《信息、控制与系统》系列教材

模 式 识 别

边肇祺等 编著

清华大学出版社

内 容 简 介

本书是清华大学自动化系教材，主要讨论统计模式识别理论和方法，包括贝叶斯决策理论、线性和非线性判别函数、近邻规则、经验风险最小化、特征提取和选择以及聚类分析等．还介绍了模式识别在一维数字信号和二维数字图象识别中的应用．每章后附有习题，适于教学和自学．

本书可供计算机信息处理、自动控制、地球物理、生物医学工程等领域中从事模式识别工作的科技人员和高校师生参考．

模 式 识 别

边肇祺等 编著

责任编辑 蔡鸿程

☆

清华大学出版社出版

北京 清华园

中国科学院印刷厂印刷

新华书店北京发行所发行

☆

开本：787×1092 1/16 印张：19 字数：485千字

1988年6月第1版 1988年6月第1次印刷

印数：0001～10000 定价：3.55元

ISBN 7-302-00203-7/O·40(课)

《信息、控制与系统》系列教材

出 版 说 明

《信息、控制与系统》系列教材是一套关于信息、控制和系统学科的基本理论和应用技术的高等学校教材. 选题范围包括信号和信息处理、模式识别、知识工程、控制理论、自动化技术、传感技术、自动化仪表、系统理论、系统工程、机器人控制、智能控制、计算机应用和控制等方面. 主要读者对象为自动控制、计算机、过程自动化、无线电等系科的高年级大学生和研究生,以及在这些领域和部门工作的科学工作者和工程技术人员.

信息、控制与系统科学是在本世纪上半叶形成和发展起来的新兴学科. 它们的应用和影响已经遍及众多的部门和领域, 贯穿其中的许多思想和方法已用于经济和社会现象的研究, 而以这些学科为理论基础的自动化技术的广泛应用更是实现现代化的重要标志之一. 这套系列教材,正是在这样的客观要求下,为适应教学和科研工作的需要而组织编写和出版的. 它以清华大学自动化系近年来经过教学实践的新编教材为主, 力求反映这些学科的基本理论和最新进展, 并且反映清华大学在这些学科领域中科学研究和教学研究的成果. 我们希望这套系列教材, 既能为在校大学生和研究生的学习提供较为系统的教科书,也能为广大科技人员提供有价值的参考书.

编辑和出版这套教材是一次尝试. 我们热忱欢迎选用本系列教材的老师、学生和科技工作者提出批评、建议.

《信息、控制与系统》系列教材编委会

一九八七年三月

前　言

这本教材是清华大学自动化系信息处理和模式识别教研组在近几年给大学生和研究生讲授“模式识别”课程的基础上，参考了国外同类教材及有关文献编写而成，重点讨论统计识别方法．为了使读者能够对模式识别的应用有一定的了解，本书最后两章讨论了模式识别在一维数字波形和二维数字图象中的应用．这一部分内容在很大程度上涉及到数字信号处理和数字图象处理中的一些方法，因此只是为了给有兴趣的读者作参考，讲授时完全可以略去而不影响对全书的理解．

本书是在常迥教授的热情支持和帮助下写成的，参加编写的有边肇祺、阎平凡、杨存荣、高林、刘松盛和汤之永等同志．边肇祺、阎平凡和杨存荣对全书原稿进行了大量修改和校正．我们在编写过程中反复进行讨论，力求把这本教材写好，但仍会有错漏之处．希望读者能及时把发现的问题告诉我们，并希望为今后进一步提高本书的质量提出宝贵意见．

目　　录

第一章 绪 论

模式识别是六十年代初迅速发展起来的一门学科. 它所研究的理论和方法在很多科学和技术领域中得到了广泛的重视，推动了人工智能系统的发展，扩大了计算机应用的可能性. 二十多年来,取得了大量研究成果,在很多地方得到了成功的应用. 但由于问题的复杂性，现有的理论和方法离开要求还有一段距离. 为了帮助读者比较容易地掌握后面各章所述的内容，本章主要讨论模式识别的一些基本概念和问题，以利于对模式识别问题的本质有所了解.

