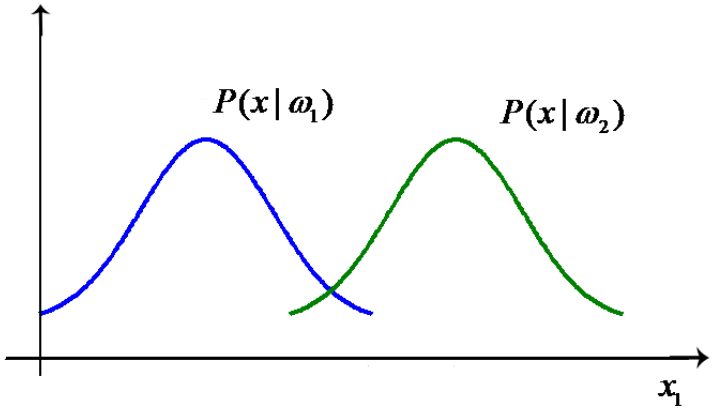
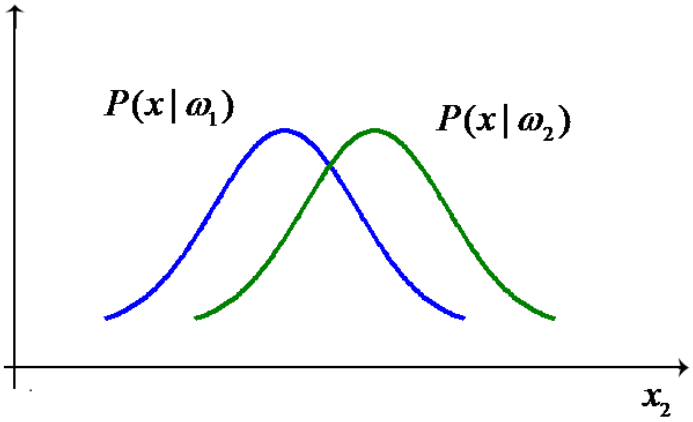
第8章　特征提取和特征选择

8.1. 引言

前面讨论分类器设计的时候，一直假定已给出了样本集 X，并且 X 中各样本的每一维都是该样本的一个特征。根据前面章节的讨论我们知道，不同的特征的作用是不同的，它强烈地影响到分类器的设计及其性能。图 8.1 给出的是一个两类分类问题在两个特征 x_1，x_2 时的数据分布示例。很明显，这两个特征中的 x_1 更易于分类。通常情况下，不同类别的样本的特征取值如果差别很大，那就比较容易设计出具有较好性能的分类器。因此，特征选择是模式识别中的一个关键问题。



（a）



（b）

图 8.1 一个两类分类问题在两个特征 x_1， x_2 时的数据分布示例

由于在很多实际问题中常常不容易找到那些最重要的特征，或受条件限制不能对它们进行测量，这就使特征选择和提取的任务复杂化而成为构造模式识别系统最困难的任务之一。这个问题已经越来越受到人们的重视。

任何识别过程的第一步，不论用计算机还是由人去识别，都要首先分析各种特征的有效性并选出最有代表性的特征。显然特征选择和提取这一任务应在设计分类器之前进行。但

是从教学经验看，在讨论分类器设计以后讲述特征选择和提取更有利于加深对这个问题的理解。

可以把特征分为三类：①物理的，②结构的，③数学的。人们通常利用物理和结构特征来识别对象，因为这样的特征容易被视觉、触觉以及其他感觉器官所发现。当我们需要设计模式识别产品时，我们常常需要考虑被识别对象的物理的和结构的特征。但在使用计算机去构造识别系统时应用这些特征有时比较复杂，因为一般说来用硬件去模拟人类感觉器官是很复杂的，而机器在抽取数学特征的能力方面则又比人强得多。这种数学特征的例子有很多，如统计平均值、相关系数、协方差阵的本征值及本征向量等。

8.1.1. 一些基本概念

在一些书籍和文献中，使用“特征提取”，“特征选择”等术语时的含义不完全相同。例如，“特征提取”在有的文献中专指特征的形成过程，有的则指从形成、经选择或变换直到得出有效特征这一全过程。在实际应用中，通过对对象进行测量，可以得到对象的一种描述，即用测量空间中的一个点来代表这个对象。例如，通过摄像机可以把一个物体转换为一个二维灰度阵列，即一幅图像。在一些识别任务中，不直接在测量空间中进行分类器设计。这一方面是因为测量空间的维数很高(一个 256×256 灰度图像相当于 256×256 维测量空间中的一个点)，不适宜于分类器的设计。更重要的是这样一种描述并不能直接反映对象的本质，并且它随摄像机位置、光照等因素的变化而变化。因此为了进行分类器设计，需要把图像从测量空间变换到维数大大减少的特征空间，被研究的图像或现象在这个特征空间中就由一个特征向量来表示。

实际上这样一种变换常常分成几个步骤进行，因此在一些文献中还采用特征提取和特征选择这样的术语。为了方便起见，我们对几个常用的有关名词作些说明。

特征形成：根据被识别的对象产生出一组基本特征，它可以是计算出来的(当识别对象是波形或数字图像时)，也可以是用仪表或传感器测量出来的(当识别对象是实物或某种过程时)，这样产生出来的特征叫做原始特征，有些书中用原始测量(或一次测量，或观察)这一名词，我们认为在很多情况下有些原始测量就可以作为原始特征，而有些情况则不然，例如识别对象是数字图像时，原始测量就是各点灰度值，但有时候我们不用各点灰度作为特征，而是需要经过计算产生一组原始特征。

特征提取：原始特征的数量可能很大，或者说样本是处于一个高维空间中，通过映射(或变换)的方法可以用低维空间来表示样本，这个过程叫特征提取。映射后的特征叫二次特征，它们是原始特征的某种组合(通常是线性组合)。所谓特征提取在广义上就是指一种变换。若 Y 是测量空间，X 是特征空间，则变换 A：Y→X 就叫做特征提取器。

特征选择：从一组特征中挑选出一些最有效的特征以达到降低特征空间维数的目的，这个过程叫特征选择。

以细胞自动识别为例，通过图像输入得到一批包括正常及异常细胞的数字图像，我们的任务是根据这些图像区分哪些细胞是正常的，哪些是异常的。首先要找出一组能代表细胞性质的特征。为此可以计算细胞总面积、总光密度、胞核面积、核浆比、细胞形状、核内纹理等，这样可得到很多原始特征，这一过程就是特征的形成。这样产生出来的原始特征可能很多(例如几十甚至几百个)，或者说原始特征空间维数很高，需要压缩维数以便分类。一种方式是用变换的方法把原始特征变换为较少的新特征，这就是特征提取。另一种方式就是从原始特征中去挑选出一些最有代表性的特征来，这就是特征选择。最简单的特征选择方法是

根据专家(这里是有经验的细胞学家)的知识挑选那些对分类最有影响的特征，另一个可能则是用数学的方法进行筛选比较，来找出最有分类信息的特征。

有时特征提取和选择并不是截然分开的。例如，可以先将原始特征空间映射到维数较低的空间，在这个空间中再进行选择以进一步降低维数。也可以先经过选择去掉那些明显没有分类信息的特征，再进行映射以降低维数。对于“特征提取”、“特征压缩”、“特征选择”在具体问题下的含义通过上下文是可以弄清楚的。

8.1.2. 一些生物特征识别应用问题举例

利用人体本身所拥有的生物特征来进行身份识别的技术叫做人体生物特征识别技术（Biometrics）。需要指出的是，与前面的特征一词的含义不完全相同，该小节中的“特征”是指一组特征的数据，如：指纹，人脸。作为人的生物特征的指纹，人脸又可以根据需要提取出很多用于分类的特征。本小节从几个生物特征识别的应用问题来讨论特征的形成过程中应该考虑的一些问题。

生物特征识别技术的发展主要源于社会上对于人的身份认证的需求。可以用来识别人的身份的人体生物特征需要满足以下几个条件：

(1)普遍性：这种特征是每个人都具有的；

(2)独特性：任意两个人的这种特征都不相同，因此，该特征可以用来区分不同的人；

(3)稳定性：这种特征至少是稳定不变的，即随着时间或者环境的变化不会发生大的变化；

(4)可采集性：这种特征数据可以被方便的采集和量化；

(5)可接受性：该特征数据的采集和使用容易被用户所接受；

(6)性能要求：基于该特征的系统应该具有较高的识别率；

(7)安全性：这种特征数据不容易被模仿或仿造；

人体特征分为两大类：生理特征和行为特征。下面我们对其中的几种进行讨论。

- DNA：即脱氧核糖核酸，存在于人体细胞中，是人的生理特征。DNA 具有很强的个体特异性和稳定性，从而可以用于身份识别。但是，由于在获取人的这一特征数据时，需要专业人员进行操作，还无法进行自动获取。另外，利用这种特征进行身份识别需要手工操作，需要的时间也很长。出于上述一些原因，利用 DNA 进行身份识别，其应用范围受到很大限制。目前主要用于刑侦和司法领域中。
- 指纹：指纹是人的生理特征。用指纹来进行身份识别已经有几百年的历史了，目前指纹自动识别已经应用于刑侦、海关、一些企业的考勤和门禁。研究表明，在人的指纹中，纹线的断点、分叉点之间的关系，以及指纹的类型、指纹的中心的位置等信息在人的一生中是保持不变的，具有很好的稳定性。图 8.2 所示是一张指纹及其上面的特征点。世界上几乎所有人都有指纹，并且根据计算，指纹具有很强的个体特异性。另外，目前也研制出越来越廉价的指纹获取设备，这为指纹识别的更广泛的应用提供了可能。目前，在采集指纹图像的时候，需要把人的手指放到一个专门的设备上采集。然后由软件自动提取指纹的特征和进行指纹的比对。由于指纹专家明确指出，指纹图像中纹线的断点、分叉点之间的关系，以及指纹的类型、指纹的中心的位置等信息具有特异性，因此，要实现一个指纹识别系统，核心的任务就是怎样从指纹图像中检测和定位纹线的断点、指纹的类型、指纹的中心等，并利用这些信息计算两张指纹的相似性，最后给出两张指纹图像

是否来自同一手指的判别结果。当指纹图像非常清晰时，目前的指纹识别算法可以达到非常高的识别率。而由于一些原因当指纹图像模糊不清的时候，要达到很高的识别率还是比较困难的。这是影响指纹识别进一步广泛应用的一个原因。另外，在一些人看来，指纹是人的隐私的一部分。因此，设计指纹识别产品时也要考虑这些因素。

- 人脸：人脸是人的生理特征。人脸是一种识别身份的最为自然和古老的方式。但是人脸的自动识别的研究是近几十年的事情。目前，人们通过摄像头获取人脸图像。与指纹相比较，人脸图像的获取方式更为自然，是一种非接触的图像采集方式，更易于被人们所接受。但是，我们知道，随着年龄的增长，人脸会发生变化，有时这种变化还比较大。另外，化妆，戴眼镜等也会影响人脸图像。这些对于人脸识别的应用造成了一定的困难。此外，人脸图像还强烈依赖于环境的光照，人脸的表情和朝向等因素。虽然我们具有很强的识别人脸的能力，但是，我们现在还不能“清晰地”指出哪些特征能够区别不同人的人脸，如何用算法稳定地从图像中提取这些特征更为困难。例如，我们很容易指出，人的面部器官的大小，形状以及它们之间的位置关系对于不同人来说具有区分性。但是，当人脸的姿态不同的时候，如正视摄像头和低头时，同一个人的这些几何特征会发生很大变化，因此，直接利用几何信息实现人脸识别是困难的。
- 签名：签名是一种行为特征。每个人的签名都有自己的特点，因此可以被用来进行身份识别。可以有两种方法获得个人的签名数据，通过摄像头、扫描仪等设备获取写在纸上的签字图像，或者通过写字板获取笔迹序列数据。签名识别的准确率对人的依赖性较大。有些人的签名具有很强的特异性，不容易被模仿，也具有很好的稳定性。而有些人的签名则常常变化，有些人的签名特异性不强，容易被模仿。另外，人们不是天生就会签名的，这需要训练。上述这些因素限制了签名识别的应用。当我们通过写字板获取笔迹序列数据时，可以利用笔划的位置、笔的速度等进行分类。但是，当我们得到的是签字图像时，提取笔划的位置和速度就很困难，因此，可以考虑采用签字的整体信息，如矩，纹理等。
- 步态：步态是人的一种行为特征，是指一个人走路时的姿态。一些人的走路姿态具有一定的特异性，因此，我们在日常生活中也步态用来识别一些熟悉的人。获取步态的简单方法是通过摄像头获取图像序列。但是，我们知道，步态很容易被模仿，有些人的步态会发生变化，有些人的步态不具有特异性。

图 8. 2 一张指纹图像，其中圆点表示纹线的断点或分叉点，圆点后的短线表示该点的方向。十字表示该指纹的中心

以上是生物特征识别的几个例子。实际上，还有很多其它生物特征识别问题，如：掌纹，虹膜，声音等。不同的生物特征需要不同的设备来获取数据，它们也具有不同的特点。因此，在设计一个生物特征识别产品的时候，就需要综合考虑产品的应用环境和条件，选择适当的生物特征以设计和实现产品。

