \tag{5-32}$$

定义类间离散度矩阵为

$$\boldsymbol{S}_{\mathrm{b}}=\sum_{i=1}^{c}N_i(\boldsymbol{M}_i-\boldsymbol{M})(\boldsymbol{M}_i-\boldsymbol{M})^{\mathrm{T}} \tag{5-33}$$

定义总体离散度矩阵为

$$\boldsymbol{S}_{\mathrm{t}}=\sum_{i=1}^{N}(\boldsymbol{X}_i-\boldsymbol{M})(\boldsymbol{X}_i-\boldsymbol{M})^{\mathrm{T}} \tag{5-34}$$

可以证明 $\boldsymbol{S}_{\mathrm{t}}=\boldsymbol{S}_{\mathrm{w}}+\boldsymbol{S}_{\mathrm{b}}$，聚类的目标是极大化 $\boldsymbol{S}_{\mathrm{b}}$ 和极小化 $\boldsymbol{S}_{\mathrm{w}}$，即使不同类的样本尽可能分开，而同一类的样本尽可能聚集，由此可定义如下基于离散度矩阵的 4 个聚类准则。

$$J_1=\mathrm{tr}[\boldsymbol{S}_{\mathrm{w}}^{-1}\boldsymbol{S}_{\mathrm{b}}] \tag{5-35}$$

$$J_2=|\boldsymbol{S}_{\mathrm{w}}^{-1}\boldsymbol{S}_{\mathrm{b}}| \tag{5-36}$$

$$J_3=\mathrm{tr}[\boldsymbol{S}_{\mathrm{w}}^{-1}\boldsymbol{S}_{\mathrm{t}}] \tag{5-37}$$

$$J_4=|\boldsymbol{S}_{\mathrm{w}}^{-1}\boldsymbol{S}_{\mathrm{t}}| \tag{5-38}$$

式中，tr(·)为求矩阵的迹，即方阵主对角线上各元素之和，也就是矩阵的特征值的总和。

聚类的目标是极大化 $J_i(i=1,2,3,4)$，J_i 越大表示各类之间的可分离性越好，聚类质量越好。当待分样本特征向量的维数为 d 时，$\boldsymbol{S}_{\mathrm{w}}^{-1}\boldsymbol{S}_{\mathrm{b}}$ 则为 $d\times d$ 的对称矩阵，其对应的特征值为 $\lambda_i(i=1,2,\cdots,d)$，易知

$$J_1=\sum_{i=1}^{d}\lambda_i \tag{5-39}$$

$$J_2=\prod_{i=1}^{d}\lambda_i \tag{5-40}$$

$$J_3=\sum_{i=1}^{d}(1+\lambda_i) \tag{5-41}$$

$$J_4=\prod_{i=1}^{d}(1+\lambda_i) \tag{5-42}$$

因此，在实际运算中，只要求出 $\boldsymbol{S}_{\mathrm{w}}^{-1}\boldsymbol{S}_{\mathrm{b}}$ 的特征值，即可求得 $J_i(i=1,2,3,4)$。

5.4 基于距离阈值的聚类算法

当确定了相似性测度和准则函数后，聚类的过程是依靠聚类算法来实现的。因此，聚类算法是一个试图识别数据集合聚类的特殊性质的学习过程。本节介绍两种简单的聚类分析方法，它通过对某些关键性的元素进行试探性的选取，使某种聚类准则达到最优，又称为基于试探的聚类算法。

5.4.1 最近邻规则的聚类算法

最近邻规则聚类分析问题描述为：假设已有混合样本集 $\boldsymbol{X}^{(N)}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$，给定类内距离门限阈值 T，将 $\boldsymbol{X}^{(N)}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$ 划分为 $\omega_1,\omega_2,\cdots,\omega_c$ 个类别。

最近邻规则聚类算法的基本思想是：计算样本的特征向量到聚类中心的距离，将该距离与门限阈值 T 比较，决定该样本属于哪一类别或作为新一类别的中心。

第 13 集
微课视频

第 14 集
微课视频

按照最近邻原则进行聚类，算法步骤如下。

(1) 选取距离阈值 T，并且任取一个样本作为第一个聚合中心 $\boldsymbol{Z}_1$，如 $\boldsymbol{Z}_1=\boldsymbol{X}_1$。

(2) 计算样本 $\boldsymbol{X}_2$ 到 $\boldsymbol{Z}_1$ 的距离 D_{21}：

若 $D_{21}\leqslant T$，则 $\boldsymbol{X}_2\in\boldsymbol{Z}_1$，否则令 $\boldsymbol{X}_2$ 为第二个聚合中心，即 $\boldsymbol{Z}_2=\boldsymbol{X}_2$。

设 $\boldsymbol{Z}_2=\boldsymbol{X}_2$，计算 $\boldsymbol{X}_3$ 到 $\boldsymbol{Z}_1$ 和 $\boldsymbol{Z}_2$ 的距离 D_{31} 和 D_{32}，若 $D_{31}>T$ 和 $D_{32}>T$，则建立第三个聚合中心 $\boldsymbol{Z}_3$。否则把 $\boldsymbol{X}_3$ 归于最近邻的聚合中心。依此类推，直到把所有的 N 个样本都进行分类。

(3) 按照某种聚类准则考查聚类结果，若不满意，则重新选取距离阈值 T 和第一个聚合中心 $\boldsymbol{Z}_1$，返回(2)，直到满意，算法结束。

该算法的优点是简单，如果有样本分布的先验知识用于指导阈值和起始点的选取，则可较快得到合理结果。其缺点是聚类过程中类别的中心一经选定，在聚类过程中将不再改变。同样，样本一经判定类别归属后也不再改变。因此，在样本分布一定时，该算法的结果在很大程度上取决于第一个聚合中心的选取和距离阈值的大小的确定。对于高维的样本集来说，只有经过多次试探，并对聚类结果进行检验，才能选择最优的聚类结果。

5.4.2 最大最小距离聚类算法

最大最小距离聚类算法的问题描述为：假设已有混合样本集 $\boldsymbol{X}^{(N)}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$，给比例系数 θ，将 $\boldsymbol{X}^{(N)}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$ 划分为 $\omega_1,\omega_2,\cdots,\omega_c$ 个类别。该算法的基本思想是：样本的特征向量以最大距离原则选取新的聚类中心，以最小距离原则进行类别归属。如果使用欧氏距离，除首先辨识最远的聚类中心外，其余步骤与最近邻规则算法相似。用一个例子说明该算法。

【例 5.4】　已知二类共 10 个样本，分布如图 5-8 所示。

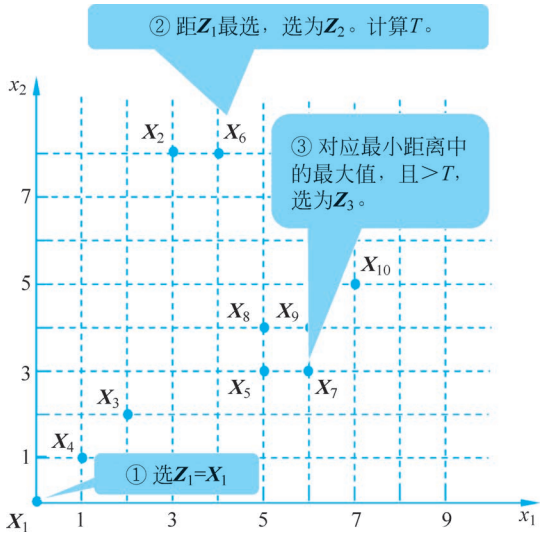


图 5-8　例 5.4 样本分布

其中，$\boldsymbol{X}_1=[0,0]^{\mathrm{T}}$，$\boldsymbol{X}_2=[3,8]^{\mathrm{T}}$，$\boldsymbol{X}_3=[2,2]^{\mathrm{T}}$，$\boldsymbol{X}_4=[1,1]^{\mathrm{T}}$，$\boldsymbol{X}_5=[5,3]^{\mathrm{T}}$，$\boldsymbol{X}_6=[4,8]^{\mathrm{T}}$，$\boldsymbol{X}_7=[6,3]^{\mathrm{T}}$，$\boldsymbol{X}_8=[5,4]^{\mathrm{T}}$，$\boldsymbol{X}_9=[6,4]^{\mathrm{T}}$，$\boldsymbol{X}_{10}=[7,5]^{\mathrm{T}}$，采用最大最小距离聚类算法求出分类结果。

