模式识别与图像处理

戚飞虎等　译

上海

模式识别与图像处理

戚飞虎　周源华
余松煜　郑志航　等译

上海交通大学出版社

序　言

在过去20年中，对模式识别和图像处理问题的兴趣有着很大的增长。这种兴趣引起了对用于设计模式识别和图像处理系统的理论方法，实验软件和硬件的日益增长的需要。在模式识别和图像处理领域中出版的书籍已超过25种。另外又出版了大量编著、会议论文集和杂志的专刊。在这个领域里有四种专门的刊物：(1) IEEE *Transaction on Pattern Analgsis and Machine Intelligence*, (2) *Pattern Recognition*, (3) *Pattern Recognition Letters*, 和(4) *Computer Vision, Graphics, and Image Processing*, 已经设计和建成了一些专用的模式识别机器和图像处理系统，供实际应用。模式识别和图像处理的应用包括：字符识别，目标检测，医学诊断，生物医学信号和图像的分析，遥感，面貌和指纹的鉴别，可靠性分析，社会经济学，考古学，语言识别和理解，机器零件的识别和自动检验等。在这部篇幅有限的著作中，我们仅提供模式识别和图像处理主要组成部分的一个概貌。

模式识别主要涉及由物质的和精神的过程所得到的度量的描述和分析。为了提供有力而有效的模式描述，通常需要用预处理来消除噪声和多余信息。然后提取一组数值的和/或非数值的特征度量以及这些度量之间的关系来表示模式。在这个表示的基础上完成对于特定目标的模式分析(分类和/或描述)。

为了确定一组好的特征度量以及它们之间的关系用于模式表示，以期有好的识别性能，必须对所研究的模式进行仔细的分析。应该全面利用有关模式的统计特征和结构特征的知识。从这个观点来看，模式识别的研究应包括模式特征的分析和识别系统的设计。

有许多数学方法被用来解决模式识别问题。这些方法基本可分为两大类：(1) 决策论或统计方法；(2) 句法或结构方法。从更一般的观点来看，在一个通常的范围内这些方法可以用模式表示和决策形成以及结构分析来讨论(在给定的模式表示的基础上)。

浏览最近出版的模式识别和图像处理著作，可以发现这些著作主要涉及图像和景物的分析。图像处理的一般目的是分析给定景物的图像和识别景物的内容。许多种景物本质上是二维的（文件就是一个明显的例子）；并且在下列应用中二维处理通常是十分恰当的，例如遥感(从很高的高空看到平坦的地面)，放射学(图像是物体的"阴影")或者显微镜学(图像是物体一个剖面)。在这些情况下，图像分析过程基本上是二维的。我们从图像中提取诸如边征，或者将图像分割成区域，于是得到一种由标有特性量的图像特征所组成的类似地示。然后利用聚类过程可改善这些地图，我们可以用抽象的关系结构来表示这些些抽象的关系结构中，例如用标有不同特性量(颜色、纹理、形状等)的示区域之间的关系。最后，这些结构可与已存的模型进行匹配，这些与一般的图像类型相应)的广义关系结构。成功的匹配可以识别图的名称对图像结构进行描述。

在另外一些情况，特别在机器人视觉

考虑待分析景物的三维特性。这里，分析中的关键一着是推断每一个图像点上表面的方向。表面方向的有关线索能够直接从图像中的阴影(即灰度级的变化)得到。或者，先对图像进行二维分割和特征提取，提取诸如表面轮廓和纹理基元等特征，然后可以由轮廓形状或纹理变化导出表面方向的线索。利用被称为"二维半示意图"的表面方向图，可以再次使用特征提取和分割技术，分割出物体的可见部分或目标，然后可用关系结构表示它们。最后这个结构可与各种模型进行匹配，以便用已知物体的名称对景物作解释。应该指出，由于只能看到物体一个方向的图像，又由于物体相互间可能部分遮挡，因此在三维情况下匹配过程更为困难。我们不是简单地将一个模型与一个观察到的结构匹配，而是去证实在适当的观察条件下这个模型可能产生这种结构。

本书由四部分组成。第一部分专门介绍模式识别中的主要技术。第二部分概述了图像处理和理解的新发展。第三部分介绍几种用于模式识别和图像处理的计算机系统和结构。第四部分的一些章节展示模式识别和图像处理的主要应用。所有这些内容对在校的大学生和工程师们都是有用的。它将不但作为在模式识别和图像处理领域里广泛而综合的信息源，也将作为一部技术参考书。

目　　录

第1章 统计模式识别

§1.1 引 言

统计模式分类的目的在于确定已知样本所属的类别。通过观察或测量，可以得到一组数，这组数构成了度量向量。这是一种随机向量，它的条件密度函数取决于它的类别。

分类器的设计由两部分组成。一是从各个类别中采集数据样本，并找出区分各类别的边界，这一过程称为分类器的设计、训练或学习；第二是利用已知类别的样本来检验经过训练的分类器。

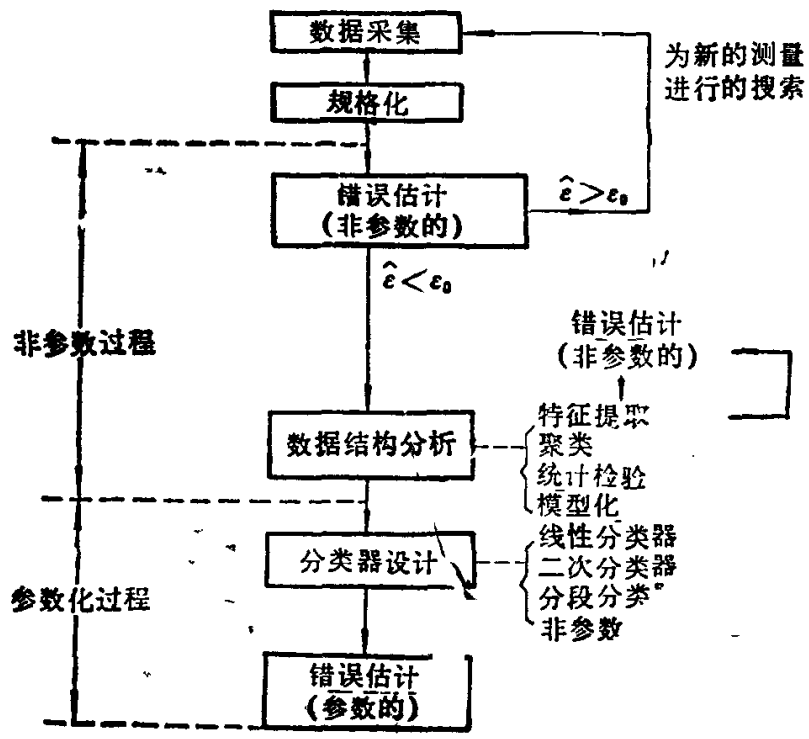


图 1.1 分类器的设计过程

图 1.1 示出了分类器的设计过程。首先，采集数据并作适当标准化处理，然后估计贝叶斯误差，即不同类别密度之间的交迭。贝叶斯误差是现行的测量空间中的最小可能误差。在后面的步骤中，特征选择和分类器设计常使误差增加。因此，如果在这一阶段中误差大得无法接受，则无法作进一步的数据处理，必须重新进行数据采集并寻找其他的度量以便使得误差可以被接受。估计的贝叶斯误差可以作为进一步运算的参考。例如，在作特征提取时，必须估计特征空间中的贝叶斯误差，并与原始度量空间中的贝叶斯误差相比较，以确定已提取的特征是否可以接受。对最终的分类器误差也是如此。