正如前面所讨论的，最简单的特征选择方法是根据专家的知识挑选那些对分类最有影响的特征。而这和具体对象联系十分密切，涉及到研究对象本身的各种物理规律，也涉及其他的研究领域，如图像处理，信号处理，计算机视觉等。这些超出了本教材的讨论范围，我们不对它们专门讨论。应当指出，在很多实际问题中，物理和结构特征对分类是非常重要的，在实际构造一个识别系统时常常把它们作为基本特征而用到分类器的设计中。而本章讨论的重点则是根据样本来选择并提取数学特征。下面我们介绍一些主要的方法。

8.2. Fisher 线性判别

在模式识别中，我们得到的常常是高维数据。例如：在图像识别问题中，如果我们把每一个像素看作一个特征，一个 100*100 的图像就是一个 10000 维的向量；在文本分析中，我们常常把每一个字或词看作是一个特征，每篇文章就是一个上万维的特征向量。

在特征提取中，我们的目的是把高维特征空间变换为低维空间，使在低维空间中更好地分类。不仅如此，应用统计方法解决模式识别问题时，一再碰到的问题之一是维数问题。在低维空间里解析上或计算上行得通的方法，在高维空间里往往行不通。因此，降低维数有时就成为处理实际问题的关键。

由于线性判别函数易于分析，所以关于这方面的研究特别多。历史上，这一工作是从 R.A.Fisher（1936 年）的经典论文开始的。

我们考虑把 d 维空间样本投影到一条直线上，形成一维空间，即把维数压缩到一维。这在数学上总是容易办到的。然而，即使样本在 d 维空间里形成若干紧凑的互相分得开的集群，若把它们投影到一条任意的直线上，也可能使几类样本混在一起而变得无法识别。但在一般情况下，总可以找到某个方向，使在这个方向的直线上，样本的投影能分开得最好。问题是如何根据实际情况找到这条最好的、最易于分类的投影线。这就是 Fisher 法所要解决的基本问题（见图 8.3）。

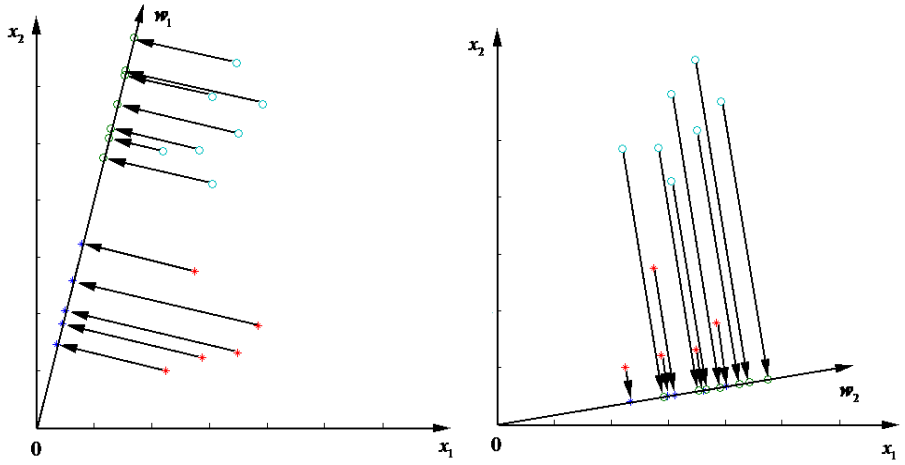


图 8. 3 Fisher 线性判别的基本原理

8.2.1. Fisher 线性判别的基本原理

首先，我们讨论从 d 维空间到一维空间的一般数学变换方法。假设有一集合 $\mathscr{X}$ 包含 N 个 d 维样本 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots \boldsymbol{x}_N$，其中 N_1 个属于 ω_1 类的样本记为子集 $\mathscr{X}_1$，N_2 个属于 ω_2 类的样本记为 $\mathscr{X}_2$。若对 $\boldsymbol{x}_n$ 的分量作线性组合可得标量

$$y_n = \boldsymbol{w}^T \boldsymbol{x}_n, n = 1, 2, \cdots, N_i \tag{8-1}$$

这样便得到 N 个一维样本 y_n 组成的集合，并可分为两个子集 $\mathscr{Y}_1$ 和 $\mathscr{Y}_2$。从几何上看，如果 $\|\boldsymbol{w}\| = 1$，则每个 y_n 就是相对应的 $\boldsymbol{x}_n$ 到方向为 $\boldsymbol{w}$ 的直线上的投影。实际上，$\boldsymbol{w}$ 的绝对值是无关紧要的，它仅使 y_n 乘上一个比例因子，重要的是选择 $\boldsymbol{w}$ 的方向。$\boldsymbol{w}$ 的方向不同，将使样本投影后的可分离程度不同，从而直接影响识别效果。因此，前述所谓寻找最好投影方向的问题，在数学上就是寻找最好的变换向量 $\boldsymbol{w}$ 的问题。

在定义 Fisher 准则函数之前，我们先定义几个必要的基本参量。

8.2.2. 在 d 维样本空间

（1） 各类样本均值向量 $\boldsymbol{m}_i$

$$\boldsymbol{m}_i = \frac{1}{N} \sum_{x \in \mathscr{X}_i} \boldsymbol{x}, i = 1, 2 \tag{8-2}$$

（2） 样本类内离散度矩阵 S_i 和总类内离散度矩阵 S_w

$$S_i = \sum_{\boldsymbol{x}\in\mathcal{Y}_i} (\boldsymbol{x}-\boldsymbol{m}_i)(\boldsymbol{x}-\boldsymbol{m}_i)^T, i=1,2 \tag{8-3}$$

$$S_w = S_1 + S_2 \tag{8-4}$$

（3） 样本类间离散度矩阵 S_b [1]

$$S_b = (\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^T \tag{8-5}$$

其中 S_w 是对称半正定矩阵，而且当 $N > d$ 时通常是非奇异的。S_b 也是对称半正定矩阵，在两类条件下，它的秩最大等于 1 。

8.2.3. 在一维投影空间

（1）各类样本均值 $\tilde{m}_i$

$$\tilde{m}_i = \frac{1}{N}\sum_{y\in\mathcal{Y}} y, i=1,2 \tag{8-6}$$

（2）样本类内离散度 $\tilde{S}_i^2$ 和总类内离散度 $\tilde{S}_w$

$$\tilde{S}_i^2 = \sum_{y\in\mathcal{Y}} (y-\tilde{m}_i)^2, i=1,2 \tag{8-7}$$

$$\tilde{S}_w = \tilde{S}_1^2 + \tilde{S}_2^2 \tag{8-8}$$

现在我们来定义 Fisher 准则函数。我们希望投影后，在一维 Y 空间里各类样本尽可能分得开些，即希望两类均值之差（$\tilde{m}_1 - \tilde{m}_2$）越大越好；同时希望各类样本内部尽量密集，即希望类内离散度越小越好。因此，我们可以定义 Fisher 准则函数为

$$J_F(\boldsymbol{w}) = \frac{(\tilde{m}_1 - \tilde{m}_2)^2}{\tilde{S}_1^2 + \tilde{S}_2^2} \tag{8-9}$$

显然应该寻找使 $J_F(\boldsymbol{w})$ 的分子尽可能大，而分母尽可能小，也就是使 $J_F(\boldsymbol{w})$ 尽可能大的 $\boldsymbol{w}$ 作为投影方向。但式(8-9)的 $J_F(\boldsymbol{w})$ 并不显含 $\boldsymbol{w}$ ，因此必须设法将 $J_F(\boldsymbol{w})$ 变成 $\boldsymbol{w}$ 的显函数。由式(8-6)可推出，

---

[1]若考虑先验概率，可以定义

$$S_w = P(\omega_1)S_1 + P(\omega_2)S_2$$

$$S_b = P(\omega_1)P(\omega_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^T$$

$$\tilde{\boldsymbol{m}}_i = \frac{1}{N_i}\sum_{y\in\mathscr{Y}_i} y = \frac{1}{N_i}\sum_{\boldsymbol{x}\in\mathscr{X}_i} \boldsymbol{w}^T\boldsymbol{x}$$

$$= \boldsymbol{w}^T\left(\frac{1}{N_i}\sum_{\boldsymbol{x}\in\mathscr{X}_i}\boldsymbol{x}\right) = \boldsymbol{w}^T\boldsymbol{m}_i \tag{8-10}$$

这样，式(8-9)的分子便成为

$$\begin{aligned}(\tilde{\boldsymbol{m}}_1 - \tilde{\boldsymbol{m}}_2)^2 &= (\boldsymbol{w}^T\boldsymbol{m}_1 - \boldsymbol{w}^T\boldsymbol{m}_2)^2 \\ &= \boldsymbol{w}^T(\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^T\boldsymbol{w} = \boldsymbol{w}^T S_b\boldsymbol{w}\end{aligned} \tag{8-11}$$

现在再来考察 $J_F(\boldsymbol{w})$ 的分母与 $\boldsymbol{w}$ 的关系。

$$\begin{aligned}\tilde{S}_i^2 &= \sum_{y\in Y_i}(y - \tilde{m}_i)^2 = \sum_{\boldsymbol{x}\in\chi_i}(\boldsymbol{w}^T\boldsymbol{x} - \boldsymbol{w}^T\boldsymbol{m}_i)^2 \\ &= \boldsymbol{w}^T\left[\sum_{\boldsymbol{x}\in\chi_i}(\boldsymbol{x} - \boldsymbol{m}_i)(\boldsymbol{x} - \boldsymbol{m}_i)^T\right]\boldsymbol{w} = \boldsymbol{w}^T S_i\boldsymbol{w}\end{aligned}$$

因此，

$$\tilde{S}_1^2 + \tilde{S}_2^2 = \boldsymbol{w}^T(S_1 + S_2)\boldsymbol{w} = \boldsymbol{w}^T S_w\boldsymbol{w} \tag{8-12}$$

将式(8-11)，式(8-12)代入式(8-9)，可得

$$J_F(\boldsymbol{w}) = \frac{\boldsymbol{w}^T S_b\boldsymbol{w}}{\boldsymbol{w}^T S_w\boldsymbol{w}} \tag{8-13}$$

(8-13)给出的就是 Fisher 准则。

下面求使 $J_F(\boldsymbol{w})$ 取极大值时的 $\boldsymbol{w}^*$。式(8-13)中的 $J_F(\boldsymbol{w})$ 是广义 Rayleigh 商，可以用 Lagrange 乘子法求解（参见附录 A4）。令分母等于非零常数，即令

$$\boldsymbol{w}^T S_w\boldsymbol{w} = c \neq 0$$

定义 Lagrange 函数为

$$L(\boldsymbol{w}, \lambda) = \boldsymbol{w}^T S_B\boldsymbol{w} - \lambda(\boldsymbol{w}^T S_w\boldsymbol{w} - c) \tag{8-14}$$

式中 λ 为 Lagrange 乘子。将式(8-14)对 $\boldsymbol{w}$ 求偏导数，得

$$\frac{\partial L(\boldsymbol{w}, \lambda)}{\partial \boldsymbol{w}} = S_b\boldsymbol{w} - \lambda S_w\boldsymbol{w}$$

令偏导数为零，得

$$S_b\boldsymbol{w}^* - \lambda S_w\boldsymbol{w}^* = 0$$

即

$$S_b\boldsymbol{w}^* = \lambda S_w\boldsymbol{w}^* \tag{8-15}$$

其中 $\boldsymbol{w}^*$ 就是 $J_F(\boldsymbol{w})$ 的极值解。因为 S_w 非奇异，式(8-15)两边左乘 S_w^{-1}，可得

$$S_w^{-1}S_b\boldsymbol{w}^* = \lambda\boldsymbol{w}^* \tag{8-16}$$

因此，求解 $\boldsymbol{w}^*$ 就转变为求 $S_w^{-1}S_b$ 的特征值问题。

但在我们这个问题中，利用式(8-5) S_b 的定义，式(8-16)左边的 $S_b\boldsymbol{w}^*$ 可以写成

$$S_b\boldsymbol{w}^* = (\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^T\boldsymbol{w}^* = (\boldsymbol{m}_1 - \boldsymbol{m}_2)R$$

式中 $R = (\boldsymbol{m}_1 - \boldsymbol{m}_2)^T\boldsymbol{w}^*$

为一标量，所以 $S_b\boldsymbol{w}^*$ 总是在向量 $(\boldsymbol{m}_1 - \boldsymbol{m}_2)$ 的方向上。由于我们的目的是寻找最好的投影方向， $\boldsymbol{w}$ 的比例因子对此并无影响，因此，从式(8-16)可得

$$\lambda\boldsymbol{w}^* = S_w^{-1}(S_b\boldsymbol{w}^*) = S_w^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2)R$$

从而可得

$$\boldsymbol{w}^* = \frac{R}{\lambda}S_w^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2) \tag{8-17}$$

忽略比例因子 $\frac{R}{\lambda}$，得

$$\boldsymbol{w}^* = S_w^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2) \tag{8-18}$$

$\boldsymbol{w}^*$ 就是使 Fisher 准则函数 $J_F(\boldsymbol{w})$ 取极大值时的解，也就是 d 维 X 空间到一维 Y 空间的最好投影方向。有了 $\boldsymbol{w}^*$，利用式(8-1)，就可以把 d 维样本 $\boldsymbol{x}_n$ 投影到一维，这实际上是多维空间到一维空间的一种映射。

