模式识别

编著 黄凤岗 宋克欧

哈尔滨工程大学出版社

模 式 识 别

黄凤岗
宋克欧 编著

哈尔滨工程大学出版社

内容简介

本书主要讨论了统计模式识别的理论和方法，包括贝叶斯分类器、线性分类器、非线性分类器、聚类分析、特征选择与提取等，并以神经网络算法贯穿其中，系统而有机地组织了全书的内容。

书中注重物理概念的阐述和具体算法的实现，并顺应模式识别的发展，介绍了一些新的研究成果。本书可作为计算机、通讯、自动控制等专业研究生或高年级本科生的教材，亦适合于相应专业的科技人员参考。

模式识别
MOSHI SHIBIE
编著　黄凤岗　宋克欧
责任编辑　张　奎

哈尔滨工程大学出版社出版发行
（哈尔滨南岗文庙街11号楼）邮编 150001
新　华　书　店　经　销
哈尔滨工业大学印刷厂印刷

开本 850×1 168 1/32 印张 5.5 字数 142 千字
1998年3月第1版　1998年3月第1次印刷
印数：1～2 000册
ISBN 7-81007-721-X
TP·50　定价：7.80元

前　言

模式识别所要处理和解决的问题是一个复杂的难以用语言准确完整表达的问题，因此模式识别系统必须具备自学习、自组织、自适应的柔性处理能力。神经网络就具备这种能力，有资格作为模式识别的基础模型。事实上，传统的一些分类算法、特征提取算法与神经网络算法是一脉相承的。基于感知器准则函数和平方误差准则函数的线性分类器学习算法可作为单层前向神经网络的学习算法。BP 网就是一种分段线性分类器。高阶神经网络是一种广义线性分类器。Kohonen 聚类网络算法与 c -均值算法本质上是一致的。对这些，本书第三、四、五章作了详细论述。

将模糊理论引入模式识别，有利于提高模式识别系统的柔性处理能力。尤其是模糊理论与神经网络相结合，可能是模式识别的重要发展方向之一，第六章对此作了讨论。

用神经网络进行特征提取也受到人们的重视，第七章作了介绍。

此外，本书还介绍了模拟退火算法、遗传算法等“热门”算法。

由于作者水平有限，书中错误与不当之处在所难免，敬请读者批评指正。

编著者

1997 年 7 月于哈尔滨

目　录

第一章 绪 论

1.1 模式识别的概念

人有这样一种能力：听到走廊里的脚步声，就知道谁来了；在人群中掠过一个人的背影，就能认出这个人是谁；留言条上的字，一看就知道是谁写的，而且尽管写得龙飞凤舞，还是能认出写的是什么意思。这种能力就是人的识别能力。

随着计算机科学的发展和计算机应用的普及，迫切希望计算机也能听懂我们说的话，看懂我们写的字，……。这种强烈愿望和不断探索实践，促使模式识别这门学科得以形成和发展。

什么是模式和模式识别呢？按照广义的定义，模式是一些供模仿用的、完美无缺的标本。模式识别就是识别出特定客体所模仿的标本。根据客体的性质，将客体分成两种类型：抽象的客体和具体的客体。论点、思想、信仰、……，是非物质的客体，对它们的研究主要属于哲学、政治学的范畴；声音、图象、文字、……，是具体的客体，它们通过对感官的刺激而被识别。研究人类对具体客体的识别机理是沿着两个方向进行的。一个方向是研究人类对客体识别能力的生物学机理，这属于生物学、生理学及心理学的范畴；另一方向是研究用计算机模拟人的识别能力，提出识别具体客体的基本理论与实用技术，这就是模式识别这一学科的研究内容。根据模式识别的研究内容，我们对模式和模式识别作如下狭义的定义：**模式是对感兴趣的客体的定量的或结构的描述；模式类是具有某些共同特性的模式的集合。模式识别是研究一些自动技术，依靠这些技术，计算机自动地（或者人进行少量干涉）把待识模式分到各自的**