一旦估计的贝叶斯误差可以被接受，则可进行数据结构的分析，它包括诸如将特征提取、聚类、统计检验、建模等多种操作。在此基础上，我们可选择一个合适的分类器对给定的数据进行处理。最后一步是对这一分类器作出评价。

在测量空间或特征空间中对贝叶斯误差进行估计时，假设类密度函数具有任一数学形式(例如高斯函数)都是不适当的，因此必须使用非参数技术。而数据结构的分析和分类器设计相结合则被认为是一种参数化过程。

本章只讨论图 1.1 中部分内容。聚类和特征提取将在其他章节中详细讨论。而且，除另作说明外，本章只讨论两类问题，但在有些地方也列出更为一般的多类问题的结论。

§1.2 贝叶斯分类

在这一节中，将给出假设已知 $p_i(x)$ 和 P_i 时的分类算法和它产生的误差。这一分类算法亦称假设检验。

一、似然比分类器[1,2]

根据 $q_1(X)>q_2(X)$ 还是 $q_1(X)<q_2(X)$ 将 X 分为 ω_1 类或 ω_2 类，可以使错分的概率达到最小。即

$$q_1(X)\underset{\omega_2}{\overset{\omega_1}{\gtrless}}q_2(X)。 \qquad \text{(贝叶斯分类器)} \tag{1}$$

对 X 来说，风险为

$$r^*(X)=\min[q_1(X),q_2(X)]。 \quad \text{(贝叶斯风险)} \tag{2}$$

取期望即得总的误差率

$$\varepsilon^*=\mathrm{E}\{r^*(X)\}=P_1\int_{\Gamma_2}p_1(X)\mathrm{d}X+P_2\int_{\Gamma_1}p_2(X)\mathrm{d}X,$$

$$\text{(贝叶斯误差率)} \tag{3}$$

$)\mathrm{d}X,\varepsilon_2=\int_{\Gamma_1}p_2(X)\mathrm{d}X$，分别称为 ω_1 误差和 ω_2 误差，Γ_i 是 X 被分为 ω_i

$$X)\neq\max_i q_i(X)\rightarrow X\in\omega_k, \tag{4}$$

$$\{1-\max_i q_i(\boldsymbol{X})\}。 \tag{5}$$

$p_i(X)/p(X)$，采取负对数，则可得贝叶斯分类器的比较简

便的形式，即

$$h(X) = -\ln[p_1(X)/p_2(X)] \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} \ln[P_1/P_2]。 \tag{6}$$

$h(X)$ 加上一个阈值的分类器称为似然比分类器。

当 $\boldsymbol{X}$ 是高斯分布，且对 ω_i 有 M_i 和 $\sum_i$，则

$$-\ln p_i(X) = \frac{1}{2}(X - M_i)^T \sum_i^{-1}(X - M_i) + \frac{1}{2}\ln|\sum_i| + (n/2)\ln 2\pi。 \tag{7}$$

可以按照如下的不同要求改变分类器的阈值。

1．最小代价的贝叶斯分类器[2,3]

设 c_{ij} 是将 ω_i 的样品分为 ω_j 的代价，则把 X 分为 ω_i 的期望代价为

$$c_i(X) = \sum_{j=1}^{L} c_{ji} q_j(X)。 \tag{8}$$

分类规则和造成的代价为

$$c_k(X) = \min_i c_i(X) \rightarrow X \in \omega_k, \tag{9}$$

$$c^* = \mathrm{E}\{\min_i c_i(\boldsymbol{X})\}。 \tag{10}$$

对两类问题，有

$$h(X) = -\ln[p_1(X)/p_2(X)] \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} \ln[(c_{12} - c_{11})P_1/(c_{21} - c_{22})P_2], \tag{11}$$

这是使用新的阈值的似然比分类器。

2．Neyman-Pearson 检验[1]

设 ε_1、ε_2 分别是 ω_1 和 ω_2 的出错概率，如式(3)所示。在 ε_2 等于某一常数 ε_0 的情况下，似然比分类器使 ε_1 达最小。阈值必须适当选择以满足 $\varepsilon_2 = \varepsilon_0$，通常由经验而定。图 1.2 是似然比分类器在取不同阈值时 ε_1 对 $1 - \varepsilon_2$ 的曲线。它称之为运行特性，经常用来直观地观察怎样通过改变阈值对两类误差进行折衷。在 Neyman-Pearson 检验中，图中曲线上的黑点是运行点，其相应的阈值被选择。当 ω_2 表示相对于另一类(ω_1) 来说为待识别类时，ε_1、ε_2 和 $1 - \varepsilon_2$ 分别称为伪警、漏警和检出概率。

3．最小最大检验[1]

选择似然比分类器的阈值使下式成立则可以使期望代价在 P_i 改变时保持不变。

$$(c_{11} - c_{22}) + (c_{12} - c_{11})\varepsilon_1 - (c_{21} - c_{22})\varepsilon_2 = 0,$$

特别当 $c_{11} = c_{22}$，$c_{12} - c_{11} = c_{21} - c_{22}$ 时，选择阈值使 $\varepsilon_1 = \varepsilon_2$，这种分类器可以消除因 P 变化而产生很大误差的可能性。

以上三种情况同样使用似然比分类器，仅仅阈值不同。这可以解释为用人为的 P_i 代替(6)中真实的 P_i。因此，所有这些情况从理论上可以认为是贝叶斯分类器，而 P_i 有不同的含义。下面讨论几个有关的问题。