以上做的全部工作是将 d 维空间的样本集 $\mathscr{X}$ 映射成一维样本集 $\mathscr{Y}$，这个一维空间的方向 $\boldsymbol{w}^*$ 是相对于 Fisher 准则 $J_F(\boldsymbol{w})$ 为最好的。但至此，我们还没有解决分类问题。然而，我们已将 d 维分类问题转化为一维分类问题了。实际上，只要确定一个阈值 y_0，将投影点 y_n 与 y_0 相比较，便可做出决策。在此，我们只简单介绍几种一维分类问题的基本原则。

(1) 当维数 d 和样本数 N 都很大时，可采用贝叶斯决策规则，从而获得一种在一维空间的“最优”分类器。

(2) 如果上述条件不满足，也可利用先验知识选定分界阈值点 y_0，如选择

$$y_0^{(1)} = \frac{\tilde{m}_1 + \tilde{m}_2}{2} \tag{8-19}$$

$$y_0^{(2)} = \frac{N_2\tilde{m}_1 + N_1\tilde{m}_2}{N_1 + N_2} = \tilde{m} \tag{8-20}$$

8.1 $$y_0^{(3)} = \frac{\tilde{m}_1 + \tilde{m}_2}{2} + \frac{\ln(P(\omega_1)/P(\omega_2))}{N_1 + N_2 - 2} \qquad (8\text{-}21)$$

式中 $P(\omega_1)$ 和 $P(\omega_2)$ 分别为 ω_1 类和 ω_2 类样本的先验概率。这样，对于任意给定的未知样本 $\boldsymbol{x}$，只要计算它的投影点 y，

$$y = \boldsymbol{w}^{*T}\boldsymbol{x}$$

再根据决策规则

$$y \gtrless y_0 \rightarrow \boldsymbol{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$$

就可判断 $\boldsymbol{x}$ 属于什么类别。

8.3. 类别可分离性判据

Fisher 线性判别是要通过线性变换找这样一个特征，该特征使得 Fisher 准则最大。这是特征提取的一个特殊情况。更一般地看，特征选择与提取的任务是求出一组（而不限制为一个）对分类最有效的特征。当然，和 Fisher 线性判别一样，我们也需要一个定量的准则(或称判据)来衡量特征对分类的有效性。具体说来，把一个高维空间变换为低维空间的映射是很多的，哪种映射对分类最有利，需要一个比较标准。从 D 个原始特征中选择出 d 个特征的各种可能组合也是很多的，哪种组合的分类效果最好，也要有一个比较标准。

大家可能很自然地想到，既然我们的目的是设计分类器，那末用分类器的错误概率作为标准就行了，也就是说，使分类器错误概率最小的那组特征，就应当是一组最好的特征。从理论上说，这是完全正确的，但在实用中却有很大困难。

回想一下错误概率的计算公式(参考“2.4 关于分类器的错误率问题”)就会发现，即使在类条件分布密度已知的情况下错误概率的计算也很复杂，何况实际问题中这一分布常常不知道，这使得直接用错误概率作为标准来分析特征的有效性比较困难。

我们希望找出另一些更实用的标准以衡量各类间的可分性(有的地方叫类别可分离性判据)，并希望可分性判据满足下列几条要求：

(1) 与错误概率(或错误概率的上界及下界)有单调关系，这样使判据取最大值的效果一般说来其错误概率也较小。

(2) 当特征独立时有可加性。即

$$J_{ij}(x_1, x_2, ..., x_d) = \sum_{k=1}^{d} J_{ij}(x_k)$$

这里 J_{ij} 是第 i 类和第 j 类的可分性准则函数，J_{ij} 愈大，两类的分离程度就愈大，$x_1, x_2, ..., x_d$ 是一定类别相应特征的随机变量。

(3) 度量特性：

$J_{ij} > 0$，当 $i \neq j$ 时

$J_{ij}=0$，当$i=j$时

$J_{ij}=J_{ji}$

(4) 单调性，即加入新的特征时，判据不减小。

$$J_{ij}(x_1,x_2,...,x_d)\le J_{ij}(x_1,x_2,...,x_d,x_{d+1})$$

很多人在这方面做了不少工作，提出了各种判据，但还没有取得完全满意的结果，下面对几种常用的判据进行讨论。

8.3.1. 基于类内类间距离的可分离性判据

直观上看，Fisher 准则希望变换后，样本的类内差异小，类间差异大。我们把这一直观扩展到一般的特征提取准则中，即希望变换后样本的类间离散度尽量大，类内离散度尽量小。这样，人们提出下面各种判据

$$J_2=\mathrm{tr}(S_w^{-1}S_b) \tag{8-22}$$

$$J_3=\ln[\frac{|S_b|}{|S_w|}] \tag{8-23}$$

$$J_4=\frac{\mathrm{tr}S_b}{\mathrm{tr}S_w} \tag{8-24}$$

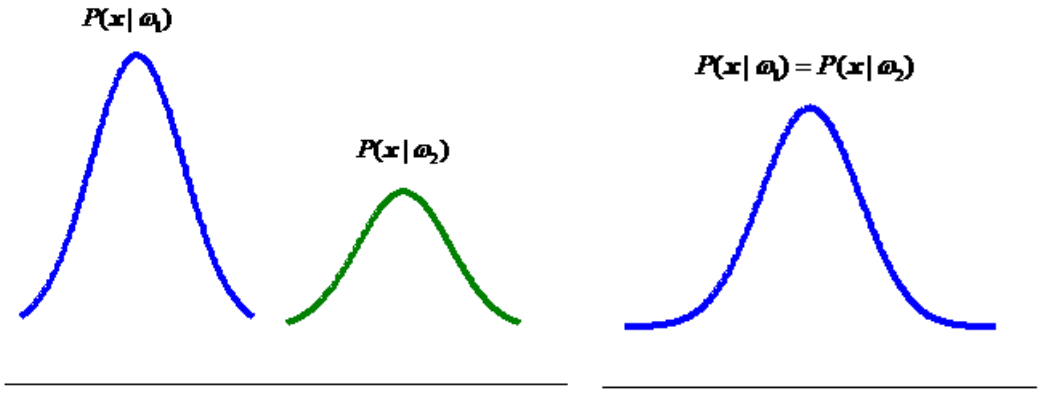
$$J_5=\frac{|S_w+S_b|}{|S_w|} \tag{8-25}$$

其中S_b为类间离散度矩阵，S_w为类内离散度矩阵。

8.3.2. 基于概率分布的可分性判据

上面介绍的距离准则是直接从各类样本算出的，没有考虑各类的概率分布，不能确切表明各类交叠的情况，因此与错误概率没有直接联系。优点是计算方便，直观概念清楚。下面讨论一些基于概率分布的可分性判据。

先研究一下两类的情况：如图 8. 4，其中图 8. 4 (a)为完全可分的情况，图 8. 4(b) 为完全不可分的情况。



(a) (b)

图 8.4 完全可分与完全不可分情况

假定先验概率相等，若对所有使 $P(\boldsymbol{x}\,|\,\omega_2)\neq 0$ 的点有 $P(\boldsymbol{x}\,|\,\omega_1)=0$，如图 8.4(a)，则两类为完全可分的；相反，如果对所有 $\boldsymbol{x}$ 都有 $P(\boldsymbol{x}\,|\,\omega_1)=P(\boldsymbol{x}\,|\,\omega_2)$，如图 8.4(b) ，则两类完全不可分。

分布密度的交叠程度可用 $P(\boldsymbol{x}\,|\,\omega_1)$ 及 $P(\boldsymbol{x}\,|\,\omega_2)$ 这两个分布密度函数之间的距离 J_p 来度量。任何函数 $J(\bullet)=\int g[p(\boldsymbol{x}\,|\,\omega_1),p(\boldsymbol{x}\,|\,\omega_2),P_1,P_2]d\boldsymbol{x}$，如果满足下述条件：

(1) J_p 为非负，即 $J_p\geq 0$；

(2) 当两类完全不交叠时 J_p 取最大值，即若对所有 x 有 $P(\boldsymbol{x}\,|\,\omega_2)\neq 0$ 时 $P(\boldsymbol{x}\,|\,\omega_1)=0$，则 $J_p=Max$；

(3) 当两类分布密度相同时， J_p 应为零，即若 $P(\boldsymbol{x}\,|\,\omega_1)=P(\boldsymbol{x}\,|\,\omega_2)$，则 $J_p=0$。都可用来作为类分离性的概率距离度量。

下面讨论一些常用的概率距离度量。

1. Bhattacharyya 距离和 Chernoff 界限

Bhattacharyya 距离的定义是：

$$J_B=-\ln\int[p(\boldsymbol{x}\,|\,\omega_1)p(\boldsymbol{x}\,|\,\omega_2)]^{1/2}d\boldsymbol{x} \tag{8-26}$$

它与错误概率的上界有直接关系。因为

$$P_e=P(\omega_1)\int_{\mathcal{R}}p(\boldsymbol{x}\,|\,\omega_1)d\boldsymbol{x}+P(\omega_2)\int_{\overline{\mathcal{R}}}p(\boldsymbol{x}\,|\,\omega_2)d\boldsymbol{x}$$

$$= \int_{-\infty}^{\infty} \min\{P(\omega_1)p(\boldsymbol{x}\mid\omega_1), P(\omega_2)p(\boldsymbol{x}\mid\omega_2)\}d\boldsymbol{x}$$

$$\leq \int_{-\infty}^{\infty} \{P(\omega_1)P(\omega_2)p(\boldsymbol{x}\mid\omega_1)p(\boldsymbol{x}\mid\omega_2)\}^{1/2}d\boldsymbol{x}$$

$$= [P(\omega_1)P(\omega_2)]^{1/2}\int_{-\infty}^{\infty} \{p(\boldsymbol{x}\mid\omega_1)p(\boldsymbol{x}\mid\omega_2)\}^{1/2}d\boldsymbol{x}$$

$$= [P(\omega_1)P(\omega_2)]^{1/2}\exp\{-J_B\}$$

式中 $\mathscr{R}$ 为 $P(\boldsymbol{x}\mid\omega_2) > P(\boldsymbol{x}\mid\omega_1)$ 的区域，而 $\overline{\mathscr{R}}$ 为 $P(\boldsymbol{x}\mid\omega_1) > P(\boldsymbol{x}\mid\omega_2)$ 的区域。

另一与此相似的判据称为 Chernoff 界限 J_c：

$$J_C = -\ln\int p^s(\boldsymbol{x}\mid\omega_1)p^{1-s}(\boldsymbol{x}\mid\omega_2)d\boldsymbol{x} \tag{8-27}$$

式中 s 是在［0，1］区间取值的一个参数。

不难看出，取 $s = 0.5$ 时 $J_C = J_B$。

2. 散度

我们已经知道，两类密度函数的似然比或负对数似然比对分类来说是一个重要的度量，设有二类 ω_i 及 ω_j，其对数似然比为：

$$l_{ij}(\boldsymbol{x}) = \ln\frac{p(\boldsymbol{x}\mid\omega_i)}{p(\boldsymbol{x}\mid\omega_j)} \tag{8-28}$$

它可以提供 ω_i 对 ω_j 类的可分性信息，对 ω_i 类的平均可分性信息应为

$$I_{ij}(\boldsymbol{x}) = E[l_{ij}(\boldsymbol{x})] = \int_X p(\boldsymbol{x}\mid\omega_i)\ln\frac{p(\boldsymbol{x}\mid\omega_i)}{p(\boldsymbol{x}\mid\omega_j)}d\boldsymbol{x} \tag{8-29}$$

同样对 ω_j 类的平均可分性信息为：

$$I_{ji}(\boldsymbol{x}) = E[l_{ji}(\boldsymbol{x})] = \int_X p(\boldsymbol{x}\mid\omega_j)\ln\frac{p(\boldsymbol{x}\mid\omega_j)}{p(\boldsymbol{x}\mid\omega_i)}d\boldsymbol{x} \tag{8-30}$$

因此，可定义散度 J_D 为区分 ω_i 类和 ω_j 类的总的平均信息，它等于两类平均可分信息之和：

$$J_D = I_{ij} + I_{ji} = \int_X [p(\boldsymbol{x}\mid\omega_i) - p(\boldsymbol{x}\mid\omega_j)]\ln\frac{p(\boldsymbol{x}\mid\omega_i)}{p(\boldsymbol{x}\mid\omega_j)}d\boldsymbol{x} \tag{8-31}$$

散度与 Bhattacharyya 距离都满足本节开始所提出的类别可分离性判据的条件。

8.3.3. 基于熵函数的可分性判据

最佳分类器由后验概率确定，所以可由特征的后验概率分布来衡量它对分类的有效性。如果对某些特征，各类后验概率是相等的，即

$$P(\omega_i \mid \boldsymbol{x}) = \frac{1}{c}$$

其中 c 为类别数，则我们将无从确定样本所属类别，或者我们只能任意指定 $\boldsymbol{x}$ 属于某一类(假定先验概率相等或不知道)，此时之错误概率为：

$$P_e = 1 - \frac{1}{c} = \frac{c-1}{c}$$

再看另一个极端情况，如果能有一组特征使得

$$P(\omega_i \mid \boldsymbol{x}) = 1 \text{ 且 } P(\omega_j \mid \boldsymbol{x}) = 0, \forall j \neq i$$