1.1 模式识别和模式的概念

人们在生产和生活等活动中总是不断地在进行模式识别. 上班坐汽车,要找汽车站. 医生看病，通过问诊化验作出病情诊断. 这些都是在进行模式识别. 本书讨论的模式识别是指用计算机的方法来实现人的模式识别能力,更具体地说,就是实现人对各种事物或现象的分析、描述、判断和识别.

人们在观察各种事物或接受各种客观现象的时候，常常把它们分成由各个相似的但又不完全相同的事物或现象组成的类别. 在同一类别中的事物或现象尽管可能是不全一样的,但它们总在某些方面具有相似之处. 例如数字"4"可以有各种各样的写法，但它们都是属于同一类别. 更为重要的是,即使对于某种写法的"4",人们过去从未见过,也很容易把它分到数字"4"这一类别中去. 人脑的这种思维能力就构成了"模式"的概念. 在这里,模式和集合的概念是分不开的,只要认识这个集合中的有限数量的事物或现象，就可以识别属于这个集合的任意多的事物或现象. 为了强调从一些个别的事物或现象推断出事物或现象的总体，我们把这样一些个别的事物或现象叫作模式. 也有的作者认为应该把整个的类别叫作模式,因为它是一个抽象概念，如"汽车"、"河流"、"肺炎"、"癌细胞"等都属于模式的范畴,而把具体的对象,如停在汽车场上的号码是 31-1245 的那辆汽车叫样本，即把从模式组成模式类的过程叫作从样本组成模式的过程. 这种名词上的不同含义是很容易从上下文上弄清楚的.

模式类和模式就相当于集合论中的子集和元素. 在一个集合M中可以定义一个关系R. 我们知道,假使对所有的$x \in M$，xRx 存在,则称这个关系是自返的,假使对于$x, y \in M$，只要 xRy 存在，yRx 也同时存在,则称这个关系是对称的,满足自返和对称的关系就是相似关系. 假使对于$x, y, z \in M$，只要 xRy，yRz 存在，xRz 也同时存在,则称这个关系是传递的,例如相等关系就是一种传递关系(当然也是自返的和对称的). 自返、对称和传递关系同时存在时称作等价关系. 满足等价关系的集合必定可以划分为若干子集，即 $M = \bigcup_i M_i$，且 $M_i \cap M_j = \phi (i \neq j)$. 在同一子集 M_i (或称等价类) 中的各个元素在一定意义上是不可区分的，因此一个子集就相当于一个模式类. 满足等价关系的各个模式

类间有明确的界限，或者说是可以区分的．但客观实际常常不能把所研究的对象划分为各个明确的类别，即在类别的邻接区域不满足等价关系，但一个模式类中的各模式一定存在着自返和对称关系．

1.2 模式识别系统

有两种基本的模式识别方法，即统计模式识别方法和结构(句法)模式识别方法，与此相应的模式识别系统都由两个过程所组成，即设计和实现．设计是指用一定数量的样本(叫作训练集或学习集)进行分类器的设计．实现是指用所设计的分类器对待识别的样本进行分类决策．本书只讨论统计模式识别方法．基于统计方法的模式识别系统主要由四个部分组成：数据获取，预处理，特征提取和选择，分类决策．具体参见图 1.1．

