$$\gamma(x,y)=\frac{\sum_s\sum_t[f(s,t)-\bar{f}(s,t)][w(x+s,y+t)-\bar{w}]}{\left\{\sum_s\sum_t[f(s,t)-\bar{f}(s,t)]^2\sum_s\sum_t[w(x+s,y+t)-\bar{w}]^2\right\}^{\frac{1}{2}}} \tag{12.2.8}$$

这里 $x=0,1,2,\cdots,M-1,y=0,1,2,\cdots N-1,\bar{w}$ 是 w 中的像素平均值(只计算一次),$\bar{f}$ 是 f 中与 w 当前所在位置相重合的区域平均值,总和的值通常由 f 和 w 两者的坐标代入后求得。相关系数 $\gamma(x,y)$在 -1 到 1 之间取值,与 f 和 w 幅值中的区间变化相独立(见习题 12.5)。

例 12.2 通过相关系数进行对象匹配

图 12.9 说明了刚才讨论过的概念。图 12.9(a)是 $f(x,y)$。图 12.9(b)是 $w(x,y)$。相关系数 $\gamma(x,y)$如图 12.9(c)所示。在 f 和 w 之间找到最佳匹配的地方,相关系数 $\gamma(x,y)$的值更大(更亮)。

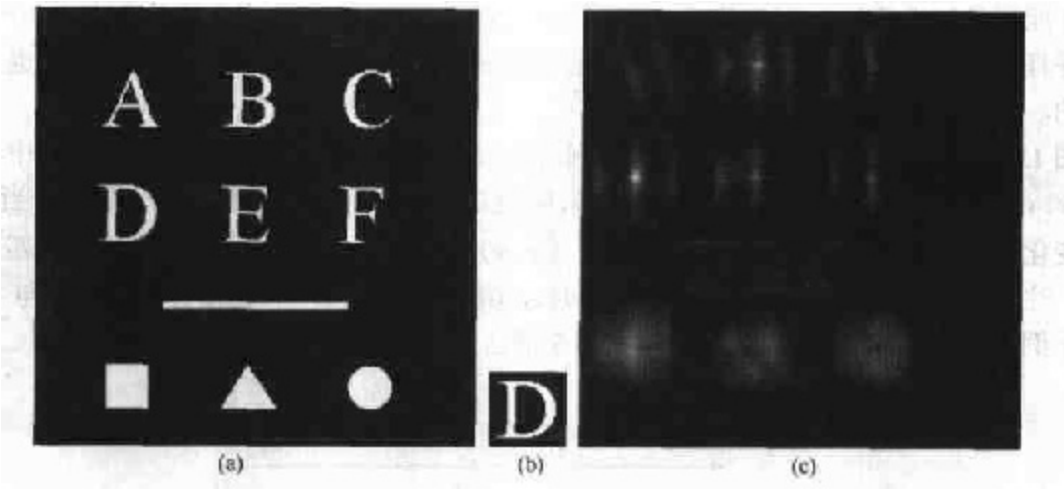

图 12.9 (a)图像,(b)子图,(c)(a)和(b)的相关系数。注意当子图(b)和(a)中的字母"D"一致时,(c)中出现最高值(亮点)

虽然相关函数对于幅度变化可以通过使用相关系数归一化,但要得到归一化的尺寸变化和旋转变化是困难的。尺寸的归一化涉及空间定标,这个过程本身会增加大量的计算。旋转变化的归一化更为困难。如果从 $f(x,y)$中可以提取其旋转变化方式的线索,就可以简单地对 $w(x,y)$ 进行旋转使它同 $f(x,y)$的旋转角度对准。然而,如果其旋转性质是未知的,那么寻找最佳匹配就要求对 $w(x,y)$进行全方位的旋转。这一过程是不实际的,当出现不确定的或不受约束的旋转变化时,相关很少使用。

在 4.6.4 节中我们曾提到,相关也可以通过 FFT 算法在频域内执行。如果 f 和 w 大小相同,这种方法比在空间域中直接实现图像的相关更有效。当 w 比 f 小得多时,可以使用式(12.2.7)。一种由 Campbell[1969]发明的折中估计法指出,如果 w 中的非零项数目小于 132(大约 13×13 像素的子图),则直接使用式(12.2.7)比 FFT 算法更为有效。当然,这个数目取决于使用的机器和算法,但是,它确实指出了在频域作为一种替代时应考虑子图的大致尺寸。在频域中使用相关系数更为困难。通常它是在空间域中直接计算出来的。

12.2.2 最佳统计分类器

在这一节中,我们将讨论一种用于模式识别的概率方法。在大多数处理测量和判断物理

事件的场合，对概率的考虑在模式识别中变得十分重要。因为，在这种情况下通常会产生随机的模式分类。正如下面讨论表明的那样，有可能推导出一种分类方法，这种方法在感觉上是最佳的，平均来说使用它产生分类错误的概率很低(见习题 12.10)。