(1) 独立的度量集　当 $\boldsymbol{X}$ 是由统计独立的度量集 $x_1, x_2, \cdots, x_N$ 组成时，贝叶斯分类器变为

$$-\ln[p_1(X)/p_2(X)] = \sum_{i=1}^{N} -\ln[p_{i1}/p_{i2}]$$

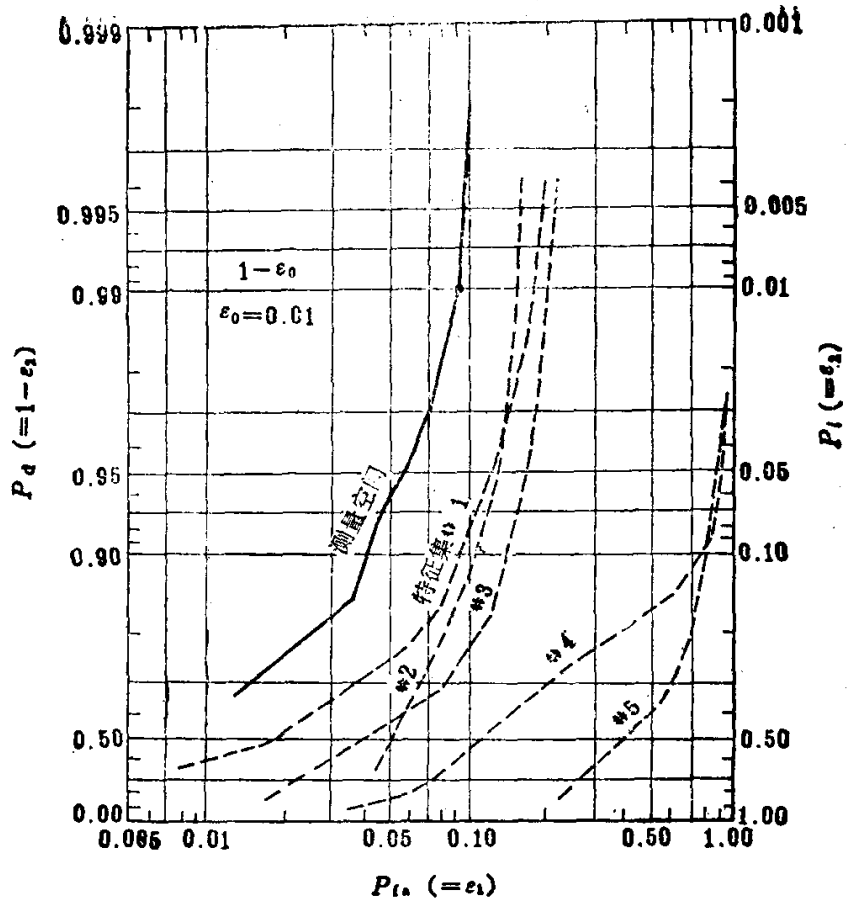


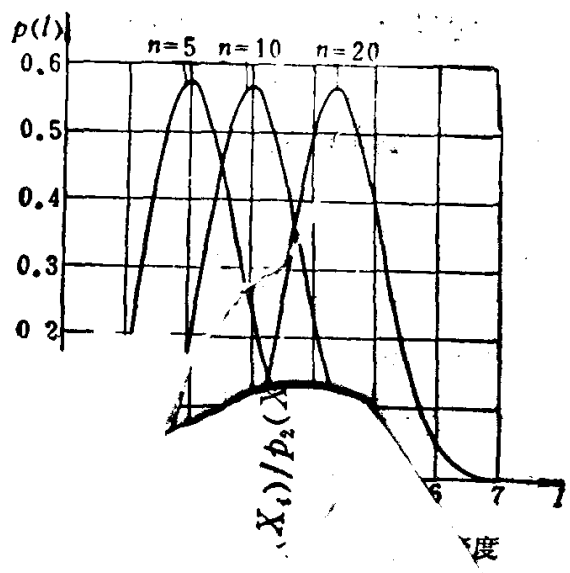
图 1.2　运行特性

对分类来说，这种情况表明了怎样将似乎无关的信息，例如雷达和红外信号结合在一起。

（2）一类分类器　当一个已被明确定义的类从其他可能的类别（有时没有被明确定义）中分出时，边界也许仅由一类的知识来确定。一个典型的例子是 $M=0, \sum=I$ 时的高斯分布的超球边界。但是，这一概念的采用必须十分小心，尤其在高维情况下。让我们考虑 $M=0, \sum=I$ 的高斯分布的半径 l 的边界密度：

$$p_l(l)=\frac{n}{2^{n/2}\Gamma(1+n/2)}\,l^{n-1}e^{-l^2/2}, \tag{14}$$

式中 $\Gamma(\cdot)$ 为gamma 函数。图1.3 示出了在 $n=5、10、20$ 时这一密度函数的曲线。这些密度



s 取值为 $0 \leqslant s \leqslant 1$，对 s 的任意值，Chernoff 上界都大于贝叶斯误差。但它与其他界限之间的不等式关系不存在。Bhattacharyya 上界是Chernoff 上界在 $s=0.5$ 时的特例。对 s 作最优化，可使 Chernoff 上界比 Bhattacharyya 上界小。但是，实际上在大多数情况下它们的差别不大，且通常取 s 为 0.5。

§1.3 贝叶斯误差的估计

本节介绍贝叶斯误差的非参数估计法。它们在 kNN 法和 Parzen 法中用来估计密度函数。

一、密度估计

下面介绍两种密度估计方法。

1. Parzen 法

设包含 X 的局部确定区域 $\Gamma(X)$ 的体积为 v，计得 $\Gamma(X)$ 中的样本数为 $\boldsymbol{k}(X)$，则密度函数用下式估计

$$\hat{\boldsymbol{p}}(X)=\frac{\boldsymbol{k}(X)}{Nv} \quad \text{且} \quad Pr\{\boldsymbol{k}=k\}=\begin{bmatrix} N \\ k \end{bmatrix} u^k(1-u)^{N-k} \tag{40}$$

式中 $u=\int_{\Gamma(X)} p(X)\mathrm{d}X$ 是某一样本落在 $\Gamma(X)$ 中的概率。如果 $\lim_{N\to\infty}k=\infty$ 且 $\lim_{N\to\infty}k/N=0$，则上式中 $\hat{\boldsymbol{p}}(X)$ 是渐近无偏和一致的[$\lim_{N\to\infty}\mathrm{Var}\{\hat{\boldsymbol{p}}(X)\}=p^2(X)(1-k/N)/k\simeq p^2(X)/k$][6]。

对已知样本 $X_1, X_2, \cdots, X_N$，式 (40) 还可表示成

$$\hat{\boldsymbol{p}}(X)=\frac{1}{N}\sum_{i=1}^{N} g(X-\boldsymbol{X}_i), \tag{41}$$

式中 $g(X-\boldsymbol{X}_i)$ 是一个均匀的核函数，在 $\Gamma(X)$ 中 $g=1/v$，在 $\Gamma(X)$ 外 $g=0$。一旦采用式(41)的形式，核函数不再要求是均匀的。最容易实现的核函数是高斯核：

$$g(X-X_i)=\frac{1}{\sqrt{(2\pi)^n h^n \sqrt{|\sum|}}}\exp\left[-\frac{1}{2h^2}(X-X_i)^T\sum{}^{-1}(X-X_i)\right]。 \tag{42}$$

为了具有渐近无偏性、一致性、均匀一致性，式中 h 必须分别满足 $\lim_{N\to\infty}h^n=0$，$\lim_{N\to\infty}Nh^n=\infty$，和 $\mathrm{Lim}_{N\to\infty}Nh^{2n}=\infty$[1]。$\sum$ 决定了核函数的形状，但并不知道怎样选择 $\sum$。