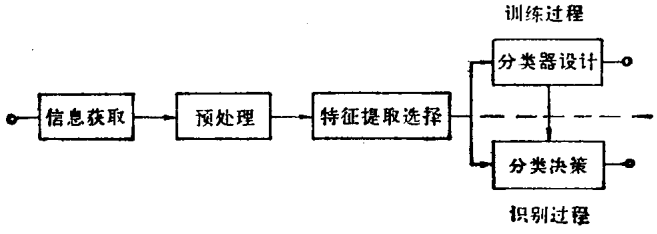

图 1.1

下面我们简单地对这几个部分作些说明．

一、数据获取

为了使计算机能够对各种现象进行分类识别，要用计算机可以运算的符号来表示所研究的对象．通常输入对象的信息有下列三种类型，即

1. 二维图象　如文字、指纹、地图、照片这类对象．

2. 一维波形　如脑电图、心电图、机械震动波形等．

3. 物理参量和逻辑值　如在疾病诊断中病人的体温，各种化验数据，或对症状有无的描述，如疼与不疼，可用逻辑值即 0 和 1 表示．

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．这就是数据获取的过程．

二、预处理

预处理的目的是去除噪声、加强有用的信息、并对输入测量仪器或其他因素所造成的退化现象进行复原．

三、特征提取和选择

由图象或波形所获得的数据量是相当大的，如一个文字图象可以有几千个数据，一个心电图波形也可能有几千个数据，一个卫星遥感图象的数据量就更大．为了有效地实现分类识别，就要对原始数据进行变换，得到最能反映分类本质的特征．这就是特征提取和选择的过程．一般我们把原始数据组成的空间叫测量空间，把分类识别赖以进行的空间

叫特征空间，通过变换，可把在维数较高的测量空间中表示的模式变为在维数较低的特征空间中表示的模式.

四、分类决策

分类决策就是在特征空间中用统计方法把被识别对象归为某一类别. 基本作法是在样本训练集基础上确定某个判决规则，使按这种判决规则对被识别对象进行分类所造成的错误识别率最小或引起的损失最小. 本书只讨论第三和第四部分的理论基础和方法，第一第二部分是数字信号处理和图象处理等课程的研究课题.

1.3 关于模式识别方法的一些问题

这一节里，我们讨论有关模式识别方法的一些有兴趣的问题，这有助于对模式识别进一步了解.

一、学习

为了能够对模式分类，往往需要先进行学习. 一个孩子学习认字时，教师首先让他看一个个字符并同时告诉他字符的名称. 即使没有教师，这个孩子一遍遍地观察各个字符，也会把它们分成相应的类别，除了名称不确定以外，分类的结果与有教师的情况是一致的. 这也是一种学习过程，是一种自学习过程. 广义地说，模式识别的学习问题就是研究如何用机器实现人脑的这种学习能力的一个控制论中的问题. 图 1.2 上有 12 个两类分类问题，其中每一个问题都有六个样本，位于同一列上的三个样本属于同一类. 读者很容易找到区分这两类的特征，并且能画出属于某一类的无穷多个其他样本，说明人对于这类问题有很强的学习能力. 但机器是否能作到这一点？机器接近人脑学习能力的可能性有多大？机器是否有可能自己发展这种学习能力？假使我们能够精确地描述我们进行智力活动的过程，或者说能够精确地表述出我们对所观察的事物或现象所作出的反应，那么原则上说是可以把这种能力赋予机器的. 可惜我们现在不能作到这一点，这样我们就只能通过把样本输入机器，使它按照某种给定的学习方法学会模式识别了. 另一方面，由于人的感受系统的感受能力（例如对声音和图象的接受频带）以及在计算速度、高维空间结构的分析能力等