基础

特定模式 $\mathbf{x}$ 来自 ω_i 类的概率表示为 $p(\omega_i/\mathbf{x})$。如果模式分类器判断 $\mathbf{x}$ 来自类 ω_j，而实际上它来自类 ω_i，分类器就会出现一次失败的分类，表示为 L_{ij}。如果考虑模式 $\mathbf{x}$ 可能属于类 W 中的任何一种模式，则在将模式指定为类 ω_j 时的平均失效率为：

$$r_j(\mathbf{x}) = \sum_{k=1}^{W} L_{kj} p(\omega_k/\mathbf{x}) \tag{12.2.9}$$

这个公式在决策理论的术语上叫做条件平均风险或条件平均失效。

由基本概率理论我们知道 $p(A/B) = [p(A)p(B/A)]/p(B)$。应用此表达式，我们用下列形式书写式(12.2.9)：

$$r_j(\mathbf{x}) = \frac{1}{p(\mathbf{x})} \sum_{k=1}^{W} L_{kj} p(\mathbf{x}/\omega_k) P(\omega_k) \tag{12.2.10}$$

这里 $p(\mathbf{x}/\omega_k)$是来自类 ω_k 的模式的概率密度函数，$P(\omega_k)$是类 ω_k 出现的概率。由于 $1/p(\mathbf{x})$ 是正的并且对所有的 $r_j(\mathbf{x})$，$j = 1,2,\cdots,W$ 都是如此，它可以从式(12.2.10)中被省略而不影响函数从最小值到最大值的相对顺序。平均失效率的表示就化简为：

$$r_j(\mathbf{x}) = \sum_{k=1}^{W} L_{kj} p(\mathbf{x}/\omega_k) P(\omega_k) \tag{12.2.11}$$

分类器有 W 个可能的类从任何已给出的未知模式进行选择。如果分类器对每一个模式 $\mathbf{x}$ 计算 $r_1(\mathbf{x}), r_2(\mathbf{x}), \cdots, r_W(\mathbf{x})$并且以最低的失效率对每一个模式制定相应的类，则对所有判决的总体平均失效率就会降至最低。这种能将总体平均失效率降至最低的分类器称做贝叶斯分类器。因此，贝叶斯分类器是在 $r_i(\mathbf{x}) < r_j(\mathbf{x})$，$j = 1,2,\cdots,W$ 且 $j \neq i$ 时将一个未知模式 $\mathbf{x}$ 归属给类 ω_i 的。换句话说，$\mathbf{x}$ 是在下列公式成立时才归属类 ω_i 的。

$$\sum_{k=1}^{W} L_{ki} P(\mathbf{x}/\omega_k) P(\omega_k) < \sum_{q=1}^{W} L_{qj} p(\mathbf{x}/\omega_q) P(\omega_q) \tag{12.2.12}$$

对所有的 $j; j \neq i$。通常，对一个正确决策分类“失败”赋予 0 值，对任何不正确决策分类失败通常赋予一个相同的非零值(即赋 1 值)。在这样的条件下，失效函数的形式为：

$$L_{ij} = 1 - \delta_{ij} \tag{12.2.13}$$

这里当 $i = j$ 时 $\delta_{ij} = 1$，当 $i \neq j$ 时 $\delta_{ij} = 0$。式(12.2.13)表明对非正确决策的一个单位失效和对正确决策的一个 0 失效。将式(12.2.13)代入式(12.2.11)得到式：

$$\begin{aligned} r_j(\mathbf{x}) &= \sum_{k=1}^{W} (1 - \delta_{kj}) p(\mathbf{x}/\omega_k) P(\omega_k) \\ &= p(\mathbf{x}) - p(\mathbf{x}/\omega_j) P(\omega_j) \end{aligned} \tag{12.2.14}$$

如果对所有 $j \neq i$ 时下列公式成立

$$p(\mathbf{x}) - p(\mathbf{x}/\omega_i)P(\omega_i) < p(\mathbf{x}) - p(\mathbf{x}/\omega_j)P(\omega_j) \tag{12.2.15}$$

或者等价地有如下公式成立时

$$p(\mathbf{x}/\omega_i)P(\omega_i) > p(\mathbf{x}/\omega_j)P(\omega_j) \quad j = 1,2,\cdots,W; j \neq i \tag{12.2.16}$$

则贝叶斯分类器将一个模式 $\mathbf{x}$ 赋予类 ω_i。

参考导出式(12.2.1)的讨论,可以了解到对于 0-1 失效函数的贝叶斯分类器不过是下列函数的决策值的计算:

$$d_j(\mathbf{x}) = p(\mathbf{x}/\omega_j)P(\omega_j) \quad j = 1,2,\cdots,W \tag{12.2.17}$$