此时可以肯定 $\boldsymbol{x}$ 划归 ω_i 类，而错误概率为零。

可见后验概率分布愈集中，错误概率就愈小。后验概率分布愈平缓(接近均匀分布)则分类错误概率就愈大。

为了衡量后验概率分布的集中程度，需要规定一个定量指标，我们可以借助于信息论中关于熵的概念。

设 ω 为可能取值 ω_i，$i = 1, 2, ..., c$ 的一个随机变量，它的取值依赖于分布密度为 $p(\boldsymbol{x})$ 的随机向量 $\boldsymbol{x}$ (特征向量)，即给定 $\boldsymbol{x}$ 后 ω_i 的概率是 $p(\omega_i \mid \boldsymbol{x})$。现在进行观察变量 $\boldsymbol{x}$ 以及相应的 ω 值的实验。我们的问题是：给定某一 $\boldsymbol{x}$ 后，我们从观察 ω 的结果中得到了多少信息?或者说 ω 的不确定性减少了多少?显然，假使对某一 $\boldsymbol{x}$，有 $P(\omega_i \mid \boldsymbol{x}) = 1$，且 $P(\omega_j \mid \boldsymbol{x}) = 0, \forall j \neq i$。则观察结果必然是 $\omega = \omega_i$。因此观察的结果并未使我们得到任何信息。反之若对所有的 i， $p(\omega_i \mid \boldsymbol{x})$ 都相等，我们只能任意猜测 ω 的可能结果，而从观察到实际发生的 ω_i 事件中得到的信息量就不再等于零，而相应于观察前的不确定程度。

从特征提取的角度看，显然用具有最小不确定性的那些特征进行分类是有利的。在信息论中用“熵”作为不确定性的度量，它是的函数 $p(\omega_1 \mid \boldsymbol{x})$， $p(\omega_2 \mid \boldsymbol{x})$， ...， $p(\omega_c \mid \boldsymbol{x})$ 即

$$H = J_c[p(\omega_1 \mid \boldsymbol{x}), ..., p(\omega_c \mid \boldsymbol{x})]$$

这个函数应该有下列性质：

(1) 熵为正且对称：

$$H_c(P_1, P_2, ..., P_c) = H_c(P_2, P_1, ..., P_c) = ... = H_c(P_c, ..., P_1) \geq 0;$$

(2) 若 $P_{i_0} = 1$，且 $P_i = 0 (1 \leq i \leq c, i \neq i_0)$，则 $H_c(P_1, P_2, ..., P_c) = 0$；

(3) $H_{c+1}(P_1,P_2,...,P_c,0)=H_c(P_1,P_2,...,P_c)$；

(4) 对于任意的概率分布，$P_i\geq 0$，$(i=1,\cdots,c)$，$\sum_{i=1}^{c}P_i=1$，有

$$H_c(P_1,P_2,\cdots,P_c)=H_c(\frac{1}{c},\frac{1}{c},\cdots,\frac{1}{c})$$；

(5) 对所有事件，熵函数是连续函数。

满足上述性质的一族信息度量是如下形式的广义熵：

$$J_c^{\alpha}[P(\omega_1\mid \boldsymbol{x}),P(\omega_2\mid \boldsymbol{x}),\cdots P(\omega_c\mid \boldsymbol{x})]=(2^{1-\alpha}-1)^{-1}[\sum_{i=1}^{c}P^{\alpha}(\omega_i\mid \boldsymbol{x})-1] \qquad (8\text{-}32)$$

式中α是一个实的正参数，$\alpha\neq 1$。

不同的α值可以得到不同的熵分离度量。例如，当α趋近于1时，据 L’Hospital 法则有：

$$J_c^{1}[P(\omega_1\mid \boldsymbol{x}),P(\omega_2\mid \boldsymbol{x}),\cdots P(\omega_c\mid \boldsymbol{x})]$$

$$=\lim_{\alpha\to 1}(2^{1-\alpha}-1)^{-1}[\sum_{i=1}^{c}P^{\alpha}(\omega_i\mid \boldsymbol{x})-1]$$

$$=-\sum_{i=1}^{c}P(\omega_i\mid \boldsymbol{x})\log_2 P(\omega_i\mid \boldsymbol{x})$$

这就是 Shannon 熵。

当$\alpha=2$时，得到平方熵：

$$J_c^{2}[P(\omega_1\mid \boldsymbol{x}),P(\omega_2\mid \boldsymbol{x}),\cdots P(\omega_c\mid \boldsymbol{x})]=2[1-\sum_{i=1}^{c}P^{2}(\omega_i\mid \boldsymbol{x})] \qquad (8\text{-}33)$$

显然，为了对所提取的特征进行评价，我们要计算空间每一点的熵函数。在熵函数取值较大的那一部分空间，不同类的样本必然在较大的程度上互相重叠。因此熵函数的期望值

$$J(\bullet)=E\{J_c^{\alpha}[P(\omega_1\mid \boldsymbol{x}),P(\omega_2\mid \boldsymbol{x}),\cdots,P(\omega_c\mid \boldsymbol{x})]\}$$

可以表征类别的分离程度，它可用来作为所提取特征的分类性能的评价指标。我们在关于用K-L变换进行特征提取的第9章中还将用到熵的概念。

8.3.4. 类别可分离性判据的直接应用举例

类别可分离性判据在图像分割中得到了有效的应用。例如，图像二值化(使图像中与物体对应的像素为1，与背景对应的像素为0)的一个常用的方法就是确定一个灰度阈值，使图像中灰度值小于(或大于)此阈值的像素为物体(取值为1)，否则为背景(取值为0)。其实质就是一个两类分类问题：灰度就是特征值，阈值就是分界点。 Otsu 用类内方差、类间方差和总方差的比值作为判据给出了一个很有效的分类方法，其基本内容如下：

设图像有L个灰度，n_i是灰度为i的像素数，图像像素总数$N=n_1+n_2+\cdots+n_L$。具

有给定灰度 i 的像素的概率 $p_i = \frac{n_i}{N}$

显然　$p_i \geq 0$，$\sum_{i=1}^{L} p_i = 1$

设二值化的阈值为 k，图像中像素灰度值小于或等于 k 的平均灰度值为 g_0，大于 k 的所有像素灰度值的平均值为 g_1，则类间方差是

$$\sigma_B^2 = \omega_0 (g_0 - k)^2 + \omega_1 (g_1 - k)^2$$

其中 $\omega_0 = \sum_{i=1}^{k} p_i$，$\omega_1 = \sum_{i=k+1}^{L} p_i = 1 - \omega_0$。

定义 σ_B^2 为可分离性判据，因此可以方便地求出 k 值，使 σ_B^2 为最大，从而完成图像二值化。上述思路可以容易地扩展到把图像分成三类或更多的情况。

8.4. 特征提取

8.4.1. 按类内类间距离度量的特征提取方法

在上一个小节中，我们给出了几个可分离性判据：J_2，J_3，J_4，J_5，其都涉及到离散度矩阵。这里我们对离散度矩阵作些解释。一个在 d 维空间由 N 个点组成的聚类，其散布的程度可以用离散度来衡量。离散度的计算和我们规定的中心点有关。设 $\boldsymbol{a}$ 是 d 维空间中所选的中心点，从聚类的 N 个点中取出 d 个点，以 $\boldsymbol{a}$ 点为引点，作一个超平行四边形。例如，图 8.5 是二维空间中的一个平行四边形。对所有 N 中的 d 个点求出相应的 C_N^d 个超平行四边形的体积平方和对 N 个点的均值，就是该聚类对于中心点 $\boldsymbol{a}$ 的离散度。假使取聚类的均值点作为引点，则离散度为

$$S = |C|$$

式中 C 是聚类的协方差矩阵，就是我们上面所讨论的离散度矩阵。

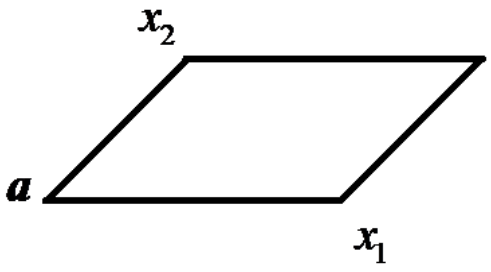


图 8.5 离散度的几何解释

从上述的离散度概念，我们可以看到，只有当类别数超过空间的维数时，类间离散度矩

阵才有相应的几何意义，因此我们在 J_5 中引入了一个总离散度

$$S_T = |\Sigma| = |S_w + S_b|$$

式中

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T$$

在上面的各种判据中， J_2， J_3 与 J_5 在任何非奇异线性变换下不变， J_4 则与坐标系有关。对此我们以 J_2 为例证明如下：

所谓线性变换的不变性就是说在用这个判据求得一个 d 维子空间后，对任何非奇异的 $d \times d$ 变换阵 A，$J_2(d)$ 是不变的，或者说在 d 维空间的任何线性变换都不改变 $J_2(d)$ 的值。

$$\mathrm{tr}(S_w^{*-1}S_b^*) = \mathrm{tr}\{[AS_wA^T]^{-1}[AS_bA^T]\} = \mathrm{tr}[(A^T)^{-1}S_w^{-1}A^{-1}AS_bA^T]$$

$$= \mathrm{tr}[(A^T)^{-1}S_w^{-1}S_bA^T] = \mathrm{tr}[S_w^{-1}S_bA^T(A^T)^{-1}] = \mathrm{tr}\ S_w^{-1}S_b$$

用以上判据进行特征提取的步骤如下：

假设我们有 D 个原始特征： $\boldsymbol{y} = [y_1, y_2, \cdots, y_D]^T$ ， 希望通过线性映射压缩为 d 个特征： $\boldsymbol{x} = [x_1, x_2, \cdots, x_d]^T$ ， 其变换关系为

$$\boldsymbol{x} = W^T\boldsymbol{y}\text{，}\ W \text{ 为 } D \times d \text{ 矩阵}$$

令 S_w， S_b 为原空间(即 $\boldsymbol{y}$ 的)离散度矩阵， S_w^*， S_b^* 为映射后(即 $\boldsymbol{x}$ 的)离散度矩阵：

$$S_b^* = W^TS_bW\text{，}\ S_w^* = W^TS_wW$$

经变换后的 J_2 变为

$$J_2(W) = \mathrm{tr}[(\mathrm{W}^TS_wW)^{-1}W^TS_bW] \tag{8-34}$$

将此式对 W 的各分量求偏导数并令其为零可以确定一个 W 值。我们不进行详细推导，只给出最后结论。对 J_2， J_3，J_5 来说使判据达最大的变换 W 如下：

设矩阵 $S_w^{-1}S_b$ 的本征值为 λ_1， λ_2， $\cdots$， λ_D，按大小顺序排列为：

$$\lambda_1 \geq \lambda_2 \geq \cdots \lambda_D$$

则选前 d 个本征值对应的本征向量作为 W 。即

此时 $J_2(W)$ 为：

$$J_2(W)=\sum_{i=1}^{d}\lambda_i \tag{8-35}$$

此结论对 J_4 判据也适用。

例 8　1 给定先验概率相等的两类，其均值向量分别为：$\mu_1=[1,3-1]^T$ 和 $\mu_2=[-1,-1,1]^T$，协方差矩阵是

$$\sum_1=\begin{bmatrix}4&1&0\\1&4&0\\0&0&1\end{bmatrix},\quad \sum_2=\begin{bmatrix}2&1&0\\1&2&0\\0&0&1\end{bmatrix}$$

求用 J_5 判据的最优特征提取。

解　根据前面的分析，应先求 $S_w^{-1}S_b$，再求此矩阵的本征阵。

令　混合均值 $\boldsymbol{\mu}=\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)=[0,1,0]^T$

类间离散度矩阵：

$$S_b=\frac{1}{2}\sum_{i=1}^{2}(\boldsymbol{\mu}_i-\boldsymbol{\mu})(\boldsymbol{\mu}_i-\boldsymbol{\mu})^T=\frac{1}{4}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T$$

类内离散度矩阵：

$$S_w=\frac{1}{2}(\sum_1+\sum_2)=\begin{bmatrix}3&1&0\\1&3&0\\0&0&1\end{bmatrix}$$

$$S_w^{-1}=\frac{1}{8}\begin{bmatrix}3&-1&0\\-1&3&0\\0&0&8\end{bmatrix}$$

由于 $S_w^{-1}S_b$ 的秩是 $c-1=1$，因此 $S_w^{-1}S_b$ 只有一个非零本征值，W 是 $D\times 1$ 矩阵，即 $W=\boldsymbol{w}$。

为求 $S_w^{-1}S_b$ 的本征值应解方程：

$$S_w^{-1}S_b\boldsymbol{w}=\lambda_1\boldsymbol{w}$$

或

$$\frac{1}{4}S_w^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{w}=\lambda_1\boldsymbol{w}$$

$\frac{1}{4}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{w}$ 是标量，所以

$$\boldsymbol{w}=S_w^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)=\frac{1}{4}(1,5,-8)^T$$

这就是所要求的解。

现在进一步讨论 J_5。由于 $\sum$ 和 S_w 是对称矩阵，因此存在矩阵 U，使

$$U^T \sum U = \Lambda \tag{8-36}$$

和

$$U^T S_w U = I \tag{8-37}$$

从而有

$$J_5 = \frac{\left|U^T \sum U\right|}{\left|U^T S_w U\right|} = \prod_{j=1}^{d} \lambda_j$$