2. kNN 方法

对这种方法，k 是确定的，$\Gamma(X)$ 则逐渐扩展直到找到 NN。设 $\boldsymbol{v}$ 是随机变量，则

$$\hat{\boldsymbol{p}}(X)=\frac{k-1}{N\boldsymbol{v}(X)}, \quad \text{且} \quad p_{\boldsymbol{v}}(u)=\frac{N!}{(k-1)!(N-k)!}u^{k-1}(1-u)^{N-k}$$

如果能够在一个小的局部区域 $\Gamma(X)$ 中假设 $p(X)$ 是线性的，则 $p(X)v(X)$。式(43)中 $\hat{\boldsymbol{p}}(X)$ 在 $\lim_{N\to\infty}k=\infty$ 及 $\lim_{N\to\infty}k/N$ [$\lim_{N\to\infty}\mathrm{var}\{\hat{\boldsymbol{p}}(X)\}=p^2(X)\{(k-1)(N-1)/(k-$ k，Parzen 估计的方差小于 kNN 估计的方差。

须大于 2。但是对样本分散的低密度区域，Parzen 密度估计则比较困难。对于一小固定的核函数，估计结果视样本是否集中而发生明显变化。而 kNN 方法根据样本密度采取大小可变的核。

假设 $\Gamma(X)$ 是超球体，其半径为 $d_{k\mathrm{NN}}$，且假设 $u(X)=p(X)\,v(X)$，则到 k NN 的距离的 m 阶矩可以用下式计算，

$$\mathrm{E}\{\boldsymbol{d}^{m}_{k\mathrm{NN}}\mid X\}=\frac{\Gamma(k+m/n)}{\Gamma(k)}\,\frac{\Gamma(N+1)}{\Gamma(N+1+m/n)}\,\frac{\pi^{n/2}}{\Gamma(1+n/2)}\,p^{-m/n}(X) \tag{44}$$

$$\mathrm{E}\{\boldsymbol{d}^{m}_{k\mathrm{NN}}\}=\frac{\Gamma(k+m/n)}{\Gamma(k)}\,\frac{\Gamma(N+1)}{\Gamma(N+1+m/n)}\,\frac{\pi^{n/2}}{\Gamma(1+n/2)}\int p^{1-m/n}(X)\mathrm{d}X。 \tag{45}$$

对某些 $p(X)$，式(45)的积分可计算如下，

$$\text{Gauss：}(2\pi)^{m/2}|\sum|^{m/2n}(1-m/n)^{-n/2}, \tag{46}$$

$$\text{均匀分布：}(2\pi)^{m/2}|\sum|^{m/2n}\Gamma^{-m/n}(1+n/2)(1+n/2)^{m/2}。 \tag{47}$$

当 m/n 很小时，仅仅 n 和 $|\sum|$ 对矩有影响，k、N、$p(X)$ 的影响极小。

两个连续的平均后的 k NN 距离之比只决定于 k 和 n 而与 N 和 $p(X)$ 无关，即

$$\frac{\mathrm{E}\{\boldsymbol{d}_{(k+1)\mathrm{NN}}\mid X\}}{\mathrm{E}\{\boldsymbol{d}_{k\mathrm{NN}}\mid X\}}=\frac{\mathrm{E}\{\boldsymbol{d}_{(k+1)\mathrm{NN}}\}}{\mathrm{E}\{\boldsymbol{d}_{k\mathrm{NN}}\}}=1+\frac{1}{kn_1}。 \tag{48}$$

由式(48)计算的 n_1 只取决于邻域信息，因此它表明了局部维数(或内在的维数)。

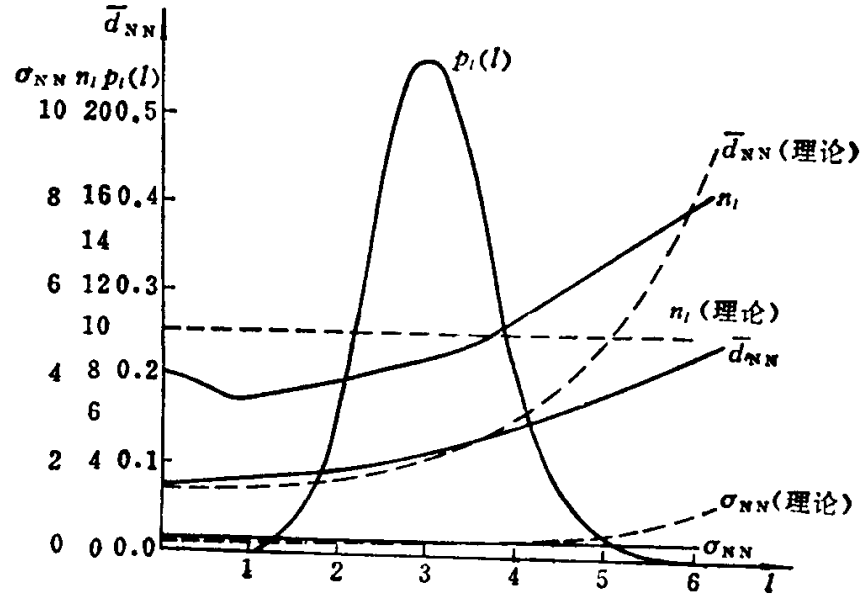


图 1 4　$\mathrm{E}\{\boldsymbol{d}_{\mathrm{NN}}\mid l\}$，$\mathrm{Var}\{\boldsymbol{d}_{\mathrm{NN}}\mid l\}$ 和 n_1 对半径 l 的变化曲线

1.4 表明了对于 $M=0$ 和 $\sum=1$ 的 10 维高斯分布及这些矩和 n_1 是怎样随位置（由半
的。论曲线由式（44）计算，经验曲线由产生的样本得到。$\mathrm{E}\{\boldsymbol{d}_{\mathrm{NN}}\mid l\}$ 和

下。

，…，$X_{k\mathrm{NN}}$，然后根据属于两个类别的近邻点的多

少对 X 进行分类[4,pp,261,280,333]。

(2) 若 k 为偶数，做法同 (1)。但当属于 ω_1 的邻点数与属于 ω_2 的邻点数相等。则 X 被拒识。这一情况不作为错分[6]。关于拒识的进一步讨论，见参考文献[3]。

假设对大的样本集有 $q_i(X_{kNN})=q_i(X)$，则对 X 而言，渐近风险可以表示如下[4,p,235]，

$$r_{2k}(X)=\sum_{i=1}^{k}\frac{1}{i}\binom{2i-2}{i-1}[q_1(X)q_2(X)]^i, \qquad \text{NN 为偶数,} \tag{49}$$

$$r_{2k-1}(X)=r_{2k}(X)+\frac{1}{2}\binom{2k}{k}[q_1(X)q_2(X)]^k, \quad \text{NN 为奇数。} \tag{50}$$

