模式识别导论

齐敏 李大健 郝重阳 编著

清华大学出版社

模式识别导论

齐敏 李大健 郝重阳 编著

清华大学出版社
北京

内 容 简 介

本书按照统计模式识别、句法模式识别、模糊模式识别法和神经网络模式识别法四大理论体系组织全书，其中统计模式识别是模式识别的经典内容和基础知识，模糊模式识别法和神经网络模式识别法两部分反映了模式识别学科发展的新进展，附录部分归纳了书中需要用到的概率知识、向量和矩阵运算的常用公式，以及供上机练习用的模式样本数据。

本书内容由浅入深，便于教师根据不同情况选择教学内容。同时讲解详细，配有丰富的图表和例题，有助于读者阅读与理解。提供了习题和计算机作业，供学习时使用。

本书可作为高等院校电子信息类专业高年级本科生和研究生的教材，也可供从事模式识别工作的广大科技人员参考。

图书在版编目（CIP）数据

模式识别导论/齐敏，李大健，郝重阳编著．—北京：清华大学出版社，2009.6
ISBN 978-7-302-20066-6

Ⅰ．模…　Ⅱ．①齐…　②李…　③郝……　Ⅲ．模式识别　Ⅳ．O235

中国版本图书馆 CIP 数据核字（2009）第 062498 号

责任编辑：袁勤勇　李玮琪
责任校对：焦丽丽
责任印制：孟凡玉

出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　**邮　　编**：100084
社　总　机：010-62770175　　**邮　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：北京密云胶印厂
装 订 者：北京市密云县京文制本装订厂
经　　销：全国新华书店
开　　本：185×260　　**印　　张**：16.75　　**字　　数**：405 千字
版　　次：2009 年 6 月第 1 版　　**印　　次**：2009 年 6 月第 1 次印刷
印　　数：1～3000
定　　价：25.00 元

前言

模式识别是一门既具有较系统的理论体系，又仍处在迅速发展之中的边缘学科，其应用几乎遍及各个领域。

这一学科涉及许多较为深奥的数学理论，对刚涉足这一领域的许多初学者来说，理解起来有一定的困难。本书既可以作为基础教材，又反映学科发展方向，以此为基调，在选材上立足于“精”，在讲解上立足于“透”。笔者结合多年教学经验，在模式识别理论的成熟部分，注重对基本概念的透彻讲解，选择经典和实用的模式识别方法和算法进行讨论。在内容的安排上，注意由浅入深，讨论问题时尽量减少数学推导和证明，通过实际运用加深学生对算法的理解，目的是使初学者能够比较容易地尽快掌握模式识别的基本理论和方法。

在材料组织上，兼顾计算机模式识别的基础和发展两个方面，按照统计模式识别、句法模式识别、模糊模式识别法和神经网络模式识别法四大理论体系组织全书内容。其中统计模式识别方法是核心，模糊模式识别和神经网络模式识别是学科的新发展。在具体章节安排上，统计模式识别部分包括属于非监督分类的聚类分析方法（第 2 章）、监督分类中的判别函数概念和几何分类法（第 3 章）以及基于统计决策的概率分类法（第 4 章），为简化分类器还讨论了特征选择和提取的方法（第 5 章）。第 6 章为句法模式识别，对概念和方法进行了简要讨论。第 7 章为模糊模式识别部分，鉴于本方向蓬勃的发展趋势，对其内容进行了充实和细化，充分地阐述了基本概念，详细讨论了其中的典型算法。第 8 章为神经网络模式识别法，介绍了几种典型的用于模式识别的神经网络模型和算法。为方便学习，在附录部分增加了必要的相关知识介绍，以备查阅。同时书中还配有丰富的图表，有助于阅读。

本书是作者在多年教学实践的基础上经过总结扩充改编而成的。参加编写的有齐敏、李大健和郝重阳同志，全书由齐敏同志负责审订和修改。模式识别学科发展迅速，本书对该学科研究发展的新进展亦有所涉及，限于水平和经验，书中错误及不当之处在所难免，敬请读者批评指正。

编　者

2009 年 2 月

目录

目录

目录

目录

目录

第1章 绪 论

1.1 模式和模式识别的概念

从广义方面讲，模式(pattern)是一个客观事物的描述，即一个可用来仿效的完善的例子。模式识别(pattern recognition)按照哲学的定义，是指一个“外部信息到达感觉器官并被转换成有意义的感觉经验”的过程。

例如，桌上的玻璃杯里装着某种物质，人们对它进行仔细观察，在这个过程中，眼睛、鼻子、皮肤等不同的感觉器官接收到一些来自这个物体的所谓的外部信息：无色、透明、液体、冒气、无臭、温度较高，这些感觉信息被送到大脑后，经过处理，转换成了感觉经验——热水，这实际上就是一个模式识别的过程。

人是一个深不可测的信息处理系统，具有超级模式识别能力。事实上，每个人每天都在进行模式识别。例如，一个人到一个新的城市里去找公共汽车站，就是在做模式识别。再例如，在一群嘈杂的人群中，我们能够区别出熟悉的朋友的声音；我们还能够认识不同的人书写的“不是很潦草”的字符；等等。这些其实都是模式识别过程。不同的人或同一个人在不同的时间写出的字是不完全相同的，有时还会有很大差别，但我们能够识别，这是因为在人的头脑中有这样一个仿制的模型，这就是模式。模式是由大量的取样、学习、归纳而成的，人们将所看到的信息与此模式比较，从而判断此信息是否属于该类模式。因此，模式识别问题通常表现为对一组过程或事件的判别或分类(pattern classification)。人类具有的模式识别功能可否由机器来实现呢？这正是本书所要研究的内容。