由式(8-36)和式(8-37)，有 $S_w^{-1} \sum U = U\Lambda$

因此 Λ 是 $S_w^{-1} \sum$ 的本征值矩阵。

因 $S_w^{-1} \sum = I + S_w^{-1} S_b$

所以 $(I + S_w^{-1} S_b)U = U\Lambda$

设 $S_w^{-1} S_b$ 的本征值矩阵是 $\tilde{\Lambda}$

则 $\tilde{\Lambda} = \Lambda - I$

因此

$$J_5 = \prod_{j=1}^{d} (1 + \tilde{\lambda}_j)$$

用 J_5 判据在处理多类问题时，不至于选取那些只对两类有很好可分性而对其余各类效果不大的特征。对于 J_2 来说，只要有一个 $\tilde{\lambda}_j$ 很大就会发生这种情况。从计算的角度看，J_4 和 J_1 则是最方便的，不需要存储任何矩阵。

8.4.2. 按概率距离判据的特征提取方法

在 8.3.2 小节中，我们介绍了一些基于概率分布的距离的判据，它们具有一系列优点，但正如文中第三部分指出的，这些判据只有当概率分布密度属于某种参数形式时才可写成便于计算的解析形式，我们下面研究当分布为多维正态且只有两类的情况。在 8.3.2 中所列的判据中 J_C 与 J_D 是常用的，J_B 是 J_C 的特殊情况，所以下面只研究 J_C 和 J_D 二个判据。

设原始特征 $\boldsymbol{y}$ 与二次特征 $\boldsymbol{x}$ 之间有映射关系：

$$\boldsymbol{x} = W^T \boldsymbol{y}$$

则原空间中一个矩阵 A 经映射后变为 A^*，它与 A 有以下关系：

$$A^* = W^T A W$$

映射后的概率距离也相应地变为 $J_C(W)$， $J_D(W)$。

我们下面只推导用 Chernoff 概率距离判据 J_C 进行特征提取的有关公式，使读者对这种方法有一个比较清楚的了解，对于某些其他的概率距离判据，我们直接给出有关结论。

当两类都是正态分布时 $J_C(W)$ 的表达式可写为：

$$\begin{aligned} J_C(W) = {} & \frac{1}{2} s(1-s) \mathrm{tr}\left\{ W^T M W [(1-s) W^T \sum_1 W + s W^T \sum_2 W]^{-1} \right\} \\ & + \frac{1}{2} \ln \left| (1-s) W^T \sum_1 W + s W^T \sum_2 W \right| - \frac{1}{2} (1-s) \ln \left| W^T \sum_1 W \right| \\ & - \frac{1}{2} s \ln \left| W^T \sum_2 W \right| \end{aligned} \tag{8-38}$$

式中 $M = (\mu_2 - \mu_1)(\mu_2 - \mu_1)^T$

注意， $J_C(W)$ 是一个标量，它可以对 W 的各分量求偏导数，从而得到梯度矩阵 $J_C^{'}(W)$ 如下：

$$\begin{aligned} J_C^{'} = {} & s(1-s)\{ M W [(1-s) W^T \sum_1 W + s W^T \sum_2 W]^{-1} \\ & - [(1-s) \sum_1 W + s \sum_2 W][(1-s) W^T \sum_1 W \\ & + s W^T \sum_2 W]^{-1} W^T M W [(1-s) W^T \sum_1 W + s W^T \sum_2 W]^{-1} \} \\ & + [(1-s) \sum_1 W + s \sum_2 W][(1-s) W^T \sum_1 W + s W^T \sum_2 W]^{-1} \\ & - (1-s) \sum_1 W (W^T \sum_1 W)^{-1} - s \sum_2 W (W^T \sum_2 W)^{-1} \end{aligned}$$

令上式等于零，并假定 $[(1-s) W^T \sum_1 W + s W^T \sum_2 W]$ 不等于零，则最优变换矩阵W一定满足下式：

$$\begin{aligned} & M W - [(1-s) \sum_1 W + s \sum_2 W][(1-s) W^T \sum_1 W + s W^T \sum_2 W]^{-1} W^T M W \\ & + \frac{1}{s(1-s)} [(1-s) \sum_1 W + s \sum_2 W] - \frac{1}{s} \sum_1 W [(1-s) I \\ & + s (W^T \sum_1 W)^{-1} W^T \sum_2 W] - \frac{1}{1-s} \sum_2 W [(1-s)(W^T \sum_2 W)^{-1} \end{aligned}$$

$$\times W^T \sum_1 W + sI] = 0$$

上式可进一步简化为：

$$MW - [(1-s)\sum_1 W + s\sum_2 W][(1-s)W^T \sum_1 W + sW^T \sum_2 W]^{-1} W^T MW$$

$$+\sum_1 W[I - (W^T \sum_1 W)^{-1} W^T \sum_2 W] + \sum_2 W[I - (W^T \sum_2 W)^{-1} W^T \sum_1 W] = 0 \quad (8\text{-}39)$$

上式对W是非线性的，因此不能直接求解而只能采用数值优化方法，但假使两类的协方差矩阵相等，或者两类的均值向量相等，我们可以得到相应的解析解：

1　$\sum_1 = \sum_2 = \sum$ 的情况

此时式(8-39)化简为

$$MW - \sum W(W^T \sum W)^{-1} W^T MW = 0 \quad (8\text{-}40)$$

设矩阵$(W^T \sum W)^{-1} W^T MW$的本征值矩阵和本征向量矩阵分别是Λ和U，即

$$(W^T \sum W)^{-1} W^T MWU = U\Lambda$$

则式(8-40)可改写为

$$\sum^{-1} MWU - WU\Lambda = 0 \quad (8\text{-}41)$$

令　　$V = WU$

显然V是$\sum^{-1} M$的本征向量矩阵，从而可得

$$W = VU^{-1}$$

是$\sum_1 = \sum_2 = \sum$条件下，用 Chernoff 概率距离判据的最优特征提取变换矩阵(或最优特征提取器)。由于J_C在任意非奇异变换下具有不变性，因此WU同样是最优特征提取器，换句话说，最优特征提取器由$\sum^{-1} M$的本征向量构成。但矩阵M的秩是 1，它只有一个非零本征值，而在这种情况下的$J_C(M)$为：

$$J_C(W) = \frac{1}{2} s(1-s)\mathrm{tr}\Lambda \quad (8\text{-}42)$$

对应于本征值为零的那些本征向量对$J_C(M)$没有影响，因此可以舍去，从而使$J_C(M)$最优的变换W是与矩阵$\sum^{-1} M$的非零本征值相应的单个本征向量$\boldsymbol{v}$，从式(8-41)可以看到

$$\boldsymbol{v} = \sum^{-1}(\mu_2 - \mu_1) \quad (8\text{-}43)$$

2　$\sum_1 \neq \sum_2$，但 $\mu_1 = \mu_2$ 的情况

这种情况下，Chernoff 概率距离简化为

$$J_C(W) = \frac{1}{2}\ln \frac{\left|(1-s)W^T \sum_1 W + sW^T \sum_2 W\right|}{\left|W^T \sum_1 W\right|^{1-s} \left|W^T \sum_2 W\right|^s} \tag{8-44}$$

相应的一次偏导数矩阵是

$$J_C^{'}(W) = \sum_1 W[I - (W^T \sum_1 W)^{-1} W^T \sum_2 W] + \sum_2 W[I - (W^T \sum_2 W)^{-1} W^T \sum_1 W]$$

令它等于零矩阵，可得

$$[\sum_2^{-1} \sum_1 W - W(W^T \sum_2 W)^{-1}(W^T \sum_1 W)][I - (W^T \sum_1 W)^{-1} W^T \sum_2 W] = 0$$

设 $I - (W^T \sum_1 W)^{-1} W^T \sum_2 W$ 不等于零。则

$$\sum_2^{-1} \sum_1 W - W(W^T \sum_2 W)^{-1} W^T \sum_1 W = 0$$

若矩阵 $(W^T \sum_2 W)^{-1} W^T \sum_1 W$ 的本征值矩阵是 Λ，本征向量矩阵是 U，即

$$(W^T \sum_2 W)^{-1} W^T \sum_1 W = U \Lambda U^{-1} \tag{8-45}$$

或

$$W^{-1} \sum_2^{-1} \sum_1 W = U \Lambda U^{-1}$$

即

$$\sum_2^{-1} \sum_1 WU - WU\Lambda = 0 \tag{8-46}$$

因此 $V = WU$ 是矩阵 $\sum_2^{-1} \sum_1$ 的本征向量矩阵，根据 Chernoff 距离判据对非奇异线性变换的不变性，最优坐标系统就是 $\sum_2^{-1} \sum_1$ 的本征向量系。

现在讨论一下怎样从坐标系中选出 d 个坐标轴使 $J_C(W)$ 最大。把从式(8-45)求得的 $W^T \sum_1 W$ 代入到式(8-44)中去，可得：

$$\begin{aligned} J_C(W) &= \frac{1}{2}\ln \frac{\left|(1-s)W^T \sum_2 WU\Lambda U^{-1} + sW^T \sum_2 W\right|}{\left|W^T \sum_2 WU\Lambda U^{-1}\right|^{1-s} \left|W^T \sum_2 W\right|^s} \\ &= \frac{1}{2}\ln \frac{\left|(1-s)U\Lambda U^{-1} + sI\right|}{\left|U\Lambda U^{-1}\right|^{1-s}} \\ &= \frac{1}{2}\ln \frac{\left|(1-s)\Lambda + sI\right|}{\left|\Lambda\right|^{1-s}} \end{aligned} \tag{8-47}$$

由于 Λ 和 I 都是对角线矩阵，所以

$$J_C(W)=\frac{1}{2}\ln\prod_{j=1}^{d}[(1-s)\lambda_j^s+s\lambda_j^{s-1}]=\frac{1}{2}\sum_{j=1}^{d}\ln[(1-s)\lambda_j^s+s\lambda_j^{s-1}] \tag{8-48}$$

为了要使 $J_C(W)$ 最大，应选择满足下列顺序关系

$$(1-s)\lambda_1^s+s\lambda_1^{s-1}\geq(1-s)\lambda_2^s+s\lambda_2^{s-1}$$

$$\geq\cdots\geq(1-s)\lambda_d^s+s\lambda_d^{s-1}$$

$$\geq\cdots\geq(1-s)\lambda_D^s+s\lambda_D^{s-1} \tag{8-49}$$

的前 d 个本征值所对应的本征向量组成变换矩阵 W 。

显然，取不同的 s 值会有不同的本征向量的排列顺序。为了得到最大的 $J_C(W)$ ，我们可以先设 $s=0.5$ ，以求出最优坐标轴 $\boldsymbol{v}_j$ ， $j=1,2,\cdots,d$ 。对于这些坐标轴，我们可以求出最优参数 s 以得到最大的 $J_C(W)$ 。对于这个新的参数 s ，重新求出最优坐标轴 $\boldsymbol{v}_j^{'}$ ，$j=1,2,\cdots,d$ 。重复上述步骤直到获得一组稳定的最优坐标轴为止。

3　一般情况

如前所述，在一般情况下得不到用解析式子表述的最优解。为了避免用数值优化方法求解，可以分别考虑两类均值向量和协方差矩阵有差别时的分类作用。下面讨论两种次优算法。

方法一　先假设两类的均值向量相等，因此用 $\sum_2^{-1}\sum_1$ 的本征向量系统作为特征提取器的候选坐标轴。然后用包含在类均值向量中的判别信息从这 D 个候选坐标轴中选出 d 个坐标轴作为特征提取器 W 。方法如下：

将式(8-45)改写为

$$U^{-1}(W^T\sum_2 W)^{-1}(U^{-1})^T[(U^{-1})^T]^{-1}W^T\sum_1 WU=\Lambda$$

即　　$$[(WU)^T\sum_2 WU]^{-1}(WU)^T\sum_1 WU=\Lambda$$

因此可选择 $V=WU$ 使 $V^T\sum_2 V=I$

以及　　$$V^T\sum_1 V=\Lambda$$

把上列两个关系式代入 $J_C(W)$ 的表达式(8-44)可得：

$$J_C(W)=\frac{1}{2}\sum_{j=1}^{d}\{s(1-s)[\boldsymbol{v}_j^T(\mu_2-\mu_1)]^2[(1-s)\lambda_j+s]^{-1}$$

$$+\ln[(1-s)\lambda_j^s+s\lambda_j^{s-1}]\}$$

要使 $J_C(W)$ 达最大，应按下列顺序

$$s(1-s)[\boldsymbol{v}_1^T(\mu_2-\mu_1)]^2[(1-s)\lambda_1+s]^{-1}+\ln[(1-s)\lambda_1^s+s\lambda_1^{s-1}]$$

$$\geq\cdots\geq s(1-s)[\boldsymbol{v}_d^T(\mu_2-\mu_1)]^2[(1-s)\lambda_d+s]^{-1}+\ln[(1-s)\lambda_d^s+s\lambda_d^{s-1}]$$

$$\geq\cdots\geq s(1-s)[\boldsymbol{v}_D^T(\mu_2-\mu_1)]^2[(1-s)\lambda_D+s]^{-1}+\ln[(1-s)\lambda_D^s+s\lambda_D^{s-1}]$$

从 $\sum_2^{-1}\sum_1$ 的本征向量系中，选前面 d 个本征向量作为 W，即 $W=[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_d]$。