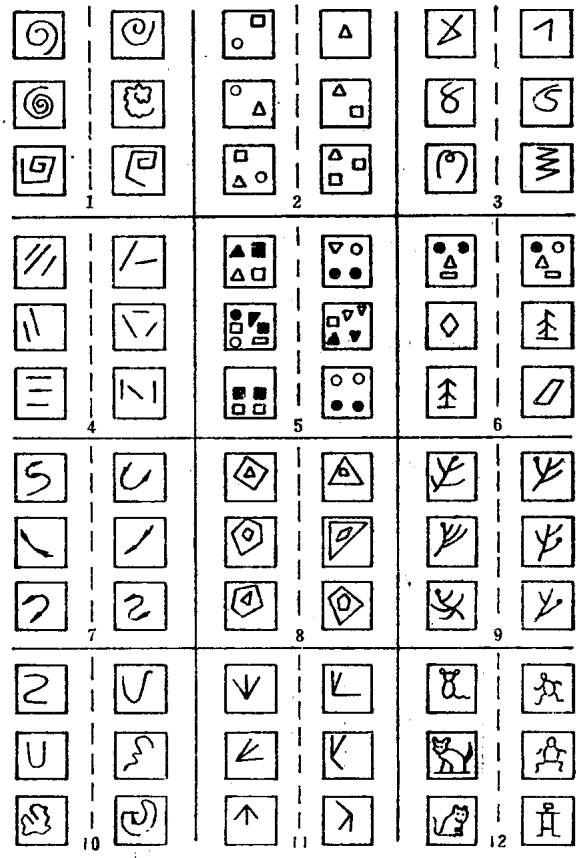

图 1.2

等方面都有很大局限性，因此用机器进行学习和识别在某些方面又可能达到人所不能达到的更高水平．下面我们简单地讨论一下一种经常在实际中得到应用的学习方法．

任何一种学习都要有一个学习目的，并以此为根据设计学习系统的学习过程．按照具体问题的要求，我们可以规定把学习系统的某个性能指标达到最优作为学习的目的．在处理和学习样本数据的基础上，通过改变学习系统的参数或者改变学习系统的结构来达到性能指标的最优化．例如我们要用木材的亮度 b 对两种木材——椅木和桦木进行分类．我们要设计这样一种学习方法，即根据已知类别的一定数量的木材，通过学习求得某个阈值 t，使当木材的亮度大于 t 时，就把它归为桦木，反之则归为椅木．阈值 t 就是这个学习系统的参数．为要求得这个参数，首先要规定学习系统的某个性能指标，即准则函数．在这个例子中，可以用对训练集样本的错分率作为准则函数．准则函数一般可以表示为

$$J(\boldsymbol{c}) = E[R(F(\boldsymbol{x},\boldsymbol{c}))]$$

式中 $\boldsymbol{x}$ 是随机向量，例如木材的亮度 b．$\boldsymbol{c}$ 是学习系统的参数，例如阈值 t．$R(\cdot)$ 是损失函数，对于上面所说的例子，假设阈值为 t，若分类正确，即对于桦木样本 $b > t$，或对椅木样本 $b < t$，则损失函数值定义为 0，若分类错误，则定义为 1．显然损失函数和 $\boldsymbol{x}$、$\boldsymbol{c}$ 以及所采用的决策规则 F 有关．E 是数学期望算子，在训练样本数有限的情况下，就是计算平均损失值．向量 $\boldsymbol{x}$ 和向量 $\boldsymbol{c}$ 之间可能由某些方程式或某些附加的约束条件联系起来．上述准则也可表示为

$$J(\boldsymbol{c}) = \int_{\boldsymbol{x}} p(\boldsymbol{x})\, R(F(\boldsymbol{x}, \boldsymbol{c}))d\boldsymbol{x}$$

式中 $p(\boldsymbol{x})$ 是向量 $\boldsymbol{x}$ 的概率密度函数．假使我们完全掌握了 $p(\boldsymbol{x})$，那么就可以用统计决策理论的方法来求取准则最优时的最佳参数 $\boldsymbol{c}^*$，而不再需要进行任何学习．若 $p(\boldsymbol{x})$ 存在但未知，为要应用统计决策理论，就需要对观察到的样本（数据）来对概率密度函数作出估计．对于大多数模式识别应用来说，学习的目的是对训练集样本的平均损失最小．从这个概念出发，发展了一系列可训练分类器的设计方法．

二、模式的紧致性

为了能在某个空间中进行分类，通常假设同一类的各个模式在该空间中组成一个紧致集．从这个紧致集中的任何一点可以均匀地过渡到同一集中的另外一点，而在过渡途中的所有各点都仍然属于这个紧致集即属于同一模式类．此外当紧致集中各点在任意方向有某些不大的移动时（相应于被观察现象有某些微小的变形）它仍然属于这个集合．为说明上述假设的意义，我们比较详细研究一下样本紧致性的概念．假定如图 1.3 所示，在三维空间中给定具有坐标 000，001，010，011，100，101，110，111的点集．我们希望用平面把它分成两个集合 A_1 和 A_2．显然解决这个问题的复杂性和这两个集合的组成情况有关．例如假使 A_1 由点 111，101，110，011 组成而 A_2 由点 000，010，100，001 组成，则只要用一个平面就能把这两个点集分开．假使 A_1 由一个点 000 组成，A_2 由一个点 111 组成，那么把它分开就更容易了．任何一个通过点 000 和点 111 连线的平面都能达到这