【解】　采用最大最小距离聚类算法求解的步骤如下。

（1）给定 θ，$0<\theta<1$，并且任取一个样本作为第一个聚合中心，$\boldsymbol{Z}_1=\boldsymbol{X}_1$。

（2）寻找新的集合中心。

计算其他所有样本到 $\boldsymbol{Z}_1$ 的距离 D_{i1}，若 $D_{k1}=\max\limits_{i}\{D_{i1}\}$，则取 $\boldsymbol{X}_k$ 为第二个聚合中心 $\boldsymbol{Z}_2$，如本例中 $\boldsymbol{Z}_2=\boldsymbol{X}_6$。计算所有样本到 $\boldsymbol{Z}_1$ 和 $\boldsymbol{Z}_2$ 的距离 D_{i1} 和 D_{i2}，若 $D_l=\max\{\min(D_{i1},D_{i2})\}$，$i=1,2,\cdots,n$，并且 $D_l>\theta\cdot D_{12}$，D_{12} 为 $\boldsymbol{Z}_1$ 和 $\boldsymbol{Z}_2$ 间距离，则取 $\boldsymbol{X}_l$ 为第三个集合中心 $\boldsymbol{Z}_3$，本例中 $\boldsymbol{Z}_3=\boldsymbol{X}_7$。如果 $\boldsymbol{Z}_3$ 存在，则计算 $D_j=\max\{\min(D_{i1},D_{i2},D_{i3})\}$，$i=1,2,\cdots,n$，若 $D_j>\theta\cdot D_{12}$，则建立第四个聚合中心。以此类推，直到最大最小距离不大于 $\theta\cdot D_{12}$ 时，结束寻找聚合中心的计算。

观察表 5-5，当取 $\theta=0.5$ 时，$\sqrt{29}$ 在 $\min(D_{i1},D_{i2})$ 中为最大的，而且 $D_l=\sqrt{29}>\theta\cdot\sqrt{80}$。所以，$\boldsymbol{Z}_3=\boldsymbol{X}_7$。

表 5-5　判断第三个聚类中心样本距离计算表

距　离	样　本									
	$\boldsymbol{X}_1$	$\boldsymbol{X}_2$	$\boldsymbol{X}_3$	$\boldsymbol{X}_4$	$\boldsymbol{X}_5$	$\boldsymbol{X}_6$	$\boldsymbol{X}_7$	$\boldsymbol{X}_8$	$\boldsymbol{X}_9$	$\boldsymbol{X}_{10}$
到 $\boldsymbol{Z}_1$ 的距离	0	$\sqrt{73}$	$\sqrt{8}$	$\sqrt{2}$	$\sqrt{34}$	$\sqrt{80}$	$\sqrt{45}$	$\sqrt{41}$	$\sqrt{52}$	$\sqrt{74}$
到 $\boldsymbol{Z}_2$ 的距离	$\sqrt{80}$	1	$\sqrt{40}$	$\sqrt{58}$	$\sqrt{26}$	0	$\sqrt{29}$	$\sqrt{17}$	$\sqrt{20}$	$\sqrt{18}$
$\min(D_{i1},D_{i2})$	0	1	$\sqrt{8}$	$\sqrt{2}$	$\sqrt{26}$	0	$\sqrt{29}$	$\sqrt{17}$	$\sqrt{20}$	$\sqrt{18}$

观察表 5-6，计算 $D_j=\max\{\min(D_{i1},D_{i2},D_{i3})\}$，$i=1,2,\cdots,n$，得 $D_j=\sqrt{8}<\theta\cdot\sqrt{80}$，结束寻找聚合中心。则图 5-8 中只有三个集合中心，$\boldsymbol{Z}_1=\boldsymbol{X}_1$，$\boldsymbol{Z}_2=\boldsymbol{X}_6$，$\boldsymbol{Z}_3=\boldsymbol{X}_7$。

表 5-6 判断第四个聚类中心样本距离计算表

距离	样本									
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
到 Z_1 的距离	0	$\sqrt{73}$	$\sqrt{8}$	$\sqrt{2}$	$\sqrt{34}$	$\sqrt{80}$	$\sqrt{45}$	$\sqrt{41}$	$\sqrt{52}$	$\sqrt{74}$
到 Z_2 的距离	$\sqrt{80}$	1	$\sqrt{40}$	$\sqrt{58}$	$\sqrt{26}$	0	$\sqrt{29}$	$\sqrt{17}$	$\sqrt{20}$	$\sqrt{18}$
到 Z_3 的距离	$\sqrt{45}$	$\sqrt{34}$	$\sqrt{17}$	$\sqrt{29}$	1	$\sqrt{29}$	0	$\sqrt{2}$	1	$\sqrt{5}$
$\min(D_{i1},D_{i2},D_{i3})$	0	1	$\sqrt{8}$	$\sqrt{2}$	1	0	0	$\sqrt{2}$	1	$\sqrt{5}$

(3) 按最近邻原则把所有样本归属于距离最近的聚合中心，有

$$\{x_1,x_3,x_4\}\in \boldsymbol{Z}_1,\quad \{x_2,x_6\}\in \boldsymbol{Z}_2,\quad \{x_5,x_7,x_8,x_9,x_{10}\}\in \boldsymbol{Z}_3$$

(4) 按照某聚类准则考查聚类结果，若不满意，则重选 θ 和第一个聚合中心 $\boldsymbol{Z}_1$，返回(2)，直到满意，算法结束。

从上述步骤可以看出，该算法的聚类结果与参数 θ 和起始点 $\boldsymbol{Z}_1$ 的选取密切相关。若无先验样本分布知识，则只有用试探法通过多次试探优化，选择最合理的一种参数选择方案和聚类结果。若有先验知识用于指导 θ 和 $\boldsymbol{Z}_1$ 选取，则算法可以很快收敛。

5.5 动态聚类算法

第 15 集
微课视频

最近邻规则和最大、最小距离聚类算法的共同缺点是：一个样本的归属一旦判定后，在后继的迭代过程中就不会改变，因此，这类算法在实际应用中有较大的局限性。与上述算法相对，动态聚类法是聚类分析中较普遍采用的方法。该算法首先选择某种样本相似性度量和适当的聚类准则函数，在对样本进行初始划分的基础上，使用迭代算法逐步优化聚类结果，当准则函数达到极值时，取得在该准则函数下的最优聚类结果。

该算法有以下两个关键问题。

(1) 首先选择有代表性的点作为起始聚合中心。若类别数目已知，则选择代表点的数目等于类别数目；若类别数未知，那么聚类过程如何形成的类别数目是一个值得研究的问题。

(2) 代表点选择好之后，如何形成初始划分是算法的另一个关键问题。

5.5.1 C 均值聚类算法

给定模式样本集 $\boldsymbol{X}^{(N)}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$，假设已知样本的类别数为 c。C 均值聚类算法的基本思想是：先选定 c 个初始聚类中心，按最短距离原则将各样本归属到 c 个类别中，然后重新计算各类别中心，调整各样本的类别归属，算法不断迭代，最终使各样本到其类别中心的距离平方和最小。

C 均值聚类算法使用的聚类准则函数是误差平方和准则 J_c，即

$$J_c=\sum_{j=1}^{c}\sum_{k=1}^{N_j}\|\boldsymbol{X}_k-m_j\|^2 \tag{5-43}$$

式中，N_j 为第 j 类的样本数，m_j 为第 j 类 ω_j 的均值。

为了使聚类结果优化，应该使准则 J_c 最小化。下面给出 C 均值算法的具体步骤。

(1) 已知混合样本 $\boldsymbol{X}^{(N)}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$，令 I 表示迭代运算次数，任选 c 个样本作为初始聚合中心 $\boldsymbol{Z}_j(I),j=1,2,\cdots,c$。