另一方面，(2)式的贝叶斯风险为

$$r^*(X)=\frac{1}{2}-\frac{1}{2}\sqrt{1-4q_1(X)q_2(X)}=\sum_{i=1}^{\infty}\frac{1}{i}\binom{2i-2}{i-1}[q_1(X)q_2(X)]^i。 \tag{51}$$

作为 $q_1(X)q_2(X)$ 函数的这些风险示于图 1.5，它们在 0～0.25 间变化。从这些曲线可以得出如下与 X 无关的不等式，

$$\frac{1}{2}r^*(X)\leqslant r_2(X)\leqslant r_4(X)\leqslant\cdots\leqslant r^*(X)\leqslant\cdots\leqslant r_3(X)\leqslant r_1(X)\leqslant 2r^*(X)。 \tag{52}$$

因为 $\varepsilon_{kNN}=E\{r_k(\boldsymbol{X})\}$，同样的不等式对渐近的 kNN 误差也成立，即

$$\frac{1}{2}\varepsilon^*\leqslant\varepsilon_{2NN}\leqslant\varepsilon_{4NN}\leqslant\cdots\leqslant\varepsilon^*\leqslant\cdots\leqslant\varepsilon_{3NN}\leqslant\varepsilon_{NN}\leqslant 2\varepsilon^*。 \tag{53}$$

图 1.5 也包括了 $\sqrt{q_1(X)q_2(X)}$，因为 $E\{\sqrt{q_1(\boldsymbol{X})q_2(\boldsymbol{X})}\}$ 是式 (35) 所示的 Bhattacharyya 上界。

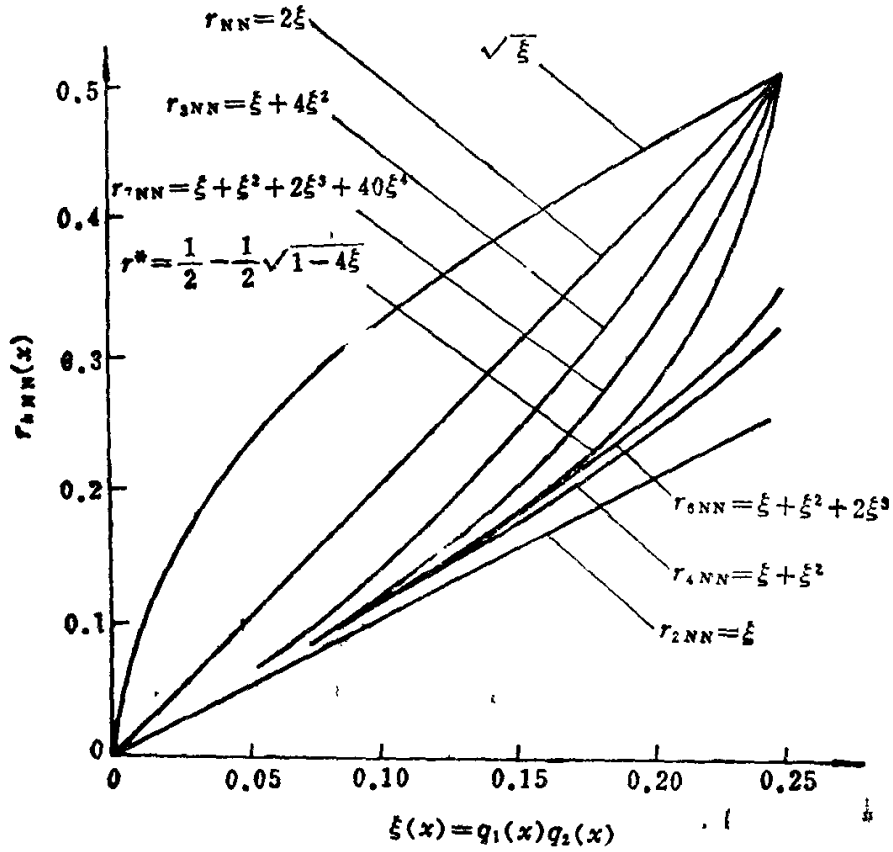


图 1.5 k 具有各种不同值时 $r_{kNN}(X)$ 的比较

当样本集大小有限时，上述不等式可能不成立。如果式(53)的不等式成立，也许就意味着可以采用 kNN 法。还要注意，$\varepsilon_{NN}=2\varepsilon_{2NN}$ $[r_1(X)=2r_2(X)]$。这一关系可以用来作为 kNN 方法有效性的另一种检验手段。

如果用下述两种方法计算 ε_{2NN}，则等式 $\varepsilon_{NN}=2\varepsilon_{2NN}$ 可以推广到多类问题中：

(1) 如果 X_{NN} 和 X_{2NN} 同类而与 X 不同类，则计 1 个误差给 X；

(2) 如果 X_{NN} 和 X_{2NN} 不同类，且两者都不与 X 同类，则计 $\frac{1}{2}$ 误差给 X。

三、有限样本集的 kNN 误差

当样本集大小有限时，kNN 错误率 $\hat{\varepsilon}_{kNN}$ 可能随 N，n，$p_i(X)$ 等的变化而发生明显变化。这是由于估计是建立在少量局部样本的基础上。

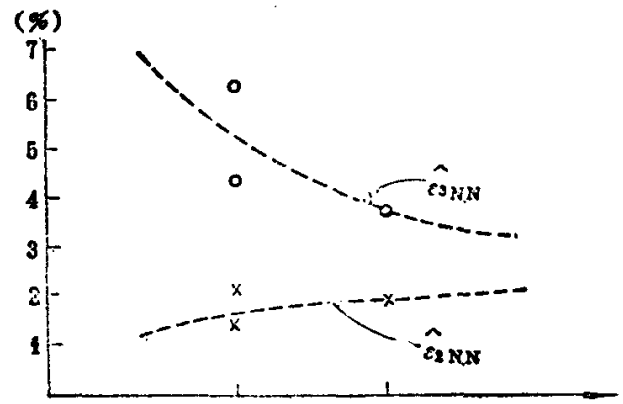


图 1.6 $\hat{\varepsilon}_{3NN}$ 和 $\hat{\varepsilon}_{2NN}$ 随 N 变化的例子

(1) N的选择　究竟需要多少样本才能得到可靠的贝叶斯误差的界限，这从理论上很难确定。但是在实际上，作 N 函数的 $\hat{\varepsilon}_{2k-1}$(上界)和 $\hat{\varepsilon}_{2k}$(下界) 的曲线将使我们对贝叶斯误差的分布和现有的 N 是否合适有一个直观而合理的了解。图 1.6 示出了一个 256 维数据集的一组曲线。对 1000 个样本，可得 $\hat{\varepsilon}_3$ 和 $\hat{\varepsilon}_2$。然后将 1000 个样本分为 500 个样本的两组，得到两个 $\hat{\varepsilon}_3$和两个 $\hat{\varepsilon}_2$。图 1.6 示出了两条给假定的贝叶斯误差定界的曲线。这幅图表明，即使增加更多的样本，比如 1000 多个，还是不能指望贝叶斯误差估计的精度有较大的提高。这个例子说明，只需要相对于维数来说较少的样本数，即可得到平滑的界限。但这在很大程度上取决于分布的结构。