模式类中去。

1.2 模式识别的方法

从上一节中模式的定义可以看出，描述模式有两种方法：定量描述和结构性描述。定量描述就是用一组数据来描述模式。比如，判断某细胞是正常细胞还是癌变细胞，我们可以抓住两个特征，一个是细胞的圆形度 x_1，一个是细胞的形心偏差度 x_2。因为，正常细胞比较圆，即圆形度大；正常细胞的细胞核中心偏离细胞中心小，即形心偏差度小；而癌变细胞的圆形度小，形心偏差度大，这两个特征能比较有效地区分正常细胞与癌变细胞，所以，我们可以用这两个特征来描述细胞。为了便于数学处理，我们把这些数据特征组成向量，称为特征向量。比如，把上述描述细胞的两个特征（圆形度 x_1 和形心偏差度 x_2）组成描述细胞的特征向量：$\boldsymbol{x}=(x_1,x_2)^{\mathrm{T}}$，其中 T 是转置符号。另一种描述模式的方法是结构性描述，即用一组基元来描述模式。比如，我们可以用一组基元来描述图 1-1 中的图形。

这个图形由四个基本元素（称为基元）组成。两个圆弧段分别用符号 a 和 c 表示，直线段用符号 b 表示。同样，为了便于用形式语言处理，我们把这些符号组成符号串：$\boldsymbol{x}=abcb$，这四个基元不仅分别表达了图形中四个线段的局部特征，而且这些基元之间的连接关系从结构上描述了这个图形。

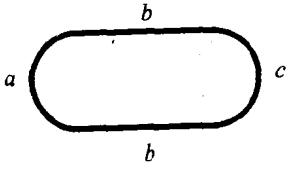

图 1-1

相应于两种模式描述方法，有两种基本的模式识别方法：统计模式识别方法和结构（句法）模式识别方法。在统计模式识别方法中，用特征向量描述模式；在结构模式识别方法中，用符号串（树）来描述模式。本书只讨论统计模式识别方法。基于统计识别法的

模式识别系统主要由五部分组成：数据获取、预处理、特征抽取、分类器设计和分类器。具体见图 1－2。

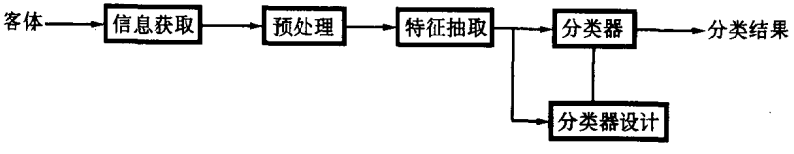

图 1－2

下面对这五个部分作些说明。

1. 数据获取

为了使计算机能够对客体进行分类识别，必须将客体用计算机所能接受的形式表示，通常从客体获得的信息有下列三种类型：

①二维图像，如文字、指纹、照片等；

②一维波形，如语音、机械振动波、心电图等；

③物理参量和逻辑值，如体温、各种实验数据等。

通过测量、采样和量化，可以用矩阵或向量表示二维图象或一维波形，这就是信息获取过程。

2. 预处理

预处理的目的是去除噪声，加强有用的信息，并对种种因素造成的退化现象进行复原。

3. 特征抽取

由信息获取部分获得的原始数据量一般是相当大的。为了有效地实现分类识别，要对原始数据进行选择或变换，得到最能反映分类本质的特征，构成特征向量。这就是特征抽取的过程。

4. 分类器设计

为了把待识模式分配到各自的模式类中去，必须设计出一套分类判别规则。基本作法是：用一定数量的样本（称为训练样本集），确定出一套分类判别规则，使得按这套分类判别规则对待识

模式进行分类所造成的错误识别率最小或引起的损失最小。这就是分类器设计的过程。

5. 分类器

分类器按已确定的分类判别规则对待识模式进行分类判别，输出分类结果。

模式识别系统中的第一和第二部分是数字信号处理和图象处理等课程的研究课题，本书只讨论第三、第四和第五部分的理论和方法。

模式识别是60年代发展起来的一门学科。目前，模式识别技术已在语音识别、文字识别、图象识别等领域得到成功的应用。但是由于问题的复杂性，离人们的期望还有一段距离。因此，模式识别仍然是一门发展中的新兴学科，新的理论和方法不断出现，同时，与其它学科相互结合和相互渗透，不断推动模式识别向前发展。模糊数学和神经网络引入模式识别就是典型的例子。由于客体的特征常常带有某些模糊，因此，1965年扎德提出模糊集合的概念后不久，模糊数学就深入到模式识别的许多环节。80年代掀起的人工神经网络研究热潮很快就波及到模式识别，出现了模糊模式识别和神经网络模式识别的提法。本书不把模糊模式识别和神经网络模式识别从统计模式识别中分离出来，而是以人工神经网络为主干线(或称为平台)将现有的模式识别方法融汇贯通起来，培育一个独特的模式识别体系结构，让读者从中掌握到模式识别的精髓，为以后从事模式识别研究打下坚实的基础。