方法二 这是一个更简单的方法。在假设类均值向量相等的情况下，按照式(8-49)选出前 $d-1$ 个坐标轴。然后再加上一个考虑类均值向量判别信息的坐标轴。用公式(8-43)的结果，把 $[(1-s)\sum_1+s\sum_2]$ 代入式中的 $\sum$ 可得：

$$\boldsymbol{w}_d=[(1-s)\sum_1+s\sum_2]^{-1}(\mu_2-\mu_1)$$

注意，这样产生的特征对准则值的作用不是可加性的。

8.4.3. 用散度准则函数的特征提取器

当只有两类时，这两类之间的散度可写为：

$$J_D(W)=\frac{1}{2}\mathrm{tr}W^TMW[(W^T\sum_1W)^{-1}+(W^T\sum_2W)^{-1}]$$
$$+\frac{1}{2}\mathrm{tr}[(W^T\sum_1W)^{-1}W^T\sum_2W+(W^T\sum_2W)^{-1}W^T\sum_1W-2I] \quad (8\text{-}50)$$

式中 $M=(\mu_2-\mu_1)(\mu_2-\mu_1)^T$。

和用 Chernoff 概率距离判据时一样，一般情况下得不到特征提取器的最优解析解。而在两类协方差阵相等的情况下，散度 J_D 和 Chernoff 距离 J_c 只差一个常数因子，因此这种情况下用 J_D 为判据的最优特征提取器和 8.4.2 小节“按概率距离判据的特征提取方法”中，在 $\sum_2=\sum_1=\sum$ 条件下用 J_c 作判据得到的 d 个坐标轴是一样的。

在 $\sum_1\neq\sum_2$，但类均值向量相等条件下，经过数学推导我们可以得到下式：

$$J_D(W)=\frac{1}{2}\mathrm{tr}(\Lambda+\Lambda^{-1}-2I)=\frac{1}{2}\sum_{i=1}^{d}(\lambda_i+\frac{1}{\lambda_i}-2) \quad (8\text{-}51)$$

式中 Λ 是矩阵 $\sum_2^{-1}\sum_1$ 的本征值矩阵。因此 W 中的各列是相应于下述排列次序

$$\lambda_1+\frac{1}{\lambda_1}\geq\lambda_2+\frac{1}{\lambda_2}\geq\cdots\geq\lambda_d+\frac{1}{\lambda_d}\geq\cdots\geq\lambda_D+\frac{1}{\lambda_D} \tag{8-52}$$

的 $\sum_2^{-1}\sum_1$ 的前面 d 个本征向量。

可以用同 8.4.2 小节类似的方法得到一般情况下特征提取器的次优解。

方法一先假设类均值向量相等，根据上面讨论的结果，求出矩阵 $\sum_2^{-1}\sum_1$ 的本征向量作为矩阵 W 的列向量，然后再考虑类均值向量不相等所带来的分类效果。由于式(8-50)在非奇异线性变换下的不变性，可令

$$W^T\sum_2 W=I$$

和

$$W^T\sum_1 W=\Lambda^{-1}$$

于是

$$J_D(W)=\frac{1}{2}\mathrm{tr}[W^TMW(\Lambda+I)+\Lambda+\Lambda^{-1}-2I]$$

$$=\frac{1}{2}\sum_{i=1}^{d}\left\{[\boldsymbol{w}_i^T(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)]^2(\lambda_i+1)+\lambda_i+\frac{1}{\lambda_i}-2\right\} \tag{8-53}$$

因此选取按下列顺序排列

$$[\boldsymbol{w}_1^T(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)]^2(\lambda_1+1)+\lambda_1+\frac{1}{\lambda_1}\geq[\boldsymbol{w}_2^T(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)]^2(\lambda_2+1)+\lambda_2+\frac{1}{\lambda_2}$$

$$\geq\cdots\geq[\boldsymbol{w}_d^T(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)]^2(\lambda_d+1)+\lambda_d+\frac{1}{\lambda_d}$$

$$\geq\cdots\geq[\boldsymbol{w}_D^T(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)]^2(\lambda_D+1)+\lambda_D+\frac{1}{\lambda_D} \tag{8-54}$$

的 $\sum_2^{-1}\sum_1$ 前 d 个的本征向量作为特征提取器。

方法二我们也可以把式(8-50)中的 $J_D(W)$ 的两个相加项分别最大化以求出 W。为了考虑包含在类平均向量中的分类信息，从 $J_D(W)$ 的第一项中可以看到，需要把 $(W^T\sum_1 W)^{-1}$ 和 $(W^T\sum_2 W)^{-1}$ 展开。由于 W 是 $D\times d$ 矩阵，因此在展式中要用广义逆

$$W^+=(W^TW)^{-1}W^T$$

来代替 W^{-1}，这样 $J_D(W)$ 可写成

$$J_D(W)=\frac{1}{2}\mathrm{tr}\{W^TMW(W^TW)^{-1}W^T[\Sigma_1^{-1}+\Sigma_2^{-1}]W(W^TW)^{-1}\}$$

$$= \frac{1}{2}\mathrm{tr}\{W^T M W[W^T(\sum_1^{-1}+\sum_2^{-1})W]^{-1}\} \tag{8-55}$$

它和 Chernoff 概率距离作为判据的表达式在形式上是一样的，因此 W 将是矩阵 $(\sum_1^{-1}+\sum_2^{-1})M$ 的本征向量系统，但是由于 M 的秩为 1，所以只用一个和矩阵 $(\sum_1^{-1}+\sum_2^{-1})M$ 的非零本征值对应的本征向量就能得到所要求的判别信息，即

$$\boldsymbol{w}_1 = (\sum_1^{-1}+\sum_2^{-1})(\mu_2-\mu_1) \tag{8-56}$$

至于其余的 $d-1$ 个向量可以从上述的在类均值向量相等的假设下，用公式(8-52)得到。

8.4.4. 多类情况

在多类情况下，最优特征提取器 W 应使广义的类别可分性判据

$$J(W) = \sum_{i=1}^{c}\sum_{j=1}^{c} J_{ij}(W) \tag{8-57}$$

最大。

要推导出最优特征变换的解析解是不大可能的。一个可行的办法是先求取一个候选坐标轴集合 $\{\boldsymbol{v}\}$。例如，它可以包括：

(1) $\boldsymbol{v} = (\sum_i^{-1}+\sum_j^{-1})(\mu_j-\mu_i), \forall i, \forall j$；

(2) $\sum_j^{-1}\sum_i$ 的所有本征向量 $\forall i, \forall j$；

(3) $\sum^{-1} M$ 的本征向量系统。

这里，i 和 j 是类别号，$\sum_i$，μ_i 是第 i 类的协方差阵和类均值向量。$\sum$ 是总的混合协方差矩阵，M 是类均值向量的离散度矩阵。

假设候选坐标轴的总数是 D。因此我们可以用搜索算法从这有 D 个坐标轴的集合中选出使 $J(W)$ 达最大的 d 个坐标轴来。

例 8.2 为说明前面结果，我们给出一个简单的例子，有两类样本，如图 8.6 所示。

ω_1 (用 ◦ 表示)： ω_2 (用 * 表示)：

$x_{11} = (0,0,0)^T$， $x_{21} = (0,0,1)^T$

$x_{12} = (1,0,0)^T$， $x_{22} = (0,1,0)^T$

$x_{13} = (1,0,1)^T$， $x_{23} = (0,1,1)^T$

$x_{14} = (1,1,0)^T$， $x_{24} = (1,1,1)^T$ 。

试利用散度判据降低维数。

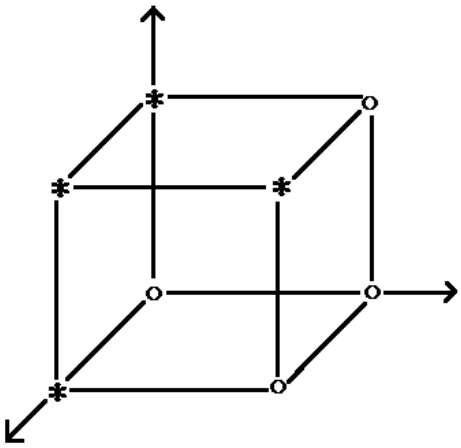


图 8.6 例 8.2 中的两类样本

解 我们用样本均值代替总体均值，样本协方差矩阵代替总体协方差矩阵。

$$\mu_1 \doteq \boldsymbol{m}_1 = \frac{1}{4}\begin{pmatrix}3\\1\\1\end{pmatrix}, \qquad \mu_2 \doteq \boldsymbol{m}_2 = \frac{1}{4}\begin{pmatrix}1\\3\\3\end{pmatrix}$$

$$\mu_1 - \mu_2 \doteq \boldsymbol{m}_1 - \boldsymbol{m}_2 = \frac{1}{4}\begin{pmatrix}2\\-2\\-2\end{pmatrix}$$

$$\Sigma_1 = \Sigma_2 = \Sigma = \frac{1}{16}\begin{bmatrix}3 & 3 & 1\\1 & 3 & -1\\1 & -1 & 3\end{bmatrix}$$

这是两类协方差阵相等的情况。

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

$$\Sigma^{-1}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T = \frac{1}{4}\begin{bmatrix}1 & -1 & -1\\-1 & 1 & 1\\-1 & 1 & 1\end{bmatrix}$$

此矩阵之非零本征值及其对应之本征向量为：

$$\lambda = \frac{3}{4}, \quad \boldsymbol{v} = \begin{pmatrix}-1\\1\\1\end{pmatrix}$$

因此 $\boldsymbol{w}^T = (-1,1,1)$

对所给样本集做变换： $\boldsymbol{x}^* = \boldsymbol{w}^T\boldsymbol{x}$ ，可得到下面一维结果(图 8.7)。

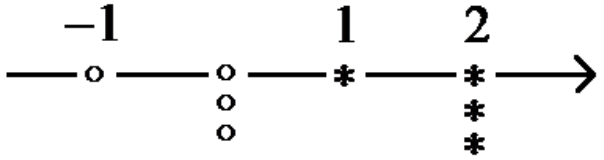


图 8.7 图 8.6 的样本投影到一维的情况

$$\omega_1 : x_{11}^* = 0 \qquad \omega_2 : x_{21}^* = 1$$

$$x_{12}^* = -1 \qquad x_{22}^* = 1$$

$$x_{13}^* = 0 \qquad x_{23}^* = 2$$

$$x_{14}^* = 0 \qquad x_{24}^* = 1$$

从图 8.7 可见，结果是完全可分的。就是说从向量 $\boldsymbol{x}$ 经 $\boldsymbol{w}$ 变换到 $\boldsymbol{x}^*$ 没有减少可分性。

8.4.5. 基于判别熵最小化的特征提取

我们在上一节中讨论了用熵作为不确定性的一种度量的表达式。例如， Shannon 熵

$$H = -\sum_{i=1}^{c} P(\omega_i \mid \boldsymbol{x}) \log P(\omega_i \mid \boldsymbol{x}) \tag{8-58}$$

这样一种概念也可以用来作为某个概率分布密度 $p(x_i)$ 偏离给定标准分布 $w(x_i)$ 的程度的度量，我们把它叫做相对熵，即

$$V(p,w) = -\sum p(x_i) \log[p(x_i)/w(x_i)] \le 0 \tag{8-59}$$

求和应在该特征所有可能的取值上进行。
相对熵越小，这两类概率分布的差别就越大，当两类概率分布完全相同时，相对熵达最大值(等于零)。因此我们可以定义判别熵 $W(p,q)$ 来表征两类分布 $p(x_i)$ 和 $q(x_i)$ 的差别大小。

$$\begin{aligned} W(p,q) &= V(p,q) + V(q,p) \\ &= -\sum p(x_i) \log p(x_i) - \sum q(x_i) \log q(x_i) \\ &\quad + \sum p(x_i) \log q(x_i) + \sum q(x_i) \log p(x_i) \le 0 \end{aligned} \tag{8-60}$$

在多类情况下，可以用

$$\sum_i \sum_j W(p^{(i)}, q^{(j)})$$

来表示各类分布之间的分离程度。这里i， j代表类别号。

对于特征提取来说，在给定维数d的条件下，我们应该求得这样d个特征，它使上述判别熵最小。

为了计算方便起见，我们可以用下列函数

$$U(p,q)=-\sum_{i}(p_i-q_i)^2\le 0 \tag{8-61}$$

来代替$W(p,q)$，而不影响选取d个最优特征的结果。

在不对概率分布作估计的情况下，可以用经过归一化处理的样本特征值来代替上式中的概率分布。

$$p_i^{(1)}=\frac{1}{N_1}\sum_{k=1}^{N_1}(x_{ki}^{(1)})^2 \tag{8-62}$$

且

$$\sum_{i=1}^{D}(x_{ki}^{(1)})^2=1$$

k是第一类样本集中的样本号， N_1是第一类的样本总数， i是特征号。由于

$$\sum_{i=1}^{D}p_i=1$$

所以这样做是合理的。

下面将证明使U取最小值的坐标系统是由矩阵

$$A=G^{(1)}-G^{(2)}$$

的满足一定条件的d个本征值相应的本征向量所组成的。

这里$G^{(1)}$和$G^{(2)}$分别是第一类样本集和第二类样本集的协方差矩阵。例如$G^{(1)}$的第i行第j列元素可用下式计算：

$$G_{ij}^{(1)}=\frac{1}{N_1}\sum_{n=1}^{N_1}x_{n_i}^{(1)}x_{n_j}^{(1)}$$

式中$x_{n_i}^{(1)}$表示第一类第n个样本的第i个坐标值(特征)。

令矩阵A的本征向量及对应的本征值依次为$\boldsymbol{u}_k$及$\lambda_k, k=1,2,\cdots,D$。由于$G^{(1)}$，$G^{(2)}$对称，所以$A$对称，且其迹为零，即

$$\mathrm{tr}A=0$$

用T表示从$\boldsymbol{u}$坐标到x坐标的变换矩阵，因为这个变换是一种旋转，所以

$$T^{-1}=T^{T} \tag{8-63}$$

式(8-61)中U在$\boldsymbol{u}$坐标系中为

$$U_0 = -\sum_{k=1}^{D} \lambda_k^2 = -\sum_{k=1}^{D} (A_0^2)_{kk}$$

式中 $(A_0^2)_{kk}$ 表示矩阵 (A_0^2) 中第 k 行 k 列的元素。而在原坐标系中则有

$$U = -\sum_{i=1}^{D} (A_{ii})^2 \geq -\sum_{i=1}^{D} (A_{ii})^2 - 2\sum_{i>k}\sum (A_{ik})^2$$

$$= -[\sum_{i=1}^{D} (A_{ii})^2 + 2\sum_{i>k}\sum (A_{ik})^2] \ ((A_{ij})\text{对称})$$

$$= -\sum_{i=1}^{D} (A^2)_{ii} \tag{8-64}$$

矩阵 $A = G^{(1)} - G^{(2)}$，即为 A_o 在 x 坐标系中的表达式。因此有

$$-\sum_{i=1}^{D} (A^2)_{ii} = -\sum_{i}\sum_{k} T_{ik} A_{0kk}^2 T_{ki}^T = -\sum_{i}\sum_{k} B_{ik}\lambda_k^2 \tag{8-65}$$