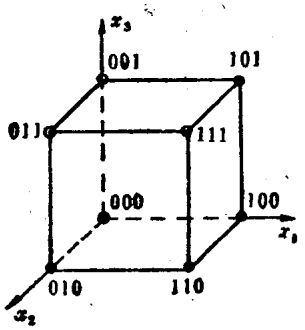

图 1.3

个目的. 但假使要分开 $A_1=\{111, 001, 100, 010\}$ 和 $A_2=\{000, 011, 101, 110\}$ 则需要用三个平面. 对于这种情况,集合 A_1 中任一个点的一个码的数值发生变化, 例如从 111 变成 101, 它就成为集合 A_2 中的成员. 对于集合 A_2 也是这样. 我们把这样的一些点叫作临界点, 而把那些改变其中任一个码值而不造成它所属集合的改变的点叫作集合的内点. 一般情况下,两个集合中具有的临界点愈多,则把它们分开就愈困难.对于 $A_1=\{000\}$, $A_2=\{111\}$ 情况.由于没有临界点,所以很容易把它们分开,而对于 $A_1=\{111, 101, 110, 011\}$, $A_2=\{000, 010, 100, 001\}$ 来说, 各有三个临界点和一个内点,就比较难于把它们分开了.

一般来说,在 D 维空间中要用超表面进行分类,假使我们用平面图来表示 D 维空间中点的分布,通常有图 1.4 所示三种情况,其中 (a) 没有临界点, (b) 有许多临界点, (c) 的临界点已经多到使分类不可能实现.

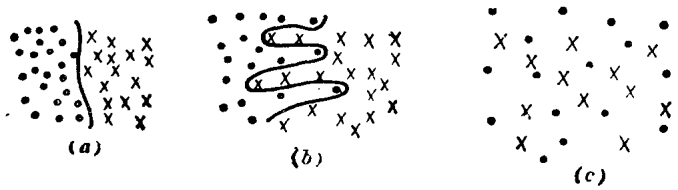

图 1.4

从上面讨论中, 我们可以看到紧致集具有下列性质: (一)临界点的数量与总的点数相比很少. (二)集合中任意两个内点可以用光滑线连接, 在该连线上的点也属于这个集合. (三)每个内点都有一个足够大的邻域,在该邻域中只包含同一集合中的点.

假使每个模式类都满足紧致性假设, 则解决模式识别问题就不会碰到什么原则上的困难. 但对于很多实际问题这个假设是不成立的. 例如图 1.2 所示的任何一个分类问题, 至少在传感器 (例如摄象机) 输出的测量空间或其他原始描述上就完全不满足紧致性要求. 另一方面,假使图 1.2 所给的识别任务已经很好解决, 我们就可以设计出一种变换方法, 使测量空间上属于同一类的所有各点都映射到特征空间上的同一点, 而把另一类的所有各点都映射到特征空间上的另外一点,且这两点使它相隔一个显著的距离. 显然,在这个特征空间中,模式类是满足紧致性假设的. 因此, 我们可以说,只要各个模式类是可分的,总存在这样一个空间,使变换到这个空间中的集合是满足紧致性要求的. 这样一种变换只能在解决识别任务的过程中来求取,是和具体问题紧密相关的,我们现在还没有统一的解决这种变换的有效理论和方法.

三、距离和相似性度量

虽然紧致性假设并没有提供解决所有模式识别问题的一般方法, 但它仍然是我们研究许多模式识别的特征提取和分类方法的基础. 紧致性本身要求在相应的空间定义距离度量,对模式来说,就是要求定义它们之间的相似性度量. 人们在日常生活中进行识别时也总是利用相似性概念,但是人们又很难对"相似"或"不相似"作出明确的定量表述,因此通常所说的相似性只具有定性的或不确定的性质. 怎样对相似性概念给以明确的定量表述是模式识别工作者要解决的任务之一.