(2) k 的选择　要从理论上解决最佳 k 值的选取是很困难的，对此，目前知之甚少。但是，经验表明，在不违反式(53)不等式的前提下，k 值选得尽可能大，则贝叶斯误差将得到好的估计。

(3) 度量选取　为了取得可靠的 $\hat{\varepsilon}_{kNN}$，必须适当选择测量远近的度量。对有明显不同的特征，用欧几里德距离会得到与 ε_{kNN} 相差很大的 $\hat{\varepsilon}_{kNN}$，在这种情况下，已不再是贝叶斯误差的界。

对于 $k=1$ 的最简单的情况，使 $\mathrm{E}\{[\hat{r}_{\mathrm{NN}}(X)-r_{\mathrm{NN}}(X)]^2|X\}$最小的最佳度量为[6]

$$\nabla^{\mathrm{T}}q_1(X)(\boldsymbol{X}_{\mathrm{NN}}-X)。 \tag{54}$$

就是说，$\Gamma(X)$ 中的局部样本必须变换到 $q_1(X)$ 的梯度向量 $\nabla q_1(X)$，具有 $\nabla q_1(X)$ 的最小分量的邻域选为最近邻(NN)。$\nabla q_1(X)$ 与似然比 $-\ln[p_1(X)/p_2(X)]$ 的梯度向量成比例，即 $\nabla q_1(X)\sim[-\nabla p_1(X)/p_1(X)+\nabla p_2(X)/p_2(X)]$。

如果 $p_1(X)$ 和 $p_2(X)$ 都是高斯分布[6]，那末

$$\nabla q_1(X)\sim\sum\nolimits_1^{-1}(X-M_1)-\sum\nolimits_2^{-1}(X-M_2)。 \tag{55}$$

这一向量可以用于许多边界表面结构与高斯分布类似的分布。

当分布与高斯分布相差甚远时，必须从样本估计 $\nabla q_1(X)$。为此，使用下式[6]

$$\nabla q_1(X)\sim\mathrm{E}\{(\boldsymbol{Y}-X)|\Gamma(X),\omega_1\}-E\{(\boldsymbol{Y}-X)|\Gamma(X),\omega_2\}。 \tag{56}$$

为了估计 $\nabla q_1(X)$，式 (56)中的期望用相应的样本平均来代替。式(56)是以下式的知识为基础的，即

$$E\{(\boldsymbol{Y}-X)|\Gamma(X),\omega_i\}=\frac{2\,d^2\nabla p_i(X)}{n+2\,p_i(X)}。 \tag{57}$$

式中假设 $\Gamma(X)$ 是一个围绕 X 的局部超球区域，d 是其半径。式(57)用来估计用于聚类和其他应用的密度函数的梯度。

(4) 极化的 2 NN 对 $k=2$，$\hat{r}_{2\mathrm{NN}}(X)-r_{2\mathrm{NN}}(X)$ 的台劳级数设有一次项，而从二次项开始。这意味着 $\varepsilon_{2\mathrm{NN}}$ 的估计从根本上说比 $\varepsilon_{\mathrm{NN}}$ 的估计可靠。同样的特点在一般的 $2k$NN 误差的估计中也可见到。

此外，用下述方法来选择 X_{NN} 和 $X_{2\mathrm{NN}}$ 能够使 $\hat{\varepsilon}_{2\mathrm{NN}}$ 更加接近 $\varepsilon_{2\mathrm{NN}}$：

通过使极化的欧几里德距离 $\|(X_i-X)+(X_k-X)\|$ 达到最小在 $\Gamma(X)$ 中的局部样本中寻找最好的对偶 X_i 和 X_k。

这一过程称为极化的 2 NN 选择规则。为了提高可靠性，可以用最佳的度量来代替欧几里德距离。由于极化的 $\hat{\varepsilon}_{2\mathrm{NN}}$ 比常规的 $\hat{\varepsilon}_{\mathrm{NN}}$ 可靠，可以用 $2\hat{\varepsilon}_{2\mathrm{NN}}$ 来代替 $\hat{\varepsilon}_{NN}$ 以得到贝叶斯误差的上界。

四、使用距离于 k NN

到目前为止，k NN 法是作为一个计数过程，即比较 ω_1 和 ω_2 邻居的数目。把距离信息用于第 k 个 NN 则可对此进行修正。

因为 $v=cd^n$，式(43)可重写为

$$-\ln\hat{\boldsymbol{P}}_i(X)=n\ln\boldsymbol{d}_i(X)+\ln\frac{cN_i}{k-1}, \tag{58}$$

式中 $\boldsymbol{d}_i(X)$ 是从 X 到 ω_i 的第 k 个 NN，即 $X_{k\mathrm{NN}}^{(\omega_i)}$ 的距离。比较式(58)和式(7)表明 $n\ln\boldsymbol{d}_i(X)$ 是 Mahalanobis 距离 $\frac{1}{2}(X-M_i)^{\mathrm{T}}\sum_i^{-1}(X-M_i)$ 的无参形式。将式(58)代入式(6)，贝叶分类器变为

$$n\ln d_2(X)\underset{\omega_2}{\overset{\omega_1}{\gtrless}}n\ln d_1(X)+\ln\frac{N_1 i}{N_2}$$

因此，λ 被分为 $d_1(\lambda)$ 最小的一类。因为式(59)可以写成 $d_2(X)\gtrless d_1(X)(N_1P_2/N_2P_1)^{1/n}$，所以式中的常数项表明了对各种不同的 N_i 和 P_i 应该怎样调整距离的尺度。

1．X 的二维显示[6]

以 $n\ln d_1(X)$ 和 $n\ln d_2(X)$ 作为 x 轴和 y 轴，可以将 X 表示成图 1.7 的样子。图中，有取自 ω_1 的样本 50 个，ω_2 的样本 150 个，它们是根据[1]中所用的八维高斯分布产生出来的。移动 45° 线可以从这一图中发现加于距离的适当的权。这种显示的优点在于它的灵活性。一旦样本被显示，则可以通过改变边界的斜率，移动线，甚至选择比较复杂的曲线来形成边界线。