第二章　贝叶斯分类器

模式识别的分类问题就是根据待识客体的特征向量值及其它约束条件将其分到某个类别中去。贝叶斯决策理论是处理模式分类问题的基本理论之一。本章要讨论的贝叶斯分类器在统计模式识别中被称为最优分类器。采用贝叶斯分类器必须满足下列两个先决条件：

①要决策分类的类别数是一定的；

②各类别总体的概率分布是已知的。

在条件①中，假设要研究的分类问题有 c 个类别，各类别状态用 ω_i 来表示，$i=1,2,\cdots,c$。在条件②中，假设待识客体的特征向量值 $\boldsymbol{x}$ 所对应的状态后验概率 $P(\omega_i/\boldsymbol{x})$ 是已知的；或者，对应于各个类别 ω_i 出现的先验概率 $P(\omega_i)$ 和类条件概率密度函数 $p(\boldsymbol{x}/\omega_i)$ 是已知的。

2.1　最小错误率贝叶斯决策

我们从一个实际例子讲起。假如我们要进行细胞识别，识别某细胞属正常类 ω_1 还是属癌变类 ω_2？假设待识细胞已作过预处理，抽取出二个特征：圆形度 x_1 和形心偏差度 x_2，组成特征向量 $\boldsymbol{x}$（或称为模式 $\boldsymbol{x}$）。现在要识别模式 $\boldsymbol{x}$ 属正常类 ω_1 还是属癌变类 ω_2，根据什么规则来识别？一个直观的想法是：如果模式 $\boldsymbol{x}$ 属正常类 ω_1 的概率大于模式 $\boldsymbol{x}$ 属癌变类 ω_2 的概率，则决策模式 $\boldsymbol{x}$ 属正常类 ω_1；反之，如果模式 $\boldsymbol{x}$ 属正常类 ω_1 的概率小于模式 $\boldsymbol{x}$ 属癌变类 ω_2 的概率，则决策模式 $\boldsymbol{x}$ 属癌变类 ω_2。这句话用数学语言可表示为：

若　　$$P(\omega_1/\boldsymbol{x})\gtrless P(\omega_2/\boldsymbol{x})$$

则 $$x\in\begin{matrix}\omega_1\\ \omega_2\end{matrix}$$

其中，条件概率 $P(\omega_i/x)$称为状态的后验概率。

利用贝叶斯公式

$$P(\omega_i/x)=\frac{p(x/\omega_i)P(\omega_i)}{p(x)}$$

上面的决策规则可改写为：

若 $$\frac{p(x/\omega_1)P(\omega_1)}{p(x)}\gtrless\frac{p(x/\omega_2)P(\omega_2)}{p(x)}$$

则 $$x\in\begin{matrix}\omega_1\\ \omega_2\end{matrix}$$

其中，$p(x)>0$，将不等式两边的分母消去，决策规则可改写为：

若 $$p(x/\omega_1)P(\omega_1)\gtrless p(x/\omega_2)P(\omega_2)$$

则 $$x\in\begin{matrix}\omega_1\\ \omega_2\end{matrix}$$

其中，$p(x/\omega_1)$是正常类下模式 x 的类条件概率密度，$p(x/\omega_2)$是癌变类下模式 x 的类条件概率密度。

这样，最小错误率贝叶斯决策有两种形式，一种是后验概率形式：

若 $$P(\omega_1/x)\gtrless P(\omega_2/x)$$

则 $$x\in\begin{matrix}\omega_1\\ \omega_2\end{matrix}$$

另一种是类条件概率密度形式：

若 $$p(x/\omega_1)P(\omega_1)\gtrless p(x/\omega_2)P(\omega_2)$$

则 $$x\in\begin{matrix}\omega_1\\ \omega_2\end{matrix}$$

将二类情况推广到 c 类情况，最小错误率贝叶斯决策规则为：

(1)后验概率形式

若 $$P(\omega_i/x)>P(\omega_j/x),\quad j=1,2,\cdots,c;\quad j\neq i$$

则 $$x\in\omega_i \tag{2.1}$$

(2)类条件概率密度形式

若

$$p(\boldsymbol{x}/\omega_i)P(\omega_i)>p(\boldsymbol{x}/\omega_j)P(\omega_j),\qquad j=1,2,\cdots,c;\ j\neq i$$