(2) 计算每个样本与聚合中心的距离 $D(\boldsymbol{X}_k,\boldsymbol{Z}_j(I)),k=1,2,\cdots,n,j=1,2,\cdots,c$。若

$$D(\boldsymbol{X}_k,\boldsymbol{Z}_j(I))=\min_{j=1,2,\cdots,c}\{D(\boldsymbol{X}_k,\boldsymbol{Z}_j(I)),k=1,2,\cdots,n\} \tag{5-44}$$

则 $\boldsymbol{X}_k\in\omega_j$。

(3) 重新计算 c 个新的集合的聚类中心，即

$$\boldsymbol{Z}_j(I+1)=\frac{1}{N_j}\sum_{k=1}^{N_j}\boldsymbol{X}_k^{(j)},\quad j=1,2,\cdots,c \tag{5-45}$$

式中，N_j 为第 $I+1$ 次迭代时归属于 ω_j 类的样本数，$\boldsymbol{X}_k^{(j)}$ 为第 $I+1$ 次迭代时归属于 ω_j 类的样本，上标表示类别。

(4) 若 $\boldsymbol{Z}_j(I+1)\neq\boldsymbol{Z}_j(I),j=1,2,\cdots,c$，则 $I=I+1$，返回(2)，否则算法结束。

C 均值聚类算法特点是：①每次迭代中都要重新计算聚类中心，并考查每个样本类别归属是否正确，若不正确，就要调整，在全部样本调整完之后，进入下一次迭代。如果在某一个迭代运算中，所有的样本都被正确分类，则样本不会调整，聚合中心也不会有变化，算法达到收敛。②算法需要首先确定类别数和初始聚类中心，显然，类别数 c 和初始聚合中心的选择对聚类结果有较大影响，所以结果不是全局最优。③算法简单便于实现，当样本分布为类内团状时，一般可以达到比较好的聚类效果。

从算法的步骤可以看出，算法在迭代中没有计算 J_c 值，也就是说 J_c 不是算法结束的明显依据。算法通过对样本分类的不断调整去逐步减少 J_c 的值，当没有样本调整时，此时 J_c 不再变化，聚类达到最优。事实上，可以通过样本移动对 J_c 的影响来修改上述算法。假定 $I+1$ 次迭代时，$\boldsymbol{X}_k$ 由样本子集 $\boldsymbol{X}^{(i)}$ 移入另一个子集 $\boldsymbol{X}^{(j)}$，那么这次移动只影响两个类型 ω_i 和 ω_j 的聚类中心 $\boldsymbol{Z}_i$ 和 $\boldsymbol{Z}_j$，以及两类的类内误差平方和 J_{c_i}、J_{c_j}。移动后，ω_i 和 ω_j 的聚类中心

$$\boldsymbol{Z}_i(I+1)=\frac{1}{N_i-1}[N_i\boldsymbol{Z}_i(I)-\boldsymbol{X}_k]=\boldsymbol{Z}_i(I)+\frac{1}{N_i-1}[\boldsymbol{Z}_i(I)-\boldsymbol{X}_k] \tag{5-46}$$

$$\boldsymbol{Z}_j(I+1)=\frac{1}{N_j+1}[N_j\boldsymbol{Z}_j(I)+\boldsymbol{X}_k]=\boldsymbol{Z}_j(I)-\frac{1}{N_j+1}[\boldsymbol{Z}_j(I)-\boldsymbol{X}_k] \tag{5-47}$$

因此，有

$$J_{c_i}(I+1)=J_{c_1}(I)-\frac{N_i}{N_i-1}\|\boldsymbol{X}_k-\boldsymbol{Z}_i(I)\|^2 \tag{5-48}$$

$$J_{c_j}(I+1)=J_{c_j}(I)+\frac{N_i}{N_i+1}\|\boldsymbol{X}_k-\boldsymbol{Z}_j(I)\|^2 \tag{5-49}$$

由于样本 $\boldsymbol{X}_k$ 从 $\boldsymbol{X}^{(i)}$ 移入 $\boldsymbol{X}^{(j)}$，显然 $\boldsymbol{X}_k$ 距 $\boldsymbol{Z}_j(I)$ 比 $\boldsymbol{Z}_i(I)$ 更近，因此有

$$\frac{N_i}{N_i+1}\|\boldsymbol{X}_k-\boldsymbol{Z}_j(I)\|^2<\frac{N_i}{N_i-1}\|\boldsymbol{X}_k-\boldsymbol{Z}_i(I)\|^2 \tag{5-50}$$

那么，J_c 的值会减小为

$$J_c(I+1)=J_c(I)-\left[\frac{N_i}{N_i-1}\|\boldsymbol{X}_k-\boldsymbol{Z}_i(I)\|^2-\frac{N_i}{N_i+1}\|\boldsymbol{X}_k-\boldsymbol{Z}_j(I)\|^2\right] \tag{5-51}$$

根据上述分析，可对 C 均值算法做如下改进：

(1) 已知混合样本 $\boldsymbol{X}^{(N)}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$，令 I 表示迭代运算次数，任选 c 个样本作为初始聚合中心 $\boldsymbol{Z}_j(I),j=1,2,\cdots,c$；

(2) 计算每个样本与聚合中心的距离 $D(\boldsymbol{X}_k,\boldsymbol{Z}_j(I)),k=1,2,\cdots,N,j=1,2,\cdots,c$。若

$$D(\boldsymbol{X}_k,\boldsymbol{Z}_j(I))=\min_{j=1,2,\cdots,c}\{D(\boldsymbol{X}_k,\boldsymbol{Z}_j(I)),k=1,2,\cdots,N\}$$

则 $\boldsymbol{X}_k\in\omega_i$。

(3) 令 $I=I+1=2$，计算新的类别中心

$$\boldsymbol{Z}_j(2)=\frac{1}{N_j}\sum_{k=1}^{N_j}\boldsymbol{X}_k^{(j)},\quad j=1,2,\cdots,c \tag{5-52}$$

计算本次迭代的误差平方和 J_c，即

$$J_c=\sum_{j=1}^{c}\sum_{k=1}^{N_j}\|\boldsymbol{X}_k^{(j)}-\boldsymbol{Z}_j(2)\|^2 \tag{5-53}$$

(4) 对每个类别中的每个样本，计算 ρ_{ii}（J_c 减少的部分）和 ρ_{ij}（J_c 增加的部分）。

$$\rho_{ii}=\frac{N_i}{N_i-1}\|\boldsymbol{X}_k^{(i)}-\boldsymbol{Z}_i(I)\|^2,\quad i=1,2,\cdots,c \tag{5-54}$$

$$\rho_{ij}=\frac{N_j}{N_j+1}\|\boldsymbol{X}_k^{(i)}-\boldsymbol{Z}_j(I)\|^2,\quad i=1,2,\cdots,c\quad i\neq j \tag{5-55}$$

令

$$\rho_{il}=\min_{i\neq j}\{\rho_{ij}\} \tag{5-56}$$

若 $\rho_{il}<\rho_{ii}$，则把样本 $\boldsymbol{X}_k^{(i)}$ 移到聚合中心 ω_l 中，并修改聚合中心和 J_c 值，即有

$$\boldsymbol{Z}_i(I+1)=\boldsymbol{Z}_i(I)-\frac{1}{N_i-1}[\boldsymbol{Z}_i(I)-\boldsymbol{X}_k^{(i)}] \tag{5-57}$$

$$\boldsymbol{Z}_l(I+1)=\boldsymbol{Z}_l(I)+\frac{1}{N_l+1}[\boldsymbol{Z}_l(I)-\boldsymbol{X}_k^{(i)}] \tag{5-58}$$

$$J_c(I+1)=J_c(I)-(\rho_{ii}-\rho_{il}) \tag{5-59}$$

(5) 若 $J_c(I+1)<J_c(I)$，则 $I=I+1$，返回(4)。否则，算法结束。

【例 5.5】 现有混合样本集 $\boldsymbol{X}^{(N)}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_{20}\}$，共有样本 20 个，样本分布如图 5-9 所示，类型数目 $c=2$。试用 C 均值聚类算法进行聚类分析。