这里哪个类的决策函数的值最大,模式矢量 $\mathbf{x}$ 就归属于哪一个类。

在式(12.2.7)中给出的判别函数在将错误分类的平均失效率降低到最低的能力方面是最佳的。然而,在每个类中的模式的概率密度函数和每种类出现的概率必须是已知的。后者的要求通常并不构成问题。比如,如果所有类的出现几率大致相同,就可令 $P(\omega_j) = 1/M$。即使这个条件不正确,我们也可以通过对问题的认识推断出这些概率。概率密度函数 $p(\mathbf{x}/w_j)$ 的估计就是另一回事了。如果模式矢量 $\mathbf{x}$ 是 n 维的,那么 $p(\mathbf{x}/\omega_j)$ 就是一个 n 元函数,如果它的形式是未知的,就需要使用多元概率理论的方法对它进行估计。这类方法在实际使用中很难应用,尤其是代表每一个类的模式数目不大,或隐含的概率密度函数形式的规律性不佳时更是如此。由于这些原因,贝叶斯分类器的运用通常是基于以下假设:对不同概率密度函数的解析式及从每一类样本模式估计的必需参数。目前,对 $p(\mathbf{x}/\omega_j)$ 的假设形式普遍使用的是高斯概率密度函数。这种设定越接近真实情况,贝叶斯分类器方法在分类中越能接近最低平均失效率。

高斯模式类的贝叶斯分类器

首先,考虑一个包含两个模式类($W = 2$)的一维($n = 1$)问题,这两个模式类具有高斯密度,分别具有均值 m_1 和 m_2 与标准差 σ_1 和 σ_2。由式(12.2.17)可知贝叶斯判别函数具有如下形式:

$$\begin{aligned} d_j(x) &= p(x/\omega_j)P(\omega_j) \\ &= \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(x-m_j)^2}{2\sigma_j^2}} P(\omega_j) \quad ,j = 1,2 \end{aligned} \tag{12.2.18}$$

这里,模式此时是标量,用 x 表示。图 12.10 显示了两个类概率密度函数的曲线图。两个类之间的交界是一个单点,用 x_0 表示,使得 $d_1(x_0) = d_2(x_0)$。如果两个类出现几率相等,则 $P(\omega_1) = P(\omega_2) = 1/2$,并且决策边界就是 $p(x_0/\omega_1) = P(x_0/\omega_2)$ 处 x_0 的值。这一点如图 12.10所示,是两个概率密度函数的交集。任何处于 x_0 点右侧的模式(点)都被划归类 ω_1。同样,任何处于 x_0 点左侧的模式(点)都被划归类 ω_2。当两个类出现几率不相等时,如果类 ω_1 出现几率大则 x_0 点左移,或者相反,如果类 ω_2 出现几率大则 x_0 点右移。这就是所期望的结果,因为分类器总是试图将错误分类的失效降至最低。例如,在极端情况下,如果类 ω_2 从不出现,通过总将所有模式划归类 ω_1,分类器也不会出现错误(即点 x_0 趋于负无穷)。

在 n 维情况下,第 j 个模式类中矢量的高斯密度具有如下所示的形式:

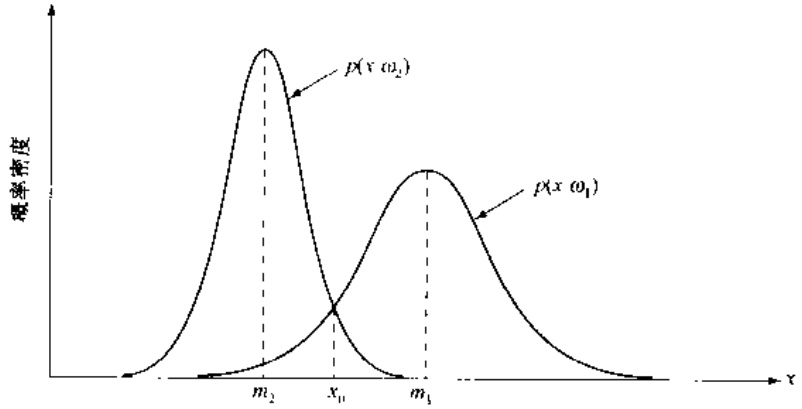

图 12.10　两个一维模式类的概率密度函数。如果两个类出现几率相等，则点 x_0 就是决策边界

$$p(\mathbf{x}/\omega_j) = \frac{1}{(2\pi)^{n/2}|\mathbf{C}_j|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_j)^T\mathbf{C}_j^{-1}(\mathbf{x}-\mathbf{m}_j)} \tag{12.2.19}$$

这里每一个密度函数都由其矢量平均数 $\mathbf{m}_j$ 和协方差矩阵 $\mathbf{C}_j$ 完全指定，它被定义如下：

$$\mathbf{m}_j = E_j\{\mathbf{x}\} \tag{12.2.20}$$

和

$$\mathbf{C}_j = E_j\{(\mathbf{x} - \mathbf{m}_j)(\mathbf{x} - \mathbf{m}_j)^T\} \tag{12.2.21}$$

这里 $E_j\{\cdot\}$ 代表类 ω_j 中模式的变元期望值。在式(12.2.19)中，n 是模式矢量的维度，$|\mathbf{C}_j|$ 是矩阵 $\mathbf{C}_j$ 的行列式。用问题中量值的平均值近似期望值 E_j，得到平均矢量值的估计值和协方差矩阵如下：

$$\mathbf{m}_j = \frac{1}{N_j}\sum_{\mathbf{x}\in\omega_j}\mathbf{x} \tag{12.2.22}$$

和

$$\mathbf{C}_j = \frac{1}{N_j}\sum_{\mathbf{x}\in\omega_j}\mathbf{x}\mathbf{x}^T - \mathbf{m}_j\mathbf{m}_j^T \tag{12.2.23}$$