则

$$\boldsymbol{x}\in\omega_i \tag{2.2}$$

例 2.1　有一家医院为了研究癌症的诊断，对一大批人作了一次普查，给每人打了试验针，然后进行统计，得到如下统计数字：

①这批人中，每 1 000 人有 5 个癌症病人；

②这批人中，每 100 个正常人有 1 人对试验的反应为阳性；

③这批人中，每 100 个癌症病人有 95 人对试验的反应为阳性。

假如正常人用 ω_1 类表示，癌症病人用 ω_2 类表示。以试验结果作为特征，特征值为阳或阴。根据统计数字，得到如下概率：

$P(\omega_1)=0.995$，$P(\omega_2)=0.005$，$p(\text{阳}/\omega_1)=0.01$，

$p(\text{阴}/\omega_1)=0.99$，$p(\text{阳}/\omega_2)=0.95$，$p(\text{阴}/\omega_2)=0.05$。

通过普查统计，该医院可开展癌症诊断。现在王某，试验结果为阳性，诊断结果是什么？

因为

$$p(\boldsymbol{x}/\omega_1)P(\omega_1)=p(\text{阳}/\omega_1)P(\omega_1)=0.01\times0.995=0.009\ 95$$

$$p(\boldsymbol{x}/\omega_2)P(\omega_2)=p(\text{阳}/\omega_2)P(\omega_2)=0.95\times0.005=0.004\ 75$$

故

$$p(\boldsymbol{x}/\omega_1)P(\omega_1)>p(\boldsymbol{x}/\omega_2)P(\omega_2)$$

所以 $\boldsymbol{x}\in\omega_1$，即王某属正常人。

上面我们介绍了贝叶斯决策规则。应用贝叶斯决策规则对模式 $\boldsymbol{x}$ 进行分类的分类器称为贝叶斯分类器 。

对于 c 类分类问题，按照决策规则可以把特征向量空间（或称模式空间）分成 c 个决策域。我们将划分决策域的边界称为决策边界，在数学上用解析形式可以表示成决策边界方程。用于表达决策规则的某些函数称为判别函数。判别函数与决策边界方程是密切相关的，而且它们都由相应的决策规则所确定。

对于 c 类分类问题，通常定义 c 个判别函数 $d_i(\boldsymbol{x})$，$i=1,2,\cdots,c$. 对照两种形式的最小错误率贝叶斯决策规则，判别函数显然可定义为：

①$d_i(\boldsymbol{x})=P(\omega_i/\boldsymbol{x})$，　$i=1,2,\cdots,c$；

②$d_i(\boldsymbol{x})=p(\boldsymbol{x}/\omega_i)P(\omega_i)$，　$i=1,2,\cdots,c$.

这样，决策规则可写为：

若　$d_i(\boldsymbol{x})>d_j(\boldsymbol{x})$，　$j=1,2,\cdots,c;j\neq i$

则　$\boldsymbol{x}\in\omega_i$

确定了判别函数，决策边界也就确定下来了。相邻的两个决策域在决策边界上其判别函数值是相等的。如果决策域 R_i 与 R_j 是相邻的，则分割这两个决策域的决策边界方程应满足：

$$d_i(\boldsymbol{x})=d_j(\boldsymbol{x})$$

一般地说，模式 $\boldsymbol{x}$ 为一维时，决策边界为一分界点；$\boldsymbol{x}$ 为二维时，决策边界为一曲线；$\boldsymbol{x}$ 为三维时，决策边界为一曲面；$\boldsymbol{x}$ 为 n 维（$n>3$）时，决策边界为一超曲面。

分类器可看成是由硬件或软件组成的“机器”，贝叶斯分类器的结构见图 2-1。

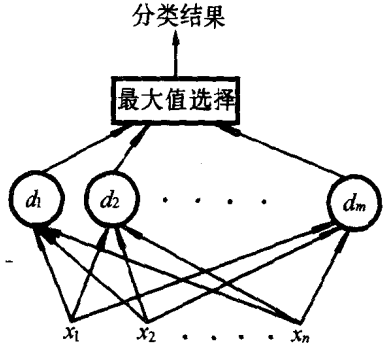

图 2-1

分类器先计算出 c 个判别函数 $d_i(\boldsymbol{x})$ 值，再从中选出对应于判

别函数值为最大的类别作为分类结果。

2.2 最小风险贝叶斯决策

在例 2.1 中，王某的试验结果为阳性，根据最小错误率贝叶斯决策，判他属正常人，那么他属正常人的概率是不是 100%呢？我们可计算出试验结果为阳性的条件下他属正常人的概率：

$$P(\omega_1/\text{阳})=\frac{p(\text{阳}/\omega_1)P(\omega_1)}{p(\text{阳}/\omega_1)P(\omega_1)+p(\text{阳}/\omega_2)P(\omega_2)}$$