给定一个输入样本集合 $\mathscr{X}$. 我们用 D 维空间中的一个点表示某个样本. 两个样本

$\boldsymbol{x}_k$ 和 $\boldsymbol{x}_j$ 之间的相似性度量 $\delta(\boldsymbol{x}_k, \boldsymbol{x}_j)$ 应满足下列要求：

(1) 相似性度量应是非负值，即 $\delta(\boldsymbol{x}_k, \boldsymbol{x}_j) \geqslant 0$.

(2) 样本本身之间的相似性度量应为最大.

(3) 相似性度量应满足对称性，即

$$\delta(\boldsymbol{x}_k, \boldsymbol{x}_j) = \delta(\boldsymbol{x}_j, \boldsymbol{x}_k)$$

(4) 在模式类满足紧致性要求下，相似性度量应是点间距离的单调函数.

在各种空间中，只要定义任何一种距离度量，就可以用这种距离度量的非增函数作为相似性度量. 例如在 D 维欧几里得空间，可以选择某个随距离增加而下降的函数 f 作为相似性度量，即

$$\delta(\boldsymbol{x}_k, \boldsymbol{x}_j) = f\left(\sqrt{\sum_{i=1}^{D}(x_{ki} - x_{ji})^2}\right)$$

也可以用其他的距离度量，如

$$\delta(\boldsymbol{x}_k, \boldsymbol{x}_j) = f\left(\sum_{i=1}^{D}|x_{ki} - x_{ji}|\right)$$

有些情况下，可以采用两个向量之间的夹角来度量相似性，例如

$$\delta(\boldsymbol{x}_k, \boldsymbol{x}_j) = \cos^{-1}\frac{\boldsymbol{x}_k^T \boldsymbol{x}_j}{\|\boldsymbol{x}_k\|\|\boldsymbol{x}_j\|}$$

上面各式中，$\boldsymbol{x}_k$, $\boldsymbol{x}_j$ 是表示模式 k 和模式 j 的向量，x_{ki}, x_{ji} 是它们相应的第 i 个分量.

四、识别的可靠性和品质

识别的可靠性是指具有某个识别品质的概率. 通常把系统的误识率作为识别系统的品质度量. 我们可以计算出识别结果和教师答案不同的那些识错的样本数，但若以它为基础来估计误识率时，显然误识率的大小和考试集(已知类别的专门用来估计误识率的样本集合)的组成有关. 我们可以假定误识率是对所有会出现的样本来计算的，但在解决实际任务的时候，通常只要求对经常出现的那些样本而不是一切样本能够进行正确识别，因此识别品质的估计应该和样本的概率密度函数 $p(\boldsymbol{x})$ 有关. 此外由于不同的错误类型会引起不同的后果，即造成不同的损失，所以识别品质的度量也可以用平均风险来度量. 为了对识别品质进行估计，所用的试验样本序列（考试集）应该是按照 $p(\boldsymbol{x})$ 随机地和独立地得到的. 在模式识别的学习阶段，用来进行学习的样本也应该是按照 $p(\boldsymbol{x})$ 随机地和独立地产生的. 但是当样本的数量不足够大，就不能保证训练集以及考试集能充分反映出 $p(\boldsymbol{x})$ 的情况. 因此我们只能在概率的意义上来讨论学习的成功与否，即模式识别系统的识别品质，或者说模式识别学习任务解决好坏的一个指标是模式识别系统具有等于或大于给定误识率的概率有多大. 设计一个模式识别系统就是要用学习样本序列通过学习过程求得一个决策规则，使该系统在规定可靠性的条件下具有不低于某一给定的识别品质.

1.4 关于本书的内容安排

正如前面已经提到的那样，本书只讨论统计模式识别方法中的特征提取和分类决策. 从教学角度出发，首先研究各种分类器的设计方法是可取的，在这个基础上就更容易理解