这里 N_j 是来自类 ω_j 的模式矢量数目，并且总和取全部向量。在这一节的后边，将给出一个例子示范如何使用这两个表达式。

协方差矩阵是对称的并且是半正定的。如 11.4 节中解释的那样，对角线元素 c_{kk} 是模式矢量第 k 个元素的方差。非对角线元素 c_{jk} 是 x_j 和 x_k 的协方差。多元高斯密度函数在协方差矩阵的非对角线元素均为零时还原为 $\mathbf{x}$ 每一个元素的单变元高斯密度的乘积。这种情况是在矢量变元 x_j 和 x_k 不相关时出现的。

依据式(12.2.17)，类 ω_j 的贝叶斯判别函数为 $d_j(\mathbf{x}) = p(\mathbf{x}/\omega_j)P(\omega_j)$。然而，由于高斯密度函数的指数形式，用这个判别函数的自然对数形式更为方便。换句话说，可以使用下列形式的公式：

$$\begin{aligned} d_j(\mathbf{x}) &= \ln[p(\mathbf{x}/\omega_j)P(\omega_j)] \\ &= \ln p(\mathbf{x}/\omega_j) + \ln P(\omega_j) \end{aligned} \tag{12.2.24}$$

从分类性能的角度来说这个表达式同式(12.2.17)是等价的。因为对数是一个单调递增函数。换句话说,式(12.2.17)中的判别函数的顺序和式(12.2.24)中的顺序是一样的。将式(12.2.19)代入式(12.2.24)得到式(12.2.25)。

$$d_j(\mathbf{x}) = \ln P(\omega_j) - \frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{C}_j| - \frac{1}{2}[(\mathbf{x}-\mathbf{m}_j)^T\mathbf{C}_j^{-1}(\mathbf{x}-\mathbf{m}_j)] \qquad (12.2.25)$$

项$(n/2)\ln 2\pi$对所有的类都是相同的,因此,可以从式(12.2.25)中消去。成为下列形式:

$$d_j(\mathbf{x}) = \ln P(\omega_j) - \frac{1}{2}\ln|\mathbf{C}_j| - \frac{1}{2}[(\mathbf{x}-\mathbf{m}_j)^T\mathbf{C}_j^{-1}(\mathbf{x}-\mathbf{m}_j)] \qquad (12.2.26)$$

对于$j=1,2,\cdots,W$。式(12.2.26)表示高斯模式类在0-1失效函数条件下的贝叶斯判别函数。

式(12.2.26)中的判别函数是超二次曲面(n维空间中的二次函数),因为式子中出现的$\mathbf{x}$项没有高于二次的。毫无疑问,一个高斯模式的贝叶斯分类器最多做到在每对模式类之间置一个二阶判别曲面。不过,如果模式域确实是高斯的,则在分类中不会有其他曲面所产生的失效率少于平均失效率。

如果所有的协方差矩阵都相等,即$\mathbf{C}_j=\mathbf{C}, j=1,2,\cdots,W$。通过展开式(12.2.26)并消去所有独立于$j$的项,会得到下列公式:

$$d_j(\mathbf{x}) = \ln P(\omega_j) + \mathbf{x}^T\mathbf{C}^{-1}\mathbf{m}_j - \frac{1}{2}\mathbf{m}_j^T\mathbf{C}^{-1}\mathbf{m}_j \qquad (12.2.27)$$

这是一个线性判别函数(超平面),$j=1,2,\cdots,W$.

另外,如果$\mathbf{C}=\mathbf{I}$,这里$\mathbf{I}$代表单位距阵,且$P(\omega_j)=1/W, j=1,2,\cdots,W$,则有公式:

$$d_j(\mathbf{x}) = \mathbf{x}^T\mathbf{m}_j - \frac{1}{2}\mathbf{m}_j^T\mathbf{m}_j, j=1,2,\cdots,W \qquad (12.2.28)$$

这些公式是最小距离分类器的判别函数,如式(12.2.5)所给出的一样。如果(1)模式类都是高斯的,(2)所有协方差矩阵等于单位矩阵,并且(3)所有类出现的几率相等,则最小距离分类器在贝叶斯意义上是最佳的。满足这些条件的高斯模式类是n维空间中外形相同的球状云团(称为超球面)。最小距离分类器在每对类之间设置一个超平面,这个超平面的特性是垂直等分连接每对超球面中心的线段。在两个维度上,类组成圆形区域,并且边界变为等分连接每对圆环中心的线段的直线。

例12.3 三维模式的贝叶斯分类器

图12.11显示了一个由三维空间中的两个类组成的简单排列。我们用这些模式说明运用贝叶斯分类器的机理,假设每个类中的模式都是服从高斯分布的样本。

对图12.11中的模式应用式(12.2.22)得到:

$$\mathbf{m}_1 = \frac{1}{4}\begin{bmatrix}3\\1\\1\end{bmatrix} \quad 和 \quad \mathbf{m}_2 = \frac{1}{4}\begin{bmatrix}1\\3\\3\end{bmatrix}$$