据式(8-63)有：

$$B_{ik} = T_{ik} T_{ki}^T = T_{ik} T_{ki}^{-1} = (T_{ik})^2$$

及

$$\sum_{i=1}^{D} B_{ik} = 1, \ \sum_{i=1}^{D} B_{ik} = 1, \ B_{ik} \geq 0$$

据式(8-64) 和式(8-65)可得

$$U \geq -\sum_{k=1}^{D} \lambda_k^2 = U_0$$

这证明了上述结论，将矩阵 A 的本征值 λ_k 排队：

$$\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_d^2 \geq \cdots\cdots \geq \lambda_D^2 \tag{8-66}$$

选前 d 个本征值对应的本征向量 $\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_d$ 为所要求的坐标轴系统，在这个坐标系统中判别熵最小。

8.5. 特征选择

我们在前面已经指出，用少数几个特征进行分类器设计，不仅在样本不多的条件下可

以改善分类器的准确率，而且在样本很多情况下，能够简化一些计算，以降低模式识别系统的代价。特征选择的任务是从一组数量为 D 的特征中选择出数量为 d ($D > d$)的一组最优特征来。为此有二个问题要解决，一是选择的标准，这可以用前面讲的可分离性判据，即要选出使某一可分性达到最大的一组特征来。另一问题是要找一个较好的算法，以便在允许的时间内找出最优的那一组特征。

如果考虑把 D 个特征中的每一个单独使用时的可分性判据都算出来，按判据值大小排队，例如：

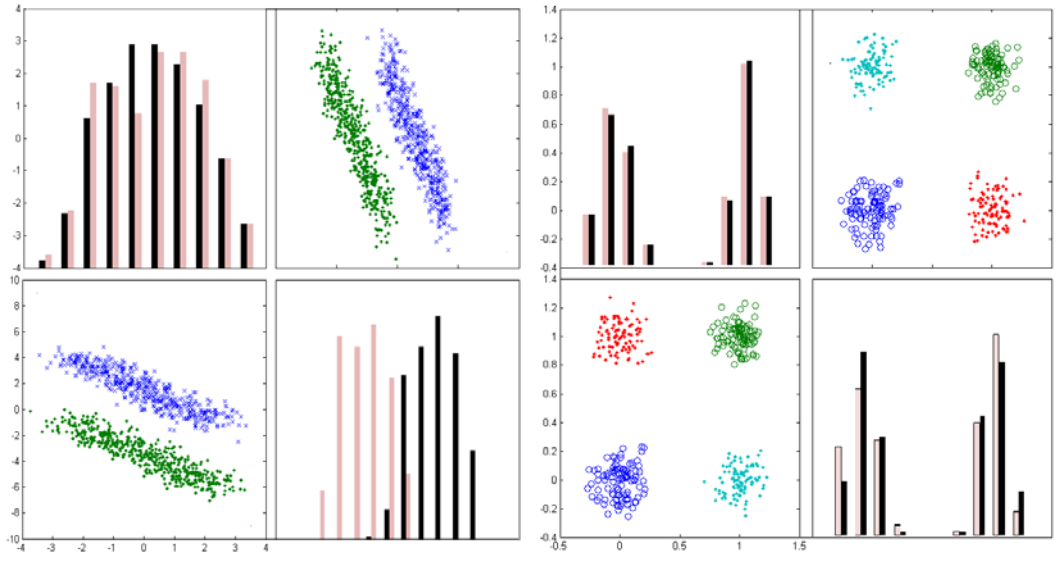
$$J(x_1) > J(x_2) > ... > J(x_D)$$

就可以提个问题：单独使用使 J 较大的前 d 个特征是否就是最优的一组特征呢?如果回答是肯定的，特征选择也就变得简单了。不幸的是，即使当所有特征都相互独立时，除了一些特殊的情况外，一般来说，前 d 个最有效的特征并非是最优的(数量为 d 的)一组特征，甚至有可能是最不好的一组特征(见习题 8. 12)。

我们可以通过图图 8. 8 来看这个问题的复杂性。图 8.8（a）的左下图给出了两个分布。这两个分布分别向横轴和纵轴投影，得到左上图和右下图。从右下图可以看出，纵轴对应的特征包含一些分类信息，两个分布有一些可分性。而横轴上两个分布几乎是重叠的，看起来，横轴特征是“无用”的。但是，这两个特征合在一起时，两类完全可分。图 8.8（b）给出的例子中，横轴和纵轴特征看起来都完全“无用”，但是，这两个特征合在一起时，两类完全可分。

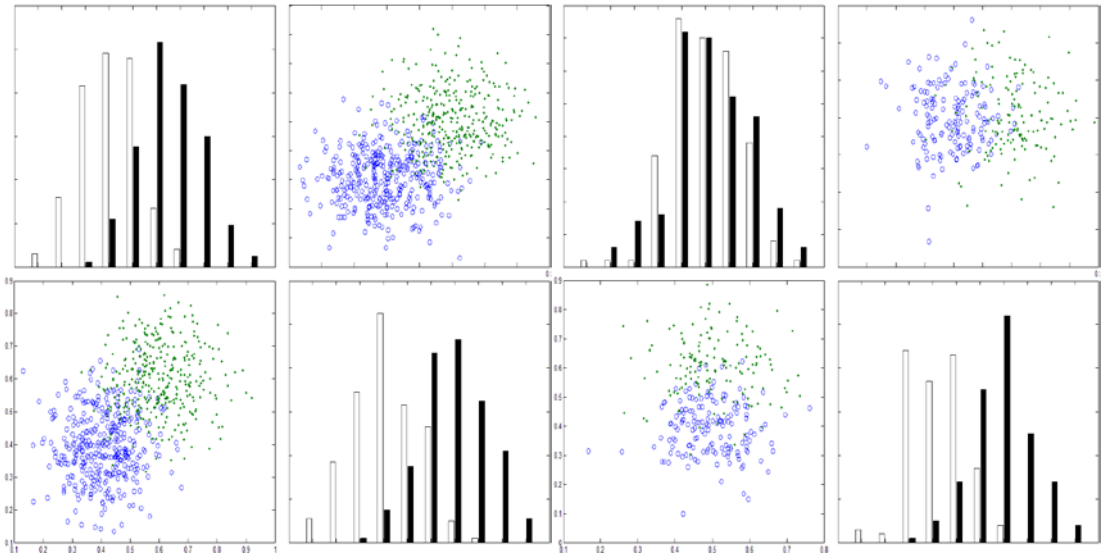
我们再看图图 8. 9 中的情况。图 8.9（a）中的两个特征分布完全相同，看起来这两个特征合起来对分类不会有更多的帮助，但实际上，图 8.9（b）告诉我们两个特征合起来后，可分性变大了。图图 8. 10 告诉我们，当两个特征完全相关的特征在包含信息上可能是冗余的，而两个不相关的特征所包含的可分性信息可能并不是冗余的。

因此，对于一组特征来说，我们仅仅从部分特征的可分性难以判断出整体的可分性，这就给特征选择造成了困难。



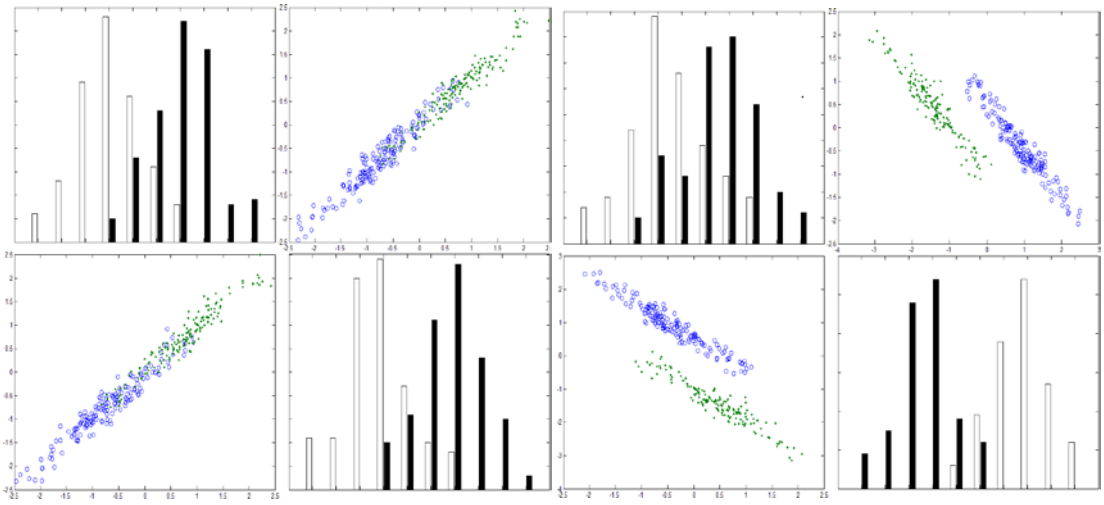
（a）　　（b）

图 8. 8　对于一组特征来说，从部分特征的可分性难以判断出整体的可分性



（a）　　（b）

图 8.9 两个相同分布的变量，其包含的信息并不冗余



（a）　　（b）

图 8.10 完全相关的变量在包含信息上是冗余的

当然，我们可以给出一种很简单的方法，即把各种可能的特征组合的 J 都算出来再加以比较，以选择最优的一组特征(这种方法叫穷举法)。这种方法计算量太大。这是因为，从 D 个特征中挑选 d 个，所有可能的组合数为：

$$q = C_D^d = \frac{D!}{(D-d)!d!}$$

如果 D=100， d =10，则 q 的数量级是 10^{13}。我们知道，这样的计算量是很大的。微积分理论告诉我们，q 的计算量就是指数级的运算量。而计算复杂性理论告诉我们，指数级运算量随变量的增长而迅速增长，是任何一台计算机所不能承受的，并且通过提高计算机的速度无助于这个问题的解决。因此，寻找一种可行的算法变得非常必要。

应当说明的是，任何非穷举的算法都不能保证所得结果是最优的，除非我们能够发现特征选择这个问题的内部结构以利于最优化搜索（至少目前我们还没有做到这一点）。因此，除非只要求次优解，否则所用算法原则上仍是穷举算法，只不过采取某些搜索技术使计算量

可能有所降低。分支定界法就是这样的一种穷举性质的搜索方法。该方法把最优特征组合的寻找转化成一个搜索过程，由于合理地组织搜索过程，使得有可能避免计算某些特征组合而不影响结果为最优。由于这一算法非常复杂，在实际中很少被使用，我们再此不对其做介绍。有兴趣读者可以阅读相关文献。

实际中我们用得更多的是次优搜索法。次优搜索法通常简单，速度快，给出的解也可以在一定程度上满足需要，因此常常被使用。下面我们介绍几种次优搜索法。

8.5.1. 次优搜索法

下面我们介绍几个算法，它们一个比一个复杂，当然得到的特征组合性能也越来越高。这里的性能只能从概率意义上理解。

1．单独最优特征组合

最简单的方法是计算各特征单独使用时的判据值并加以排队，取前 d 个作为选择结果。正如前面指出过的，即使各特征是统计独立的，这一结果也不一定就是最优结果，只有当可分性判据 J 可写为如下形式时：

$$J(X)=\sum_{i=1}^{D}J(x_i)$$

或　　$$J(X)=\prod_{i=1}^{D}J(x_i)$$

这种方法才能选出一组最优的特征来。其中一种可能是在两类都是正态分布情况下，当各特征统计独立时，用 Mahalanobis 距离作为可分性判据，就满足上述要求。

下面我们给出一个单独最优特征组合的算法Relief。该算法假设有一样本集合X 包含N 个D 维样本x。

算法（Relief）
bigin
初始化d 维权向量w = [w_1,w_2, …,w_D] 的各个权值都为0 .
for i = 1 : n
从X 中随机选择一个样本x
计算X 中离x 最近的同类样本h ,和不同类的样本m
for j = 1 : d
$w_j = w_j - diff(j;\ x;\ h)/n + diff(j;\ x;m)/n$
return　w
end
其中 $diff(j;\ x;\ h)$表示 x 和 h 这两个样本在 j 维上的差异。当该特征取离散值时，可以如下计算：

$$diff(j,x,h)=\begin{cases}0 & x_j=h_j\\ 1 & otherwise\end{cases}$$

当该特征取离散值时，可以如下计算：

$$diff(j,x,h)=\frac{|x_j-h_j|}{x_{j\max}-x_{j\min}}$$

其中 $x_{j\max}$， $x_{j\min}$ 分别表示样本在第 j 个特征上的最大值和最小值。

算法结束后选择权重最大的前 d 个特征作为最后的特征组合。根据分析可以知道，该算法采取了局部可分性的准则（for 循环语句内容），即：根据每一个样本的近邻样本（h 和 m）而不是所有样本，计算其可分性。这样做的好处之一是因为有些特征从整体上看（如：均值，方差）可分性不够好，但是局部特性反映了该特征具有可分性。当然，为了提高算法对于随机因素的敏感，可以选择样本 x 的 k 个同类近邻和不同类近邻，从而改进该算法。另外，当我们面对的是多类分类问题时，也可以修改该算法以适应多类分类问题，我们把这个算法扩展工作留给读者作为习题。