【解】 (1) $c=2$，任选 2 个集合中心，不妨取 $\boldsymbol{Z}_1(1)=\boldsymbol{X}_1,\boldsymbol{Z}_2(1)=\boldsymbol{X}_2$，即 $\boldsymbol{Z}_1(1)=[0,0]^{\mathrm{T}},\boldsymbol{Z}_2(1)=[1,0]^{\mathrm{T}}$。

(2) 选用欧氏距离作为相似性度量，计算各样本到 $\boldsymbol{Z}_1(1)=\boldsymbol{X}_1,\boldsymbol{Z}_2(1)=\boldsymbol{X}_2$ 的距离，并把各样本归属于距离最小的聚类范围内，有

$$\|\boldsymbol{X}_1-\boldsymbol{Z}_1(1)\|<\|\boldsymbol{X}_1-\boldsymbol{Z}_2(1)\|,\text{因此 }\boldsymbol{X}_1\in\omega_1$$

$$\|\boldsymbol{X}_2-\boldsymbol{Z}_2(1)\|<\|\boldsymbol{X}_2-\boldsymbol{Z}_1(1)\|,\text{因此 }\boldsymbol{X}_2\in\omega_2$$

$$\|\boldsymbol{X}_3-\boldsymbol{Z}_1(1)\|<\|\boldsymbol{X}_3-\boldsymbol{Z}_2(1)\|,\text{因此 }\boldsymbol{X}_3\in\omega_1$$

$$\cdots$$

可得

$$\omega_1:\{\boldsymbol{X}_1,\boldsymbol{X}_3\},N_1=2$$

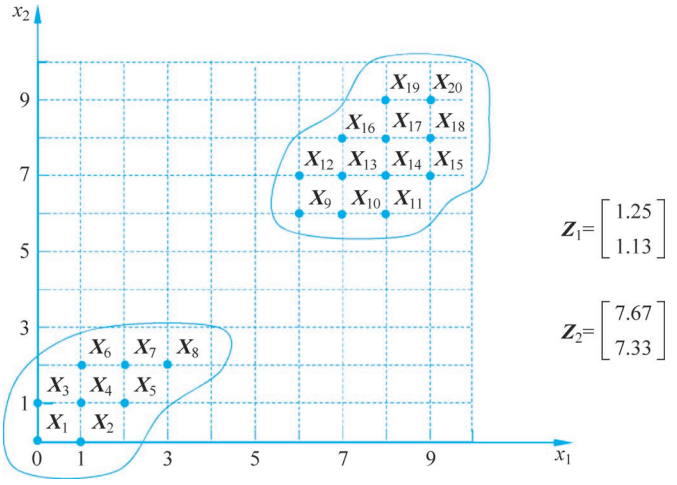


图 5-9　例 5.5 样本分布

$$\omega_2:\{\boldsymbol{X}_2,\boldsymbol{X}_4,\boldsymbol{X}_5,\cdots,\boldsymbol{X}_{20}\},N_2=18$$

(3) 计算新的聚类中心：

$$\boldsymbol{Z}_1(2)=\frac{1}{2}(\boldsymbol{X}_1+\boldsymbol{X}_3)=[0,0.5]^{\mathrm{T}}$$

$$\boldsymbol{Z}_2(2)=\frac{1}{18}(\boldsymbol{X}_2+\boldsymbol{X}_4+\boldsymbol{X}_5+\cdots+\boldsymbol{X}_{20})=[5.67,5.33]^{\mathrm{T}}$$

(4) $\boldsymbol{Z}_j(2)\neq\boldsymbol{Z}_j(1),j=1,2$。令 $I=I+1=2$，返回(2)。

计算各样本到 $\boldsymbol{Z}_j(2),j=1,2$ 的欧氏距离，有

$$\|\boldsymbol{X}_k-\boldsymbol{Z}_1(2)\|<\|\boldsymbol{X}_k-\boldsymbol{Z}_2(2)\|,\quad k=1,2,\cdots,8$$

$$\|\boldsymbol{X}_k-\boldsymbol{Z}_2(2)\|<\|\boldsymbol{X}_k-\boldsymbol{Z}_1(2)\|,\quad k=9,10,\cdots,20$$

得到新的聚类：

$$\omega_1:\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_8\},N_1=8$$

$$\omega_2:\{\boldsymbol{X}_9,\boldsymbol{X}_{10},\cdots,\boldsymbol{X}_{20}\},N_2=12$$

计算聚类中心：

$$\boldsymbol{Z}_1(3)=\frac{1}{8}(\boldsymbol{X}_1+\boldsymbol{X}_2+\cdots+\boldsymbol{X}_8)=[1.25,1.13]^{\mathrm{T}}$$

$$\boldsymbol{Z}_2(2)=\frac{1}{12}(\boldsymbol{X}_9+\boldsymbol{X}_{10}+\cdots+\boldsymbol{X}_{20})=[7.67,7.33]^{\mathrm{T}}$$

因为 $\boldsymbol{Z}_j(3)\neq\boldsymbol{Z}_j(2),j=1,2$。令 $I=I+1=3$，返回(2)。判断：$\boldsymbol{Z}_j(4)=\boldsymbol{Z}_j(3),j=1,2$，聚类结果无变化，聚类中心无变化，算法结束，聚类结果如图 5-9 所示。

(5) J_c 与 c 的关系曲线。上述 C 均值算法，其类别数假定已知为 c。对于类别数未知时，可以令 c 逐渐增加，如 $c=1,2,\cdots$，分别使用 C 均值算法，显然误差平方和 J_c 随 c 的增加而单调减少。最初，由于 c 较小，类别的分裂会使 J_c 迅速减小，但当 c 增加到一定数值时，相当于将本来就比较密集的类别再行分开，因此 J_c 的减小速度会减慢，直到 c 增加到总类别数目 N 时，$J_c=0$，J_c 与 c 的关系曲线如图 5-10 所示。

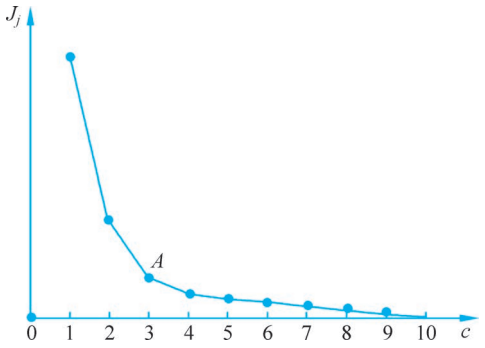


图 5-10 J_c 与 c 的关系曲线

在图 5-10 中，曲线的拐点 A 对应着接近最优的 c 值。但是并非所有的情况都容易找到 J_c 与 c 的关系曲线的拐点，此时 c 值将无法确定。下面介绍一种确定类型数目 c 的方法。

5.5.2 ISODATA 聚类算法

C 均值聚类算法的一个缺点是必须事先指定聚类的个数，在实际应用中有时并不可行，而是希望这个类别的个数也可以自动改变，于是形成了迭代自组织数据分析算法(Iterative Self-Organizing Data Analysis Techniques Algorithm，ISODATA)。ISODATA 是在 C 均值聚类算法基础上，通过增加对聚类结果的“合并”和“分裂”两个操作，并设置算法运行控制参数的一种聚类算法。ISODATA 可以通过类的自动合并(两类合一)与分裂(一类分为二)，得到较合理的类型数目，因此是目前应用比较广的一种聚类算法。

算法的基本思想：通过设定初始参数，并使用合并与分裂的机制，当某两类聚类中心距离小于某一阈值时，将它们合并为一类；当某类标准差大于某一阈值或其样本数目超过某一阈值时，将其分为两类。在某类样本数目少于某阈值时，需将其取消。如此，根据初始聚类中心和设定的类别数目等参数迭代，最终得到一个比较理想的分类结果。