类似地,对两个模式类依次使用式(12.2.23)得到两个协方差矩阵,在这种情况下两个矩阵是相等的:

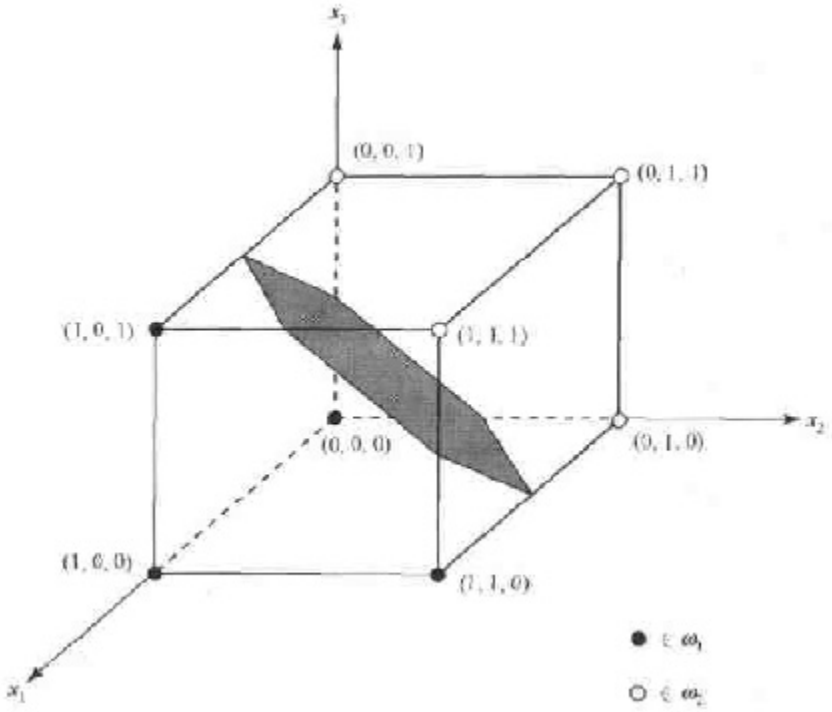

图 12.11　两个简单模式类和它们的贝叶斯判别边界(阴影处)

$$\mathbf{C}_1 = \mathbf{C}_2 = \frac{1}{16}\begin{bmatrix} 3 & 1 & 1 \\ 1 & 3 & -1 \\ 1 & -1 & 3 \end{bmatrix}$$

因为协方差矩阵相等，则贝叶斯判别函数由式(12.2.27)给出。如果假定 $P(\omega_1) = P(\omega_2) = 1/2$，然后应用式(12.2.28)会得到：

$$d_j(\mathbf{x}) = \mathbf{x}^T C^{-1}\mathbf{m}_j - \frac{1}{2}\mathbf{m}_j^T\mathbf{C}^{-1}\mathbf{m}_j$$

这里

$$\mathbf{C}^{-1} = \begin{bmatrix} 8 & -4 & -4 \\ -4 & 8 & 4 \\ -4 & 4 & 8 \end{bmatrix}$$

对 $d_j(\mathbf{x})$执行矢量矩阵展开，规定判别函数：

$$d_1(\mathbf{x}) = 4x_1 - 1.5 \quad 和 \quad d_2(\mathbf{x}) = -4x_1 + 8x_2 + 8x_3 - 5.5$$

分开两个类的决策面是：

$$d_1(\mathbf{x}) - d_2(\mathbf{x}) = 8x_1 - 8x_2 - 8x_3 + 4 = 0$$

图 12.11 显示了这个面的一部分，注意到类被有效地分开了。

贝叶斯分类器方法最成功的应用之一是，对在航天器、卫星或空间站上利用多频谱扫描器产生的遥感图像进行分类处理。对这些平台上产生的大量图像数据进行自动图像分类和分析处理，是在遥感领域相当有意义的一项任务。遥感技术应用于许多领域，包括在陆地上的应

用,如农作物产量的调查,农作物灾害检测,森林、空气质量和水质的监测,地质研究,天气预报以及许多对环境有重要意义的领域。下面的例子显示了一个典型应用。

例 12.4 使用贝叶斯分类器对多频谱数据的分类

如在 1.3.4 节和 11.4 节曾讨论过的,多频谱扫描器在选定的频段对电磁能量有响应,例如 0.40 ~ 0.44 μm,0.58 ~ 0.62 μm,0.66 ~ 0.72 μm 和 0.80 ~ 1.00 μm 波段。这些频率范围分别处于紫色光、绿色光、红色光和红外频段上,在这些频段上对一个区域进行扫描,生成 4 幅数字图像,每个频段产生一幅。如果对这些图像进行配准(实际上经常是这样做的),就可以将它们前后叠加在一起形成一幅图像呈现出来,如图 12.12 所示。因而,就如同在 11.4 节中做过的,地面上的每一个点都可以采用 $\mathbf{x} = (x_1, x_2, x_3, x_4)^T$ 表示成一个四元矢量的形式。这里 x_1 代表紫外光的像,x_2 代表绿光的像,等等。如果图像是 512 × 512 像素大小的,每幅四频段的叠加图可以用 262 144 个四维模式矢量表示。