2．顺序前进法(Sequential Forward Selection，SFS)

这是最简单的自下而上搜索方法，每次从未入选的特征中选择一个特征，使得它与已入选的特征组合在一起时所得 J 值为最大，直到特征数增加到 d 为止。

设已选入了 k 个特征构成了一个大小为 k 的特征组 X_k，把未入选的 $D-k$ 个特征 x_j， $j=1,2,\cdots,D-k$，按与已入选特征组合后的 J 值大小排列：

即若

$$J(X_k+x_1)\geq J(X_k+x_2)\geq\cdots\geq J(X_k+x_{D-k})$$

则下一步的特征组选为 $X_{k+1}=X_k+x_1$。

开始 $X_0=\varnothing$ 时，直到 $k=d$ 为止。

SFS 法考虑了所选特征与已入选特征之间的相关性，一般说来比上面讲的按单独使用时 J 值最大的选择方法好些，主要缺点是一旦某特征已入选，即使由于后加入的特征使它变为多余，也无法再把它剔除。

把 SFS 法推广为每次不止增加一个特征而是增加 r 个特征，就成为广义顺序前进法(Generalized Sequential Forward Selection，GSFS)。即每次从未入选的特征中选择出 r 个特征，使得这 r 个特征加入后 J 值达最大。

SFS 法每次只增加一个特征，它未考虑未入选特征之间的统计相关性，而 GSFS 法可以克服这个缺点，当然此时每步有 C_{D-k}^{r} 个候补特征组需要逐个计算，因而计算量变大了，相应地，它比 SFS 法更可靠，此外它也无法剔除已入选的特征。

3. 顺序后退法(Sequential Backward Selection，SBS)

这是一种自上而下的方法，从全体特征开始每次剔除一个，所剔除的特征应使剩下的特征组的 J 值最大。例如，设已剔除了 k 个特征，剩下的特征组为 $\overline{X}_k$，将 $\overline{X}_k$ 中的各特征 x_j 按

下述 J 值大小排队， $j=1,2,\cdots,D-k$ 。

若 $J(\overline{X}_k - x_1) \geq J(\overline{X}_k - x_2) \geq \cdots \geq J(\overline{X}_k - x_{D-k})$

则 $\overline{X}_{k+1} = \overline{X}_k - x_1$

和顺序前进法比较，顺序后退法有两个特点：一是在计算过程中可以估计每去掉一个特征所造成可分性的降低，二是由于顺序后退法的计算是在高维空间进行的，所以计算量比顺序前进法要大。同样此法亦可推广为广义顺序后退法(Generalized Sequential Backward Selection，GSBS)。

4 增 l 减 r 法($l-r$ 法)

为避免前面方法的一旦被选入(或剔除)就不能再剔除(或选入)的缺点，可在选择过程中加入局部回溯过程。例如，在第 k 步可先用 SFS 法一个个加入特征到 $k+l$ 个，然后再用 SBS 法一个个剔去 r 个特征，我们把这样一种算法叫增 l 减 r 法($l-r$ 法)。具体步骤如下(假设已经选了 k 个特征，得出了特征组 X_k)：

步骤 1 用 SFS 法在未入选特征组 $\overline{X}_D - X_k$ 中逐个选入特征 l 个，形成新特征组 X_{k+l} ，置 $k=k+l$ ， $X_k = X_{k+l}$ 。

步骤 2 用 SBS 法从 X_k 中逐个剔除 r 个最差的特征，形成新特征组 X_{k-r} ，置 $k=k-r$ 。若 $k=d$ 则终止算法，否则，置 $X_k = X_{k-r}$ ，转向第一步。

这里要说明一下，当 $l>r$ 时， $l-r$ 法是自下而上的算法，先执行第一步，然后执行第二步，起始时应置 $k=0$ ， $X_0 = \varnothing$ 。当 $l<r$ 时， $l-r$ 法是自上而下的算法，先执行第二步，然后执行第一步，起始时应置 $k=D$ ， $X_0 = \{x_1, x_2, \cdots, x_D\}$ 。

我们会自然地想到，若用 GSFS 及 GSBS 代替上法中的 SFS 和 SBS，则可形成一个广义的 $l-r$ 法，这里我们再展开一下，令 l ， r 由一些整数

$$l_i, i=1,2,\cdots,z_l$$

$$r_j, j=1,2,\cdots,z_r$$

组成，且

$$0 \leq l_i \leq l, \quad 0 \leq r_j \leq r$$

$$\sum_{i=1}^{z_l} l_i = l, \quad \sum_{j=1}^{z_r} r_j = r$$

然后用 GSFS 法逐次分 z_l 步增加特征，每次加入 l_i 个特征，一共加入 l 个特征；再用 GSBS 法分 z_l 步剔除特征，每次剔除 r_j 个特征，一共剔除 r 个特征。具体说就是：

$$Z_l = (l_1, l_2, \cdots, l_{z_l})$$

$$Z_r = (l_1, l_2, \cdots, l_{z_r})$$

称此法为(Z_l，Z_r)法。之所以要分 z_l 步增加 l 个特征和分 z_r 步减少 r 个特征，是为了既要考虑到入选(或剔除)特征之间的相关性，又不至于因此引起计算量过大。例如，假使一步就使特征组从原来的空集增加到 d 个特征，就成为计算量很大的穷举法，而每次只增加一个特征，又没有考虑入选特征间的相关性，因此合理地设置 Z_l 和 Z_r 可以同时对这两者(计算复杂性和特征选择的合理性)兼顾考虑。

前面所讲的各种方法都可看作为(Z_l，Z_r)法的特例，它们之间的关系，如下表所示。

（Z_l，Z_r）算法		*等效算法*
$Z_l = (1)$	$Z_r = (0)$	SFS (1,0) 算法
$Z_l = (0)$	$Z_r = (1)$	SBS（0,1）算法
$Z_l = (d)$	$Z_r = (0)$	穷举法
$Z_l = (l)$	$Z_r = (0)$	GSPS（l）算法
$Z_l = (0)$	$Z_r = (r)$	GSBS（r）算法
$Z_l = (1,1,\cdots,1)$	$Z_r = (1,1,\cdots,1)$	（l，r）算法

8.5.2. 可分性判据的递推计算

所有上述搜索算法都有一个共同点，即第 k 步特征组是在第 $k-1$ 步特征组上加入或剔除某些特征来构成的。因此我们可以分析一下，是否有可能从第 $k-1$ 步的判据值 $J(x_{k-1})$ 推算出 $J(x_k)$ 而不必完全重新计算。仔细观察式(8-22)～式(8-25)可以知道，各判据 J 都是协方差矩阵与均值向量的函数，若加上(或减去)某一特征，均值向量只增加(或减少)一个元素，协方差矩阵增加(或减少)一行一列。可见对于这些判据递推关系是存在的，即求 $J(x_k)$ 时可

在 $J(x_{k-1})$ 基础上把新加入(或剔除)特征的影响加进去即可，不必完全从头算起，这样就大大简化了计算工作。这里不再讨论具体的计算公式，需要时可参阅有关文献[2]。

8.5.3. 基于迹比值的特征选择方法

对于特征选择，我们可以根据给定的数据构造两个无向图 $\mathcal{G}_w$ 和 $\mathcal{G}_b$，其中 $\mathcal{G}_w$ 反映类内关系，图 $\mathcal{G}_b$ 反映类间关系。图 $\mathcal{G}_w$ 和图 $\mathcal{G}_b$ 分别由邻接矩阵 A_w 和 A_b 来刻画。一般来说，为了反映数据中类内关系，当数据 x_i 和 x_j 属于同一类或互为近邻时，权值 $(A_w)_{ij}$（A_w 矩阵的第 ij 个元素）取一个相对大的值，否则就取一个相对小的值。我们希望选择一组特征后，$\sum_{ij}\|\boldsymbol{y}_i-\boldsymbol{y}_j\|^2(A_w)_{ij}$ 越小越好。

类似地，为了反映数据中类间关系，当数据 x_i 和 x_j 属于不同类或距离比较远时，权值 $(A_b)_{ij}$ 取一个相对大的值，否则就取一个相对小的值。我们希望选择一组特征后，$\sum_{ij}\|\boldsymbol{y}_i-\boldsymbol{y}_j\|^2(A_b)_{ij}$ 越大越好。

因此，我们考虑采取如下的目标函数：

$$J(W_I)=\frac{\sum_{ij}\|\boldsymbol{y}_i-\boldsymbol{y}_j\|^2(A_b)_{ij}}{\sum_{ij}\|\boldsymbol{y}_i-\boldsymbol{y}_j\|^2(A_w)_{ij}} \tag{8-67}$$

式(8-67)可以写成

$$J(W_I)=\frac{\mathrm{tr}(W_I^T XL_bX^TW_I)}{\mathrm{tr}(W_I^T XL_wX^TW_I)} \tag{8-68}$$

其中 L_w 和 L_b 为拉普拉斯矩阵。$L_w=D_w-A_w$，D_w 为对角矩阵，第 i 个对角线元素为 $(D_w)_{ii}=\sum_j(A_w)_{ij}$。$L_b=D_b-A_b$，$D_b$ 为对角矩阵，第 i 个对角线元素为 $(D_b)_{ii}=\sum_j(A_b)_{ij}$。因此，只需求解如下的基于迹比值目标函数的优化问题：

$$\Phi(I)=\arg\max_{\Phi(I)}\frac{\mathrm{tr}(W_I^T XL_bX^TW_I)}{\mathrm{tr}(W_I^T XL_wX^TW_I)} \tag{8-69}$$

---

[2] C. B. Chittineni:“Efficient feature subset Selection with probabilistic distance Criterion”, Information Science, Vol. 22,No.1.Oct.1980 pp.19～35.

显然 XL_bX^T 和 XL_wX^T 都是半正定的，已有优化算法可以快速的求解这个优化问题。

值得指出的是，式(8-67)为我们提供了一个特征选择的一般化框架。不同的方法构造的 A_w 和 A_b 将导致不同的非监督，半监督以及监督的特征选择算法。Fisher 分数和拉普拉斯分数是其中的两种具有代表性的例子。

Fisher 分数中， A_w 和 A_b 按如下来构造：

$$(A_w)_{ij}=\begin{cases}\dfrac{1}{n_{c(i)}}, & \text{如果} c(i)=c(j)\\ 0, & \text{如果} c(i)\neq c(j)\end{cases} \tag{8-70}$$

$$(A_b)_{ij}=\begin{cases}\dfrac{1}{n}-\dfrac{1}{n_{c(i)}}, & \text{如果} c(i)=c(j)\\ \dfrac{1}{n}, & \text{如果} c(i)\neq c(j)\end{cases} \tag{8-71}$$

其中 $c(i)$ 表示 x_i 的类别标签， n_i 表示第 i 类中数据的个数。

在拉普拉斯分数中， A_w 和 A_b 按如下来构造：

$$(A_w)_{ij}=\begin{cases}e^{-\frac{\|x_i-x_j\|^2}{t}}, & x_i\text{和}x_j\text{为近邻；}\\ 0, & \text{otherwise}\end{cases} \tag{8-72}$$

$$\mathrm{A_b}=\frac{1}{1^TD_w1}D_w11^TD_w \tag{8-73}$$

其中 1 表示各个元素都取值为 1 的列向量。Fisher 分数是一种监督的方法，利用了类别信息来构造 A_w 和 A_b ，而拉普拉斯分数是一种非监督的方法，在构造 A_w 和 A_b 时并没有利用类别信息。基于准则(8-67)，特征子集 $\Phi(I)$ 的分数可以定义为：

$$score(\Phi(I))=\frac{\mathrm{tr}(W_I^TXL_bX^TW_I)}{\mathrm{tr}(W_I^TXL_wX^TW_I)} \tag{8-74}$$

在传统的方法中，由于直接找一个子集分数最大的特征子集比较困难，一般是采用另一种更简单的方法来求解。首先对每一个特征按如下公式进行打分：

$$score(F_i)=\frac{W_I^TXL_bX^TW_I}{W_I^TXL_wX^TW_I} \tag{8-75}$$

然后按分数排序，找到前 m 个分数最大的特征。显然这种方法找到的特征子集 $\Phi(I)$ 不能保证优化问题(8-69)达到全局最优。