$$=\frac{0.01\times0.995}{0.01\times0.995+0.95\times0.005}\approx67.7\%$$

该人属正常人的概率为 67.7%，换句话说，他属癌症病人的概率为 32.2%。

从这里可以看出，尽管采用了最小错误率贝叶斯决策，但仍然可能将正常人错判为癌症病人，也可能将癌症病人错判为正常人。这些错判都会带来一定的损失。将正常人错判为癌症病人，会给他带来短期的精神负担，造成一定的损失，这个损失比较小。如果把癌症病人错判为正常人，致使患者失去挽救的机会，这个损失就大了。这两种不同的错判所造成损失的程度是有显著差别的。所以，在决策时还要考虑到各种错判所造成的不同损失，提出了最小风险贝叶斯决策。

风险是什么？条件风险定义为：**将模式 x 判属某类所造成的损失的条件数学期望**。

仍以细胞识别为例。假定：

模式 x 本属正常类而判属正常类所造成的损失为 L_{11}；

模式 x 本属癌变类而判属正常类所造成的损夫为 L_{21}；

模式 x 本属正常类而判属癌变类所造成的损失为 L_{12}；

模式 x 本属癌变类而判属癌变类所造成的损失为 L_{22}.

根据条件风险的定义，将模式 x 判属正常类 ω_1 的条件风险为将模式 x 判属 ω_1 类所造成的损失的条件数学期望：

$$r_1(\boldsymbol{x})=L_{11}P(\omega_1/\boldsymbol{x})+L_{21}P(\omega_2/\boldsymbol{x})$$

同理，将模式 $\boldsymbol{x}$ 判属癌变类 ω_2 的条件风险为：

$$r_2(\boldsymbol{x}) = L_{12}P(\omega_1/\boldsymbol{x}) + L_{22}P(\omega_2/\boldsymbol{x})$$

我们可以根据条件风险的大小来决策。如果将 $\boldsymbol{x}$ 判属 ω_1 类的条件风险 $r_1(\boldsymbol{x})$ 小于判属 ω_2 类的条件风险 $r_2(\boldsymbol{x})$，则决策 $\boldsymbol{x}$ 属 ω_1 类；反之，如果将 $\boldsymbol{x}$ 判属 ω_1 类的条件风险 $r_1(\boldsymbol{x})$ 大于判属 ω_2 类的风险 $r_2(\boldsymbol{x})$，则决策 $\boldsymbol{x}$ 属 ω_2 类。即

若

$$L_{11}P(\omega_1/\boldsymbol{x})+L_{21}P(\omega_2/\boldsymbol{x})\lessgtr L_{12}P(\omega_1/\boldsymbol{x})+L_{22}P(\omega_2/\boldsymbol{x})$$

则

$$\boldsymbol{x}\in\begin{matrix}\omega_1\\ \omega_2\end{matrix}$$

利用贝叶斯公式，上面的决策规则改写为：

若

$$L_{11}\frac{p(\boldsymbol{x}/\omega_1)P(\omega_1)}{p(\boldsymbol{x})}+L_{21}\frac{p(\boldsymbol{x}/\omega_2)P(\omega_2)}{p(\boldsymbol{x})}\lessgtr L_{12}\frac{p(\boldsymbol{x}/\omega_1)P(\omega_1)}{p(\boldsymbol{x})}+L_{22}\frac{p(\boldsymbol{x}/\omega_2)/P(\omega_2)}{p(\boldsymbol{x})}$$

消去 $p(\boldsymbol{x})$，简化为：

若

$$L_{11}p(\boldsymbol{x}/\omega_1)P(\omega_1)+L_{21}p(\boldsymbol{x}/\omega_2)P(\omega_2)\lessgtr L_{12}p(\boldsymbol{x}/\omega_1)P(\omega_1)+L_{22}p(\boldsymbol{x}/\omega_2)P(\omega_2)$$

则

$$\boldsymbol{x}\in\begin{matrix}\omega_1\\ \omega_2\end{matrix} \tag{2.3}$$

将两类情况推广到 c 类情况：

①后验概率形式

$$r_i(\boldsymbol{x}) = \sum_{k=1}^{c}L_{ki}P(\omega_k/\boldsymbol{x}) \tag{2.4}$$

②类条件概率密度形式

$$r_i(\boldsymbol{x}) = \sum_{k=1}^{c}L_{ki}p(\boldsymbol{x}/\omega_k)P(\omega_k) \tag{2.5}$$