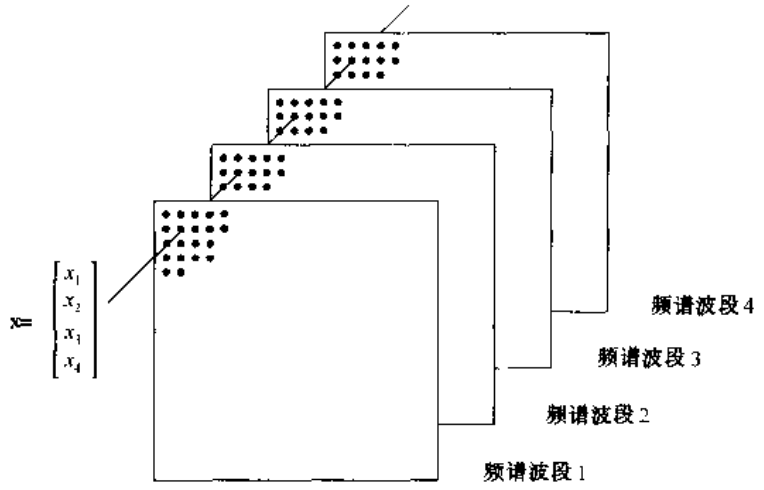

图 12.12 由多频谱扫描器生成的 4 幅数据图像经像素配准后的模式矢量格式

如以前所注意到的,高斯模式的贝叶斯分类器需要对每类平均矢量和协方差矩阵的估计。在遥感应用中,这些估计是通过像前面的范例中提到的那样,收集感兴趣地区的多频谱数据并运用这些样本而得到的。图 12.13(a)显示的是一幅从航天器上得到的典型遥感图像(本图是由一幅多频谱原图转换成的单色图)。在这种特殊情况下,问题是将所拍区域划分为植被、水和裸露的土地等不同区域。图 12.13(b)显示的是用高斯型贝叶斯分类器分类的结果。箭头显示了我们感兴趣的特征。箭头 1 指向一块绿色植被区的拐角处。箭头 2 指出了一条河。箭头 3 指向两块裸露土地之间的灌木树篱。箭头 4 指出了被系统正确识别出来的一条支流。箭头 5 指出了在图 12.13(a)中几乎无法分辨的一个小池塘。把原始图像与计算机输出的识别结果相比较可以看出,它与人类通过视觉分析得到的识别结果非常相似。

在结束这一节之前,有意思地注意到,逐点分类方法确实可以将图像分成不同的类,正如我们在前述例子中描述的那样。这种方法类似于用多个变量的门限分段法,这种方法在 10.3.7 节中有简要的描述。