具体算法步骤如下。

(1) 给定控制参数。

K：预期的聚类中心数目。

θ_n：每一聚类中最少的样本数目，如果少于此数就不能作为一个独立的聚类。

θ_s：一个聚类域中样本距离分布的标准差(阈值)。

θ_c：两个聚类中心之间的最小距离，如果小于此数，两个聚类合并。

L：每次迭代允许合并的最大聚类对数目。

I：允许的最多迭代次数。

给定 N 个混合样本 $\boldsymbol{X}^{(N)}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$，令迭代次数 $J=1$，任选 c 个样本作为初始聚合中心 $\boldsymbol{Z}_j(1),j=1,2,\cdots,c$。

(2) 计算每个样本与聚类中心距离 $D(\boldsymbol{X}_k,\boldsymbol{Z}_j(1)),k=1,2,\cdots,N,j=1,2,\cdots,c$。若

$$D(\boldsymbol{X}_k,\boldsymbol{Z}_j(1))=\min_{j=1,2,\cdots,c}\{D(\boldsymbol{X}_k,\boldsymbol{Z}_j(1)),k=1,2,\cdots,N\} \tag{5-60}$$

则 $\boldsymbol{X}_k\in\omega_j$。把所有样本都归属到 c 个聚类中去，N_j 表示类别 ω_j 中的样本数目。

(3) 若 $N_j<\theta_n,j=1,2,\cdots,c$，则舍去子集 ω_j，$c=c-1$，返回(2)。

(4) 计算修改聚合中心：

$$\boldsymbol{Z}_j(J)=\frac{1}{N_j}\sum_{k=1}^{n_j}\boldsymbol{X}_k^{(j)},\quad j=1,2,\cdots,c \tag{5-61}$$

(5) 计算类内距离平均值：

$$\overline{D}_j=\frac{1}{N_j}\sum_{k=1}^{N_j}D(\boldsymbol{X}_k^{(j)},\boldsymbol{Z}_j(J)),\quad j=1,2,\cdots,c \tag{5-62}$$

(6) 计算类内总平均距离 $\overline{D}$(全部样本对其相应聚类中心的总平均距离)：

$$\overline{D}=\frac{1}{N}\sum_{j=1}^{N_j}N_j\overline{D}_j \tag{5-63}$$

(7) 判别分裂、合并及迭代运算等步骤。

① 如果迭代运算次数已达 I 次，即最后一次迭代，置 $\theta_c=0$，跳到(11)。

② 如果 $c\leqslant\frac{K}{2}$，即聚类中心的数目等于或不到规定值的一半，则进入(8)，将已有的聚类分裂。

③ 如果迭代运算的次数是偶数，或 $c>2K$，则不进行分裂，跳到(11)，若不符合上述两个条件，则进入(8)进行分裂处理。

(8) 计算每个聚类的标准偏差向量 $\boldsymbol{\sigma}_j=(\sigma_{j1},\sigma_{j2},\cdots,\sigma_{jd})$。

每个分量为

$$\sigma_{ji}=\sqrt{\frac{1}{N_j}\sum_{x_{ji}\in\omega_j}(x_{ji}-z_{ji}(J))^2},\quad i=1,2,\cdots,d,j=1,2,\cdots,c \tag{5-64}$$

式中，x_{ji} 表示 $\boldsymbol{X}_j$ 的第 i 个分量，z_{ji} 表示 $\boldsymbol{Z}_j$ 的第 i 个分量，d 为样本特征向量维数。

(9) 求出每个聚类的最大分量：

$$\sigma_{j\max}=\max_{j=1,2,\cdots,c}\{\sigma_{ji}\},\quad j=1,2,\cdots,c \tag{5-65}$$

(10) 考查 $\sigma_{j\max}(j=1,2,\cdots,c)$ 若有 $\sigma_{j\max}>\theta_s$，并同时满足以下两条件之一：

① $\overline{D}_j>\overline{D}$ 及 $N_j>2(\theta_n+1)$(类内平均距离大于总类内平均距离，样本数目超过规定值一倍以上)。

② $c\leqslant\frac{K}{2}$。

则该聚类分裂成两个新的聚类，聚类中心分别为

$$\begin{cases}\boldsymbol{Z}_j^+(J)=\boldsymbol{Z}_j(J)+\boldsymbol{r}_j\\ \boldsymbol{Z}_j^-(J)=\boldsymbol{Z}_j(J)-\boldsymbol{r}_j\end{cases} \tag{5-66}$$

式中，$\boldsymbol{r}_j=\alpha\boldsymbol{\sigma}_j$ 或 $\boldsymbol{r}_j=\alpha[0,0,\cdots,\sigma_{j\max},\cdots,0,0]^{\mathrm{T}}$，$0<\alpha\leqslant1$。

令 $c=c+1$，$J=J+1$ 返回(2)。这里 α 的选择很重要，应使 $\boldsymbol{X}_j$ 中的样本到 $\boldsymbol{Z}_j^+(J)$ 和 $\boldsymbol{Z}_j^-(J)$ 的距离不同，但又使样本全部在这两个集合中。

(11) 计算任意两聚类中心间的距离：

$$D_{ij}=D[\boldsymbol{Z}_i(J),\boldsymbol{Z}_j(J)],\quad i=1,2,\cdots,c-1,j=1,2,\cdots,c \tag{5-67}$$

(12) 将 D_{ij} 与 θ_c 比较，并把小于 θ_c 的 D_{ij} 按递增次序排列，取前 L 个

$$D_{i_1j_1}<D_{i_2j_2}<\cdots<D_{i_Lj_L} \tag{5-68}$$

(13) 考查式(5-68),对每一个 $D_{i_l j_l}$($l=1,2,\cdots,L$),相应有两个聚类中心 $\boldsymbol{Z}_{i_l}(J)$ 和 $\boldsymbol{Z}_{j_l}(J)$,则把两类合并,合并后的聚类中心为

$$\boldsymbol{Z}_l(J)=\frac{1}{N_{i_l}+N_{j_l}}\left[N_{i_l}\boldsymbol{Z}_{i_l}(J)+N_{j_l}\boldsymbol{Z}_{j_l}(J)\right],\quad l=1,2,\cdots,L$$

$c=c$-已并掉的类数。

(14) 若 $J<I$,则 $J=J+1$,如果修改给定参数则返回(1),不修改参数返回(2),否则 $J=I$,算法结束。

在上述算法步骤中,第(8)~(10)步为分裂,第(11)~(13)步为合并,算法的合并与分裂条件可归纳如下。

(1) 合并条件:(类内样数$<\theta_n$)‖(类的数目$\geqslant 2K$)&&(两类间中心距离$<\theta_c$)。

(2) 分裂条件:$\left(\text{类的数目}\leqslant\frac{K}{2}\right)$&&(类的最大分量的标准差$>\theta_s$),即

$$(\sigma_{j\max}>\theta_s)\&\&\left[(\overline{D}_j>\overline{D})\&\&(N_j>2(\theta_n+1))\parallel\left(c\leqslant\frac{K}{2}\right)\right]$$

这里,‖ 表示“或”的关系,&& 表示“与”的关系。当类的数目满足 $\frac{K}{2}<c<2K$ 时,迭代运算的次数是偶数时合并,迭代运算的次数是奇次时分裂。