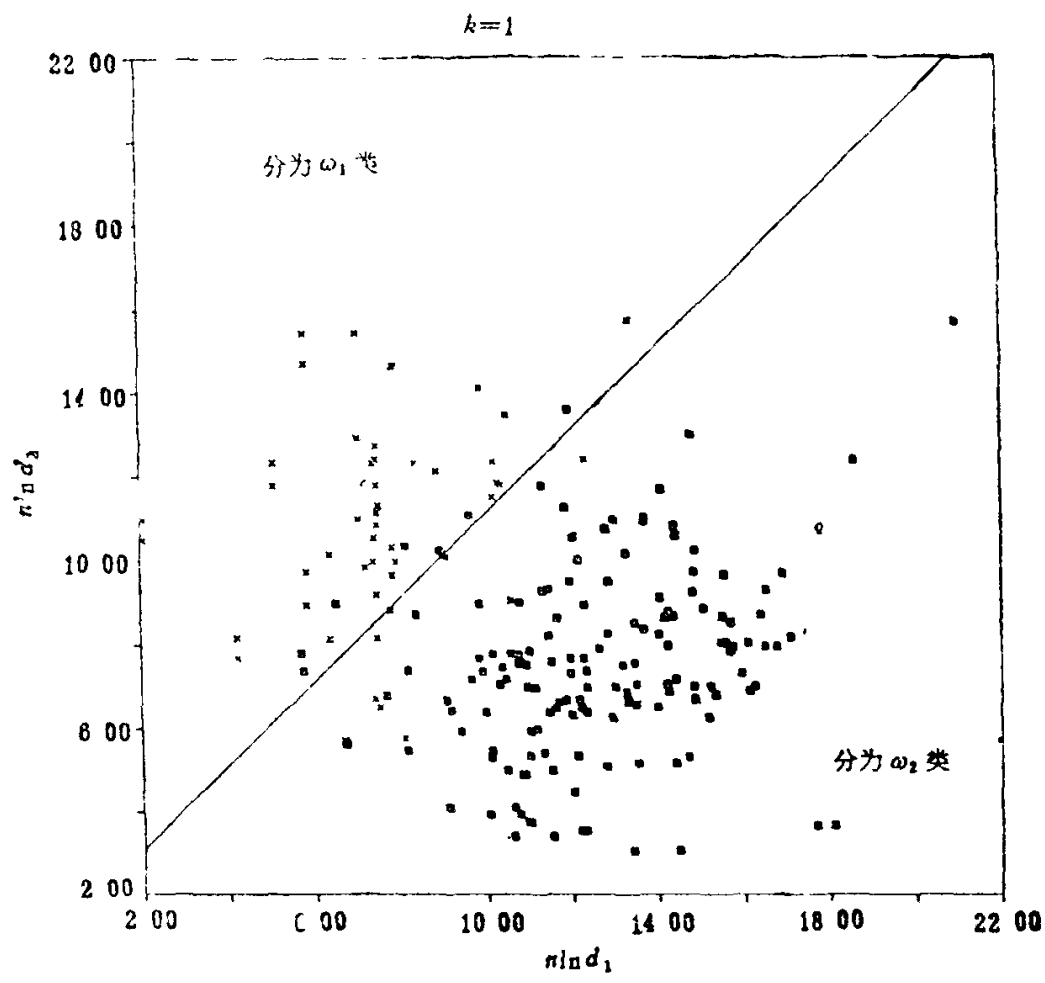


图 1.7 非参数图示，$P_1=P_2=0.5$，$N_1=50$，$N_2=150$

另外，对于 Neyman-Pearson 和极小极大检验，通过移动 45° 线分别使 $\varepsilon_2=\varepsilon_0$，$\varepsilon_1=\varepsilon_2$ 就可以找到边界。从理论上可以确定阈值．但是，在数据显示的图上形象地选择线的位置则比较灵活、直观，以至很多情况下可以获得比较好的结果。

2．贝叶斯风险等值线[6]

结合式(2)和式(43)，贝叶斯风险与 $d_1(X)$ 和 $d_2(X)$ 的关系为

$$n\ln \quad d_1(X)+\ln\frac{N_1P_2}{N_2P_1}\pm\ln\frac{r^*(X)}{1-r^*(X)}, \tag{60}$$

(X) 和 $q_1(X)>q_2(X)$ 的情况。因此，对于给定程度的

$r^*(X)$,等值线成了 45° 线。图 1.8 示出了对各种 $r^*(X)$ 值的等值风险线。这些风险线关于贝叶斯分类器 $[r^*(X)=0.5]$ 为对称。图 1.8 中还画出了混合密度函数 $p(X)$ 的等值线。