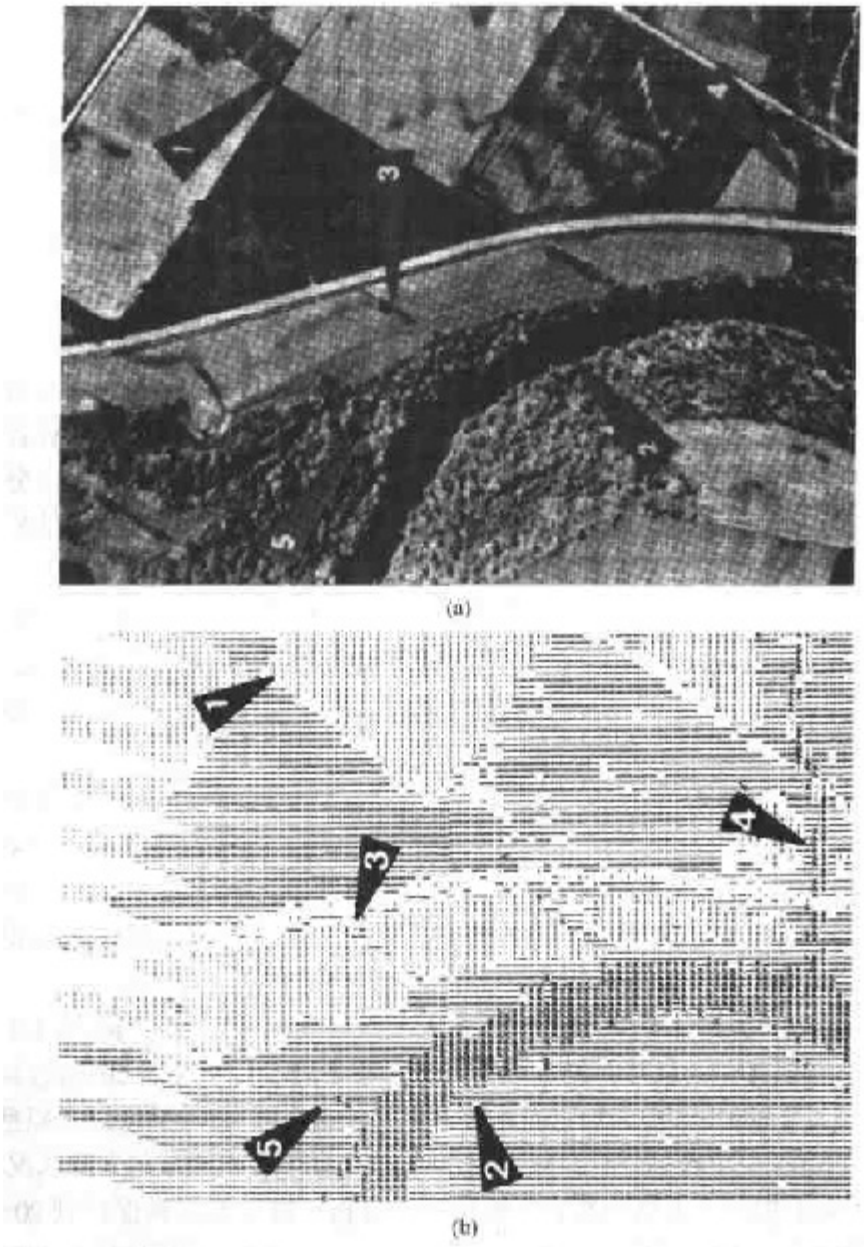

图 12.13　(a)多频谱图像,(b)打印出来的使用了贝叶斯分类器的机器分类结果(由 Purdue 大学遥感应用实验室提供)

12.2.3　神经网络

这种方法在以前的两节中有所论述。它基于使用样本模式来估计每个模式类的统计参数。最小距离分类器完全由每个类的平均矢量确定。同样,具有高斯密度的贝叶斯分类器完全由每个类的平均矢量和协方差矩阵确定。用于估计这些参数的模式(已知其所属的类)称做训练模式,一系列从每个类中得到的这类模式称做一个训练集合。使用训练集合得到判别函数的过程叫做学习或训练。

在刚刚讨论过的两种方法中,训练是简单的。每一个类的训练模式用于计算符合本类的判别函数参数。在估计了问题中的参数后,分类器的结构就被固定下来了,而且分类器的最终

效果的好坏取决于真实的模式类密度是否满足基本的统计假设。此处使用的统计假设来自所使用的分类方法。

对一个要解决的问题来说,模式类的统计特性通常是未知的或者无法估计的(回想我们在前面章节谈到的应用多元统计所遇到的困难)。实际上,这类决策理论问题最好使用直接通过训练过程生成所需判别函数这一方法来处理。对基本的概率密度函数或关于模式类的其他概率信息进行假设是必要的。在这一节中,我们将讨论各种符合这一准则的方法。

背景

以下所列材料的精髓就是使用大量非线性元素的计算单元(称为神经元),这些神经元是以被认为同大脑神经元的互联方式一样的方法组织起来的。其产生的模型有各种各样的名字,包括神经网络、神经计算机、分布式并行处理(PDP)模型、神经形态学系统、分层次自适应网络和连接模式。在这里,使用"神经网络"这一名称或简称"神经网"。使用这些网络作为工具,通过对模式训练集合的进一步描述自适应地导出判别函数系数。

神经网络受到关注可以追溯到 20 世纪 40 年代早期 McCulloch 和 Pitts [1943]的示范性工作。他们提议用二进制门限器件和包括 0-1 和 1-0 的神经状态变化的随机算法作为神经系统建模的基础。由 Hebb [1949]进行的后续工作是建立在数学模型基础上的,这个数学模型试图通过增强或建立关联来掌握"学习"这一概念。

20 世纪 50 年代中期至 20 世纪 60 年代早期,由 Rosenblatt [1959,1962]发明的一种称做学习机的装置在模式识别理论领域的研究者和开创者中引起了巨大的轰动。人们对这种称做感知器的机器给予如此关注的原因是,由于数学的发展,证明这种感知器在使用线性离散训练集合(也就是可被超平面分开的训练集合)进行训练后,可以通过有限的迭代步骤得出解答。这个解答采取超平面系数(它们能正确分离由训练集模式描述的类)的形式。

遗憾的是,在接下来研究学习模式的过程中,人们很快便大失所望。当时基本的感知器和它的一些推广对于大部分有实际意义的模式识别任务是很不够的。后来,试图通过将这种装置多层化来扩展类感知器装置的能力,尽管它们在概念上都较有吸引力,但是均缺乏原型感知器那样有效的训练算法。Nilsson [1965]总结了 20 世纪 70 年代中期学习机领域的大致状况。几年之后,Minsky 和 Papert [1969]提出了类感知器装置的局限性分析。这个看法被保持到 20 世纪 80 年代中期。这一观点在 Simon[1986]的评论中得到了证明。在这一工作中,原始论文 1984 年以法文出版,在这篇文章中,Simon 以"一个神话的诞生和死亡"为题目对感知器进行驳斥。

由 Rumelhart, Hinton 和 Williams[1986]在多重感知器新的训练运算法则的发展中得出的更新结果,使这个问题发生了很大变化。他们的基本方法,即被称做反向传播方式学习的一般性德尔塔(delta)规则,为多层机器提供了一种有效的训练方法。尽管这种训练算法与单层感知器比起来,在相似的证明方面无法得出一个集中的解,但产生的德尔塔规则在解决大量实际问题时却得到了成功的应用。这一成功奠定了多层类感知器装置作为现今神经网络应用中的主要模型之一的地位。