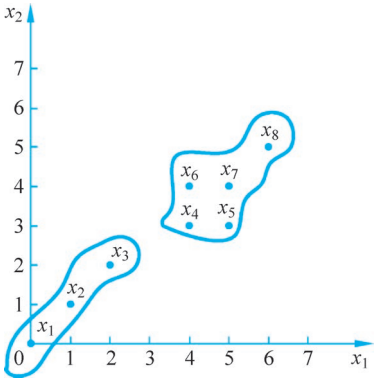


图 5-11 例 5.6 样本分布

【例 5.6】 有一混合样本集,其样本分布如图 5-11 所示,试用 ISODATA 算法进行聚类分析。

【解】 如图 5-11 所示,样本数目 $N_1=8$,取类型数目初始值 $c=1$,执行 ISODATA 算法。

(1) 给定参数 $K=2,\theta_n=2,\theta_s=1,\theta_c=4,L=0,I=4$,预选 $\boldsymbol{X}_1$ 为聚类中心,即 $\boldsymbol{Z}_1=[0,0]^{\mathrm{T}}$,令迭代次数 $J=1$。参数可任意选取,然后在迭代过程中加以调整。

(2) 因只有一个聚合中心 $\boldsymbol{Z}_1=[0,0]^{\mathrm{T}}$,故 $\omega_1:\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_8\}$,$N_1=8$。

(3) 因 $N_1=8>\theta_n$,故没有子集舍弃。

(4) 计算新聚合中心:

$$\boldsymbol{Z}_1=\frac{1}{8}\sum_{\boldsymbol{X}_i\in\omega_i}\boldsymbol{X}_i=\left[\frac{1+2+4+4+5+5+6}{8},\frac{1+2+3+3+4+4+5}{8}\right]^{\mathrm{T}}$$

$$=[3.38,2.75]^{\mathrm{T}}$$

(5) 计算类内距离平均值:

$$\overline{D}_1=\frac{1}{N_1}\sum_{\boldsymbol{X}_i\in\omega_1}\parallel\boldsymbol{X}_i-\boldsymbol{Z}_1\parallel$$

$$=\frac{1}{8}\left[\sqrt{\left(\frac{27}{8}\right)^2+\left(\frac{22}{8}\right)^2}+\sqrt{\left(\frac{19}{8}\right)^2+\left(\frac{14}{8}\right)^2}+\sqrt{\left(\frac{11}{8}\right)^2+\left(\frac{6}{8}\right)^2}+\sqrt{\left(\frac{5}{8}\right)^2+\left(\frac{2}{8}\right)^2}+\right.$$

$$\left.\sqrt{\left(\frac{13}{8}\right)^2+\left(\frac{2}{8}\right)^2}+\sqrt{\left(\frac{5}{8}\right)^2+\left(\frac{10}{8}\right)^2}+\sqrt{\left(\frac{13}{8}\right)^2+\left(\frac{10}{8}\right)^2}+\sqrt{\left(\frac{21}{8}\right)^2+\left(\frac{18}{8}\right)^2}\right]$$

$$=2.26$$

(6) 计算类内总平均距离：

$$\overline{D}=\overline{D}_1=2.26$$

(7) 因不是最后一次迭代，且满足 $c=\frac{K}{2}$，进入(8)。

(8) 计算聚类 ω_1 中的标准偏差：

$$\sigma_{11}=\sqrt{\frac{1}{8}\sum_{\boldsymbol{X}_i\in\omega_1}(x_{i1}-z_{11})^2}$$

$$=\sqrt{\frac{1}{8}\left[\left(0-\frac{27}{8}\right)^2+\left(1-\frac{27}{8}\right)^2+\left(2-\frac{27}{8}\right)^2+\left(4-\frac{27}{8}\right)^2+\sqrt{\left(5-\frac{27}{8}\right)^2+\left(4-\frac{27}{8}\right)^2+\left(5-\frac{27}{8}\right)^2+\left(6-\frac{27}{8}\right)^2}\right]}$$

$$=\sqrt{3.98}=1.99$$

$$\sigma_{12}=\sqrt{\frac{1}{8}\left[\left(\frac{22}{8}\right)^2+\left(\frac{14}{8}\right)^2+\left(\frac{6}{8}\right)^2+\left(\frac{2}{8}\right)^2+\left(\frac{22}{8}\right)^2+\left(\frac{10}{8}\right)^2+\left(\frac{10}{8}\right)^2+\left(\frac{18}{8}\right)^2\right]}=1.56$$

$$\boldsymbol{\sigma}_1=[\sigma_{11},\sigma_{12}]^{\mathrm{T}}=[1.99,1.56]^{\mathrm{T}}$$

(9) $\boldsymbol{\sigma}_1$ 中的最大偏差分量为 $\sigma_{11}=1.99$，即 $\sigma_{1\max}=1.99$。

(10) 因为 $\sigma_{1\max}>\theta_s$，且 $c=\frac{K}{2}$。所以把 ω_1 分裂成两个子类，取 $\alpha=0.5$，则 $0.5\sigma_{1\max}\approx 1$，故新的聚类中心分别为

$$\boldsymbol{Z}_1^{+}=[3.38+1,2.75]^{\mathrm{T}}=[4.38,2.75]^{\mathrm{T}}$$

$$\boldsymbol{Z}_1^{-}=[3.38-1,2.75]^{\mathrm{T}}=[2.38,2.75]^{\mathrm{T}}$$

将 $\boldsymbol{Z}_1^{+}$ 和 $\boldsymbol{Z}_1^{-}$ 改写为 $\boldsymbol{Z}_1$ 和 $\boldsymbol{Z}_2$，令 $c=c+1$，$J=J+1=2$，返回(2)。

(2)′ 按最小距离原则，重新聚类，得

$$\omega_1:\{\boldsymbol{X}_4,\boldsymbol{X}_5,\boldsymbol{X}_6,\boldsymbol{X}_7,\boldsymbol{X}_8\},N_1=5$$

$$\omega_2:\{\boldsymbol{X}_1,\boldsymbol{X}_2,\boldsymbol{X}_3\},N_2=3$$

(3)′ 因 $N_j>\theta_n$，故无合并。

(4)′ 重新计算聚类中心：

$$\boldsymbol{Z}_1=\frac{1}{N_1}\sum_{\boldsymbol{X}_i\in\omega_1}\boldsymbol{X}_i=[4.8,3.8]^{\mathrm{T}}$$

$$\boldsymbol{Z}_2=\frac{1}{N_2}\sum_{\boldsymbol{X}_i\in\omega_2}\boldsymbol{X}_i=[1.00,1.00]^{\mathrm{T}}$$

(5)′ 计算类内距离平均值：

$$\overline{D}_1=\frac{1}{N_1}\sum_{\boldsymbol{X}_i\in\omega_1}\|\boldsymbol{X}_i-\boldsymbol{Z}_1\|=1.06$$

$$\overline{D}_2=\frac{1}{N_2}\sum_{\boldsymbol{X}_i\in\omega_2}\|\boldsymbol{X}_i-\boldsymbol{Z}_2\|=0.94$$

(6)′ 计算类内总平均距离：

$$\overline{D}=\frac{1}{N}\sum_{j=1}^{2}N_j\cdot\overline{D}_j=\frac{1}{8}(5\times1.06+3\times0.94)=1.02$$

(7)′ 因是偶次迭代,故跳到(11)。

(11) 计算两个聚类中心之间的距离:

$$D_{12}=\|\boldsymbol{Z}_1-\boldsymbol{Z}_2\|=4.72$$

(12)~(13):因 $D_{12}>\theta_c$,故聚类中心不合并。

(14) 因为不是最后一次迭代,令 $J=J+1=3$,考虑是否修改参数。由上面结果可知,已获得合理的类别数目,两类别中心间距离大于类内总平均距离,每个类别都有足够比例的样本数目,且两类样本数相差不大,因此不必修改控制参数,返回(2)。

第(2)″~(6)″步与上次迭代相同。

(7)″ 所列情况均不满足,继续执行。

(8)″ 计算两个聚合的标准偏差。

$$\boldsymbol{\sigma}_1=[0.75,0.75]^{\mathrm{T}},\quad \boldsymbol{\sigma}_1=[0.82,0.82]^{\mathrm{T}}$$