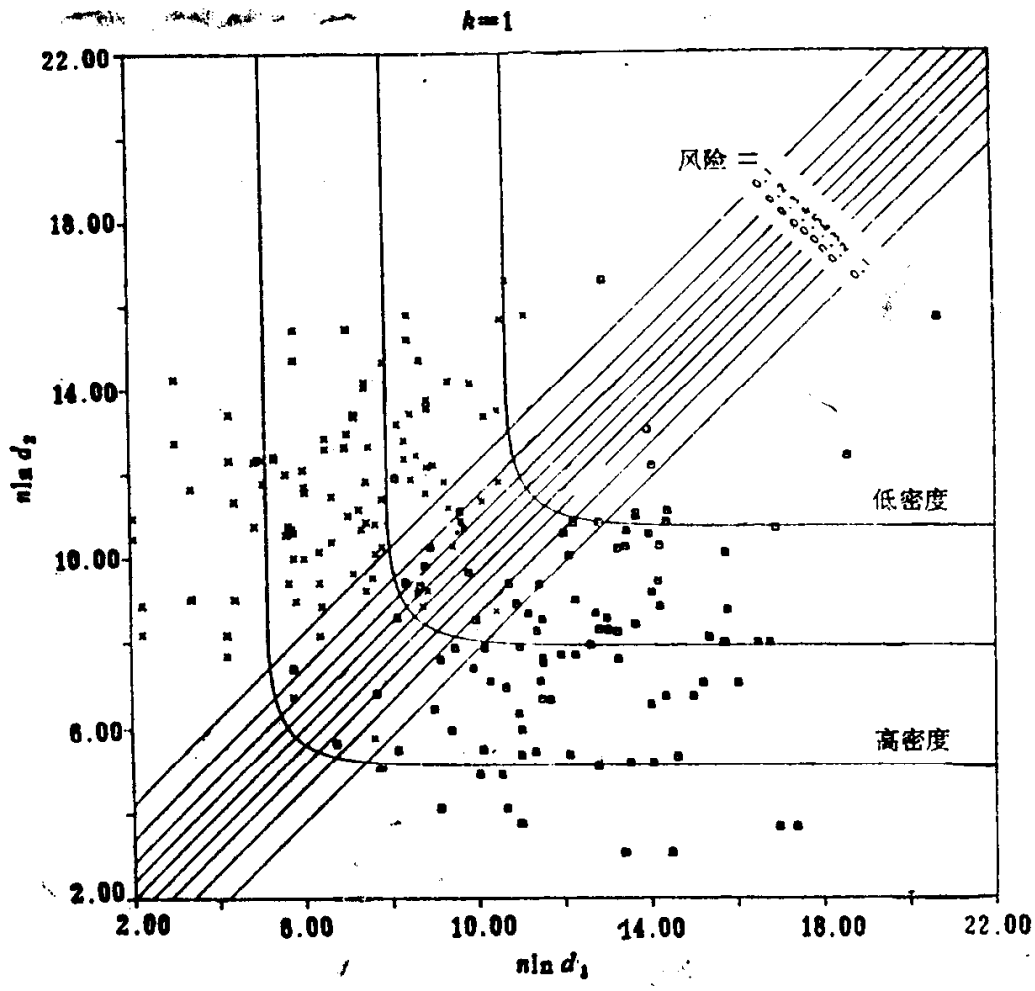


图 1.8　风险和密度等值线的非参图示

当想要拒识如 $0.4 \leqslant r^*(X) \leqslant 0.5$ 的样本时，处于两条 0.4 风险线之间的样本就被去除。被拒识的样本数除以 N 就是式(16)的 $R(0.4)$。式(17)的 $\varepsilon(0.4)$ 也可以通过计算两条 0.4风险线以外的错分样本的数目来得到。

3．贝叶斯误差的估计[6]

对第 k 个最近邻使用距离信息，贝叶斯误差本身可以估计如下

$$\hat{r}^*(X)=\min\left[\frac{\hat{P}_1\hat{p}_1(X)}{\hat{P}_1\hat{p}_1(X)+\hat{P}_2\hat{p}_2(X)},\ \frac{\hat{P}_2\hat{p}_2(X)}{\hat{P}_1\hat{p}_1(X)+\hat{P}_2\hat{p}_2(X)}\right]$$

$$=\min\left[\frac{V_2(X)}{V_1(X)+V_2(X)},\ \frac{V_1(X)}{V_1(X)+V_2(X)}\right], \tag{61}$$

式中 $\hat{P}_i=N_i/N$，$\hat{p}(X)$ 如式(43)，且 $k_1=k_2$。用样本均值代替期望就得到贝叶斯误差的估计，即

$$\hat{\varepsilon}^*=\frac{1}{N}\sum_{i=1}^{N}\hat{r}^*(X_i)。$$

已经知道这种估计比起常规的误差计算具有较小的方差。虽然并不能保证对贝

但是经验表明，这种估计的值比 $\hat{\varepsilon}_{NN}$ 和 $\hat{\varepsilon}_{2NN}$ 更接近于贝叶斯误差。为了得到可靠的估计，选择适当的度量来测量邻近的程度是重要的。

五、Parzen 法

当用 Parzen 密度估计进行分类时，在由 Parzen 法引起的误差和贝叶斯误差之间不存在已知的关系。但是如下所述，有一种比较一般的原则用来寻找贝叶斯误差的界。

设 $S=\{P_1,p_1(X),P_2,p_2(X)\}$ 是所有实际分布信息的集合，又设 $\hat{S}=\{\hat{P}_1,\hat{p}_1(X),\hat{P}_2,\hat{p}_2(X)\}$ 是它们的估计的集合。分类误差取决于这两个集合，写作 $\varepsilon(S_D,S_T)$。这意味着，贝叶斯分类器按 S_D 设计，误差则通过 S_T 来计算。S_D 和 S_T 不一定相同。则贝叶斯误差 $\varepsilon(S,S)$ 由下式定界[11]，

$$E\{\varepsilon(\hat{S},\hat{S})\}\leqslant\varepsilon(S,S)\leqslant E\{\varepsilon(\hat{S},S)\} \tag{63}$$

这一界限是根据假设 $E\{\varepsilon(S,\hat{S})\}=\varepsilon(S,S)$，即误差估计对检验密度来说是无偏的。

通过对与设计数据相同的数据进行检验可得下界，这一检验过程称为重复置换法。至于上界，对用给定的数据集设计的贝叶斯分类器，必须检验实际的分布。因为不可能知道实际的分布，所以常用数据集 $\hat{S}_T$ 来代替 S，$\hat{S}_T$ 不作设计用。虽然 $E\{\varepsilon(\hat{S},\hat{S}_T)\}$ 不能保证对贝叶斯误差定界，但这种方式的误差大于贝叶斯误差这一点通常是能够接受的。这种方法称为“失控法”(hold-out method)。另一种确定上界的方法是留一个出来的方法 (leaving-one-out method)。设有 N 个可用的样本，分别对每一个样本作检验，检验时的分类器是用除了检验样本以外的另外 $(N-1)$个样本设计的。这种检验重复 N 次。因此，用这种方法能使可用的样本比较有效地利用。另外，不必担心怎样分离样本，这是失控法的一个问题。

将上述检验过程与式(41)的 Parzen 估计相结合，可以得到下列方程[11]。

(1) 重复置换法：

$$\frac{N_1}{N}\frac{1}{N_1}\sum_{i=1}^{N_1}g(X_k-X_i^{(1)})\underset{\omega_2}{\overset{\omega_1}{\gtrless}}\frac{N_2}{N}\frac{1}{N_2}\sum_{j=1}^{N_2}g(X_k-X_j^{(2)}), \tag{64}$$

其中 $X_k\in\{X_1^{(1)},\cdots,X_{N_1}^{(1)},X_1^{(2)},\cdots,X_{N_2}^{(2)}\}$，式中 $X_j^{(i)}$ 是 ω_i 中第 j 个样本。

(2) 留一个出来的方法：

当检验样本 $X_k\in\{X_1^{(1)},\cdots,X_{N_1}^{(1)}\}$ 时，

$$\frac{N_1-1}{N-1}\frac{1}{N_1-1}\left[\sum_{i=1}^{N_1}(X_k-X_i^{(1)})-g(0)\right]\underset{\omega_2}{\overset{\omega_1}{\gtrless}}\frac{N_2}{N-1}\frac{1}{N_2}\sum_{j=1}^{N_2}g(X_k-X_j^{(2)})。 \tag{65}$$

当检验样本 $X_k\in\{X_1^{(2)},\cdots,X_{N_2}^{(2)}\}$ 时，

$$\frac{N_1}{N-1}\frac{1}{N_1}\sum_{i=1}^{N_1}g(X_k-X_i^{(1)})\underset{\omega_2}{\overset{\omega_1}{\gtrless}}\frac{N_2-1}{N-1}\frac{1}{N_2-1}\sum_{j=1}^{N_2}[g(X_k-X_j^{(2)})-g(0)]。 \tag{66}$$

当用留一个出来的方法对 ω_i 中的某个样本进行检验时会降低 ω_i 的密度，而且误差总是比重复置换法大。另外，要注意留一个出来的方法的一项附加的运算是减去常数 $g(0)$，因此，当用重复置换方法计算误差时，留一个出来的方法的误差几乎同时算出，附加的计算时间可以忽略。

Parzen 的关键参数是核函数的数学形式和大小。但是怎样选择这些参数并不太清楚。