两个模式类的感知器

在这种最基本的形式中,感知器"学习"了一个线性判别函数,这个判别函数将两个线性分离的训练集合对半分开。图 12.14(a)示意性地显示了两个模式类的感知器模型。

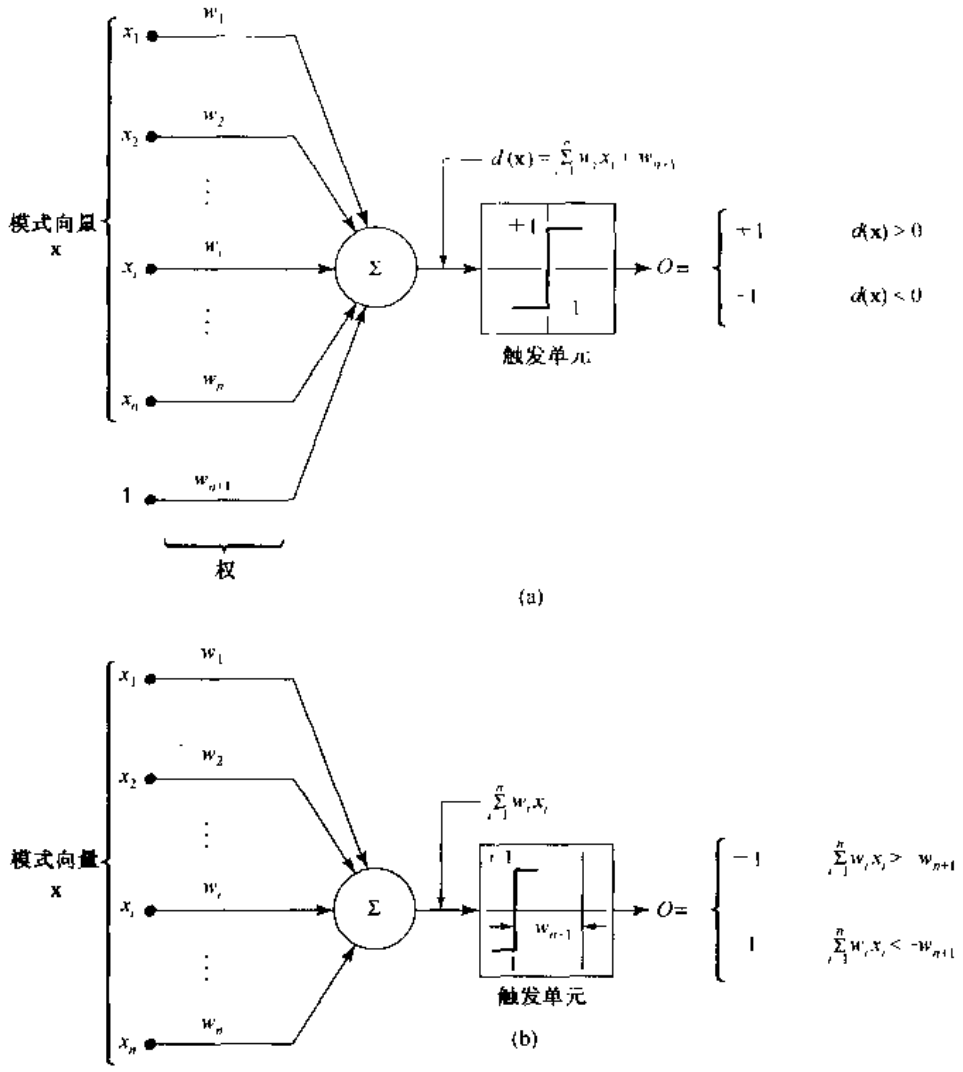

图 12.14　对两个模式类的感知器模型的等效描述

这个基本装置的响应是基于输入值的加权和，即：

$$d(\mathbf{x}) = \sum_{i=1}^{n} w_i x_i + w_{n+1} \tag{12.2.29}$$

这是一个考虑了模式矢量的分量的线性判别函数。系数 $w_i, i = 1,2,\cdots,n,n+1$，称为权，在元素相加并被输入到门限元件中之前对输入进行修正。在这方面，权值同人类大脑神经系统中的神经突触是相似的。把总连接的输出映射到最终的装置输出的函数有时称为激活函数。

当 $d(\mathbf{x}) > 0$ 时门限元件使感知器的输出为 +1，表示模式 $\mathbf{x}$ 被识别并划归类 ω_1。当 $d(\mathbf{x}) < 0$时相反的情况也是正确的。这种操作模式与前面做出的和式(12.2.2)相关的评论是一致的，那时考虑对两个模式类使用单个判别函数进行判别。当 $d(\mathbf{x}) = 0$ 时，$\mathbf{x}$ 位于分离两个模式类的判别(决策)平面上，给出了一个不能确定的条件。由感知器实现的判别边界是通过设式(12.2.29)等于 0 得到的。

$$d(\mathbf{x}) = \sum_{i=1}^{n} w_i x_i + w_{n+1} = 0 \tag{12.2.30}$$