(9)″ $\sigma_{1\max}=0.75$,$\sigma_{2\max}=0.82$。

(10)″ 因为 $c=\dfrac{K}{2}$,且 N_1 和 N_2 均小于 $2(\theta_n+1)$,分裂条件不满足。继续执行(11)。

第(11)′~(13)′步与前一次迭代结果相同。

(14)′ 因 $J<I$,令 $J=J+1=4$,故无须修改控制参数,返回(2)。

第(2)‴~(6)‴步与前一次迭代相同。

(7)‴ 因为 $J=I$,是最后一次迭代,所以令 $\theta_c=0$,跳到(11)。

第(11)″~(13)″步与前一次迭代相同。

(14)″ 因 $J=I$,故聚类过程结束。

在 ISODATA 算法中,起始聚合中心的选取对聚类过程和结果都有较大影响,如果选择得好,则算法收敛快,聚类质量高。

注意:ISODATA 与 C 均值算法的以下异同点。

(1) 都是动态聚类算法。

(2) C 均值算法较简单,ISODATA 算法较复杂。

(3) 在 C 均值算法中,类型数目固定;在 ISODATA 算法中,类型数目可变。

5.6 Python 示例

【例 5.7】 使用 Python 对最大最小距离算法进行仿真,并通过该算法对二维模式样本 x=[0 3 2 1 5 4 6 5 6 7;0 8 2 1 3 8 3 4 4 5]进行聚类,样本共有 10 个点,设置合适的阈值,最后将其分为 3 类。

```
import math
import numpy as np
import matplotlib.pyplot as plt
import matplotlib as mpl

mpl.rcParams['font.sans-serif'] = [u'SimHei']   # 设置字体为 SimHei 显示中文
mpl.rcParams['axes.unicode_minus'] = False      # 设置正常显示字符

def calcuDistance(data1, data2):
    """  计算两个模式样本之间的欧氏距离 """
```

```
    distance = 0
    for i in range(len(data1)):
        distance += pow((data1[i] - data2[i]), 2)
    return math.sqrt(distance)

def maxmin_distance_cluster(data, Theta):
    """
    :param data: 输入样本数据,每行一个特征
    :param Theta:阈值,一般设置为 0.5,阈值越小聚类中心越多
    :return:样本分类,聚类中心
    """
    maxDistance = 0
    start = 0                                    # 初始选一个中心点
    index = start                                # 相当于指针指示新中心点的位置
    k = 0                                        # 中心点计数,也即是类别

    dataNum = len(data)                          # 样本数
    distance = np.zeros((dataNum,))
    minDistance = np.zeros((dataNum,))
    classes = np.zeros((dataNum,))
    centerIndex = [index]

    ptrCen = data[0]                             # 初始选择第一个为聚类中心点
    # 寻找第二个聚类中心,即与第一个聚类中心最大距离的样本点
    for i in range(dataNum):
        ptr1 = data[i]
        d = calcuDistance(ptr1, ptrCen)
        distance[i] = d
        classes[i] = k + 1
        if (maxDistance < d):
            maxDistance = d
            index = i                            # 与第一个聚类中心距离最大的样本
            print("与第一个聚类中心的距离为:{},索引为:{}".format(distance[i], index))
                                                 # 打印欧氏距离及新的聚类中心
    minDistance = distance.copy()
    maxVal = maxDistance
    while maxVal > (maxDistance * Theta):
        k = k + 1
        centerIndex += [index]                   # 新的聚类中心
        for i in range(dataNum):
            ptr1 = data[i]
            ptrCen = data[centerIndex[k]]
            d = calcuDistance(ptr1, ptrCen)
            distance[i] = d
            # 按照当前最近邻方式分类,哪个近就分哪个类别
            if minDistance[i] > distance[i]:
                minDistance[i] = distance[i]
                classes[i] = k + 1
        # 寻找 minDistance 中的最大距离,若 maxVal > (maxDistance * Theta),则说明存在下一
        # 个聚类中心
        index = np.argmax(minDistance)
            print("最小值中的最大值:{},索引为:{}".format(minDistance[i], index))
        maxVal = minDistance[index]
```

```
    return classes, centerIndex

data = [[0, 0], [3, 8], [2, 2], [1, 1], [5, 3], [4, 8], [6, 3], [5, 4], [6, 4], [7, 5]]
Theta = 0.5
classes, centerIndex = maxmin_distance_cluster(data, Theta)
print("样本所属类别:",classes)

marker = ['o', '*', 's']
color = ['r', 'b', 'g']
data = np.array(data)

plt.figure(figsize=(10,6),dpi=120)
plt.xlim(-1, 8)
plt.ylim(-1, 9)     # 设置坐标范围

# 画出样本数据
for idc in np.unique(classes):
    tag = classes == idc
    index = int(idc - 1)
    plt.scatter(data[tag,:][:, 0], data[tag,:][:,1], c=color[index],
                marker=marker[index], label=f"第{int(idc)}类", s=120)
# 画出中心点
for i in range(len(centerIndex)):
    plt.scatter(data[centerIndex[i]][0], data[centerIndex[i]][1], c=color[i], marker='x',
s=500)
plt.legend(loc="lower right",fontsize=16)
plt.grid(True, alpha=0.6)
plt.show()
```

第 16 集
微课视频

运行结果：

```
与第一个聚类中心的距离为:0.0,索引为:0
与第一个聚类中心的距离为:8.54400374531753,索引为:1
与第一个聚类中心的距离为:2.8284271247461903,索引为:1
与第一个聚类中心的距离为:1.4142135623730951,索引为:1
与第一个聚类中心的距离为:5.830951894845301,索引为:1
与第一个聚类中心的距离为:8.94427190999916,索引为:5
与第一个聚类中心的距离为:6.708203932499369,索引为:5
与第一个聚类中心的距离为:6.4031242374328485,索引为:5
与第一个聚类中心的距离为:7.211102550927978,索引为:5
与第一个聚类中心的距离为:8.602325267042627,索引为:5
最小值中的最大值:4.242640687119285,索引为:6
最小值中的最大值:2.23606797749979,索引为:2
样本所属类别: [1. 2. 1. 1. 3. 2. 3. 3. 3. 3.]
```

运行结果如图 5-12 所示。

【例 5.8】 使用 Python 对基于函数准则的 C 均值算法进行仿真，并实现对样本的聚类。选择的样本 X 为二维模式样本，设置两个聚类中心，画出样本聚类情况，并判断[2，3]和[6，9]分别属于哪一类。

```
import numpy as np
from matplotlib import pyplot
from pprint import pprint
```

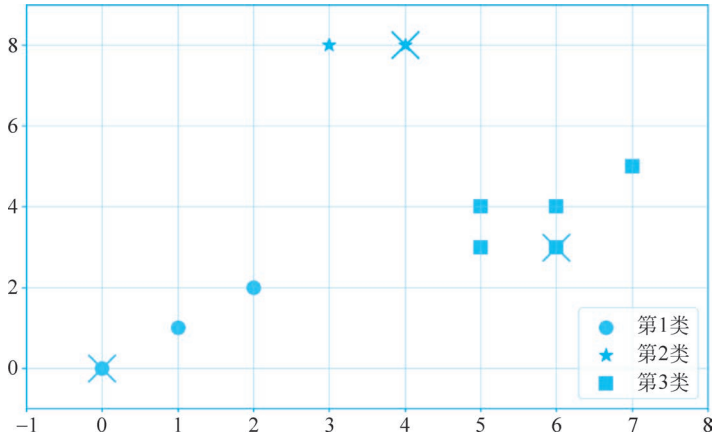


图 5-12　运行结果

```
import matplotlib as mpl

class K_Means(object):
    def __init__(self, k = 2, tolerance = 0.0001, max_iter = 300):
        """
        :param k:分组数
        :param tolerance:中心点误差
        :param max_iter:迭代次数
        """
        self.k_ = k
        self.tolerance_ = tolerance
        self.max_iter_ = max_iter

    def fit(self, data):
        """  k均值计算 """
        self.centers_ = {}                                  # 中心点
        for i in range(self.k_):
            self.centers_[i] = data[i]

        for i in range(self.max_iter_):
            self.clf_ = {}                                  # 分组情况
            for j in range(self.k_):
                self.clf_[j] = []                           # 每次迭代清空分组结果
            for feature in data:
                distances = []
                for center in self.centers_:
                    distances.append(np.linalg.norm(feature - self.centers_[center]))
                                                            # 欧氏距离
                classification = distances.index(min(distances))
                                                            # 单个数据的分组结果
                self.clf_[classification].append(feature)
                                                            # 将单个数据添加到不同组中

            print(f"第{i + 1}次迭代")
            print("中心点:",end = ' ')
            pprint(self.centers_)
```

```
            print("分组情况:",)
            pprint(self.clf_, width=120, indent=4, compact=True)
            print("--------------------------------------")

            prev_centers = dict(self.centers_)
            for c in self.clf_:
                self.centers_[c] = np.average(self.clf_[c], axis=0)
                                                        # 重新计算中心点坐标

            # 中心点是否在误差范围
            optimized = True
            for center in self.centers_:
                org_centers = prev_centers[center]      # 上一次的中心点坐标
                cur_centers = self.centers_[center]     # 这一次的中心点坐标
                if np.sum((cur_centers - org_centers) / (org_centers * 100.0 + 1e-6)) >
self.tolerance_:
                    optimized = False
            if optimized:
                break  # 两次中心点坐标比较相差无几后,结束循环

    def predict(self, p_data):
        """ k均值预测数据 """
        distances = [np.linalg.norm(p_data - self.centers_[center]) for center in self.
centers_]                                               # 欧氏距离
        index = distances.index(min(distances))         # 单个数据的分组结果
        return index

x = np.array([[0, 0], [1, 0], [0, 1], [1, 1], [2, 1], [1, 2], [2, 2], [3, 2], [6, 6], [7, 6],
[8, 6], [6, 7], [7, 7], [8, 7], [9, 7], [7, 8], [8, 8], [9, 8], [8, 9], [9, 9]])
k_means = K_Means(k=2)                                  # 分成2类
k_means.fit(x)

# 开始绘图
pyplot.figure(figsize=(10, 8), dpi=120)
for i, center in enumerate(k_means.centers_):
    # 画出中心点
    pyplot.scatter(k_means.centers_[center][0], k_means.centers_[center][1],marker='*',
s=150, label=f"第{i}类中心")

for i, cat in enumerate(k_means.clf_):
    point = np.array(k_means.clf_[cat])
    pyplot.scatter(point[:, 0], point[:, 1], c=('r' if cat == 0 else 'b'), label=f"第{i}
类")                                                    # 画出样本数据

predict = [[2, 3], [6, 9]]
for feature in predict:
    cat = k_means.predict(feature)
    pyplot.scatter(feature[0], feature[1], c=(
        'r' if cat == 0 else 'b'), marker='x',
                label = "待检测样本")                   # 画出预测数据

pyplot.legend(loc="lower right", fontsize=16)
pyplot.grid(True, alpha=0.6)
pyplot.show()
```

运行结果：

```
第 1 次迭代
中心点：{0: array([0, 0]), 1: array([1, 0])}
------------------------------------------------------------
第 2 次迭代
中心点：{0: array([0. , 0.5]), 1: array([5.66666667, 5.33333333])}
------------------------------------------------------------
第 3 次迭代
中心点：{0: array([1.25 , 1.125]), 1: array([7.66666667, 7.33333333])}
```

聚类结果如图 5-13 所示。

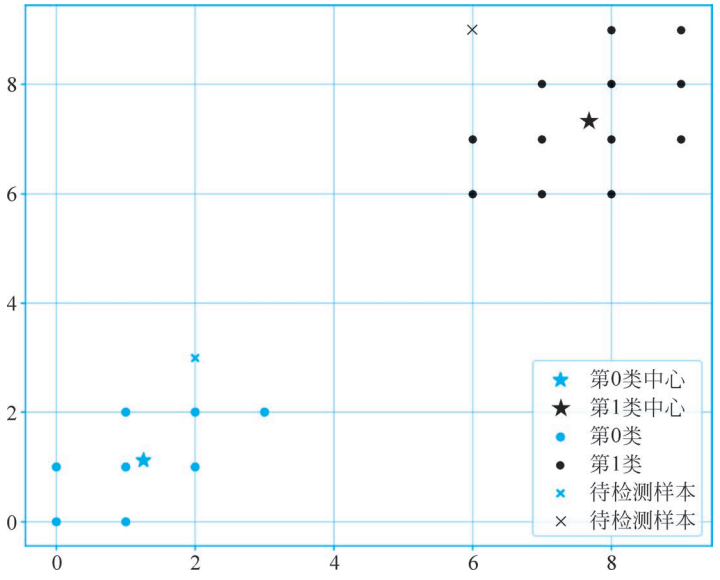


图 5-13　聚类结果

对获取的数据具有随机性的样本可采用 Bayes 理论进行分类，其前提是各类别总体的概率分布已知，要决策的分类的类别数一定。对于确定性的模式，如果类别已知（训练样本属性也已知），则可以通过第 2 章介绍的方法进行分类。然而在实际应用中，不少情况下无法预先知道样本的标签，也就是说没有训练样本，因而只能从原先没有样本标签的样本集开始进行分类器设计，这就是通常说的无监督学习方法，这就是本章介绍的聚类分析方法。聚类分析无训练过程，训练与识别混合在一起完成。

习题及思考题

5.1　证明马哈拉诺比斯距离是平移不变的、非奇异线性变换不变的。

5.2　简述有监督学习方法和无监督学习方法的异同。

5.3　请聚类下列数据（其中(x,y)代表坐标），将其分为三个簇。

$A_1(2,10),A_2(2,5),A_3(8,4),B_1(5,8),B_2(7,5),B_3(6,4),C_1(1,2),C_2(4,9)$

其距离为欧氏（欧几里得）距离。起初假设A_1、B_1、C_1为每个簇的聚类中心。用 C 均值算法给出：在第一次循环后的三个簇中心和最终的三个簇中心。

5.4 ISODATA 算法较之于 C 均值算法的优势何在？

5.5 (1) 设有 M 类模式 $\omega_j, j=1,2,\cdots,M$，试证明总体离散度矩阵 $\boldsymbol{S}_t$ 是总的类内离散度矩阵 $\boldsymbol{S}_w$ 与类间离散度矩阵 $\boldsymbol{S}_b$ 之和，即 $\boldsymbol{S}_t=\boldsymbol{S}_w+\boldsymbol{S}_b$。

(2) 设有二维样本：$\boldsymbol{x}_1=[-1,0]^T$，$\boldsymbol{x}_2=[0,-1]^T$，$\boldsymbol{x}_3=[0,0]^T$，$\boldsymbol{x}_4=[2,0]^T$ 和 $x_5=[0,2]^T$。试选用一种合适的方法进行一维特征提取 $\boldsymbol{y}_i=\boldsymbol{W}^T\boldsymbol{x}_i$。要求求出变换矩阵 $\boldsymbol{W}$，并求出变换结果 $\boldsymbol{y}_i(i=1,2,3,4,5)$。

(3) 根据(2)特征提取后的一维特征，选用一种合适的聚类算法将这些样本分为两类，要求每类样本个数不少于两个，并写出聚类过程。

5.6 (1) 试给出 C 均值算法的算法流程。

(2) 试证明 C 均值算法可使误差平方和准则 $J^{(k)}=\sum_{j=1}^{c}\sum_{x_i\in\omega_j^{(k)}}(x_i-z_j^{(k)})^T(x_i-z_j^{(k)})$ 最小。其中，k 是迭代次数；$z_j^{(k)}$ 是 $\omega_j^{(k)}$ 的样本均值。

5.7 证明：

(1) 如果 s 是类 x 上的距离相似性测度，$\forall x,y>0, s(x,y)>0$，那么对于 $\forall a>0$，$s(x,y)+a$ 也是类 x 上的距离相似性测度。

(2) 如果 d 是类 x 上的距离差异性测度，那么对于 $\forall a>0, d+a$ 也是类 x 上的距离差异性测度。