根据人类的识别能力所涉及的被识别客体的性质，可以将识别活动的对象分为两个主要类型：具体的客体和抽象的客体。具体的客体如字符、图画等，通过对感官的刺激而被识别；论点、思想等则是非物质的抽象客体，不属于本书研究的范畴。我们主要是研究对具体客体的识别，而且仅局限于研究用机器完成与识别任务有关的基本理论与实用技术。

针对所要研究的内容，可以对模式和模式识别做如下狭义的定义：模式是对某些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。模式识别是研究一种自动技术，依靠这种技术，计算机将自动(或人尽少干涉)地把待识别模式分配到各自的模式类中去。

注意，狭义的“模式”概念是指对客体的描述，不论是待识别客体，还是已知类别的客体。而在广义的“模式”定义中，模式指的是“用于效仿的完善例子”。两者所表述的范围是不同的，但无论是广义的模式还是狭义的模式，都是对事物的一种描述，也就是说，模式指的并不是事物本身，而是我们从事物获得的信息。另外，也有人习惯将模式类称为模式，而把个别具体的模式称为样本，这种用词上的不同可以从上下文区分其含义，一般不会引起误解。

模式识别是伴随着计算机的研究和应用日益发展起来的。随着计算机应用领域的不断扩大，人们将计算机称为“电脑”，几乎所有本来由人脑实现的功能，人们都试图用

“电脑”来完成。虽然在这方面已经取得了令人振奋的成就，但比起人脑来，电脑毕竟是小巫见大巫了。人脑具有极丰富的联想能力，而反观电脑，除了在联想、判断、推理能力等方面远远不及人脑外，在对外界信息的感知方面，其能力更是远不如人脑。第一次人工智能(artificial intelligence)研讨会于1956年夏天在美国召开。在早期的一次人工智能国际会议上，日本学者曾展出了一个脸谱识别器，演示时能把几个日本人的脸谱识别出来，并叫出他们的名字。有些其他国家的代表也想试试，但计算机识别不出来，显示“不是人”，引起哄堂大笑。从那次会议到现在已过去很久了，尽管可以将“不是人”改为“不认识”，但在这方面至今尚无质的飞跃。

目前，计算机处理能力的发展非常迅速，相比之下，计算机与外部的信息交换能力却低得可怜。对当前建立在冯·诺依曼体系基础上的计算机来说，其模仿人脑的识别、思维过程的能力很难有一个大的进步，因此，日本提出了第五代计算机的研究计划。

冯·诺依曼(Von Neumann)是美籍匈牙利数学家，在1946年提出了关于计算机组成和工作方式的基本设想，理论要点是：数字计算机的数制采用二进制；计算机应该按照程序顺序执行，即“程序存储”的概念。1949年研制出了第一台冯·诺依曼式计算机。到现在为止，尽管计算机制造技术已经发生了极大的变化，但是就其体系结构而言，仍然是根据冯·诺依曼的设计思想制造的，因此现在的计算机仍称为冯·诺依曼结构计算机。

第五代计算机以突破冯·诺依曼体系为前提，它与前四代计算机的本质区别是：计算机的主要功能将从信息处理上升为知识处理，使计算机具有人类的某些智能，所以第五代计算机又称为人工智能计算机。从20世纪80年代开始，日本、美国和欧洲各国纷纷进行第五代计算机的研制工作，但目前尚未形成一致结论，仍在研究当中。通常认为，第五代计算机具有以下几个方面的功能：

(1) 具有处理各种信息的能力。第五代计算机除了能像目前的计算机一样处理离散数据外，还能对声音、文字和图像等形式的信息进行识别处理。

(2) 具有学习、联想、推理和解释问题的能力。

(3) 具有对人的自然语言的理解能力。人们只需针对要处理或计算的问题，用自然语言写出要求及说明，第五代计算机就能理解其意，并按人的要求进行处理或计算，而不需要像现在这样，使用专门的计算机语言把处理过程与数据描述出来。对第五代计算机来说，只需告诉它“做什么”，而不必告诉它“怎么做”。

第五代计算机研制计划的实施促成了人们的共识，即继续展开人工智能各个领域的研究，使计算机能够更好地为人类服务。而模式识别就是其中的一个重要方面。

总之，研究和发展模式识别的目的在于提高计算机的感知能力，从而大大开拓计算机的应用范围。当然，计算机感知能力的真正提高，不仅与模式识别这一学科本身有关，而且与概率论、线性代数、模糊数学、形式语言、离散数学、工程技术学以及计算机本身的体系结构和软硬件性能等均有关系。正在研究的第五代计算机的发展方向有以下几种可能：

(1) 神经网络计算机。模拟人的大脑思维。

(2) 生物计算机。运用生物工程技术，采用蛋白分子作为芯片。

(3) 光计算机。用光作为信息载体,通过对光进行处理来完成对信息的处理过程。

因此,需要及时把握计算机科学和其他相关学科的新进展,以便对模式识别的发展状况有一个全面的了解,进而促进新理论和新方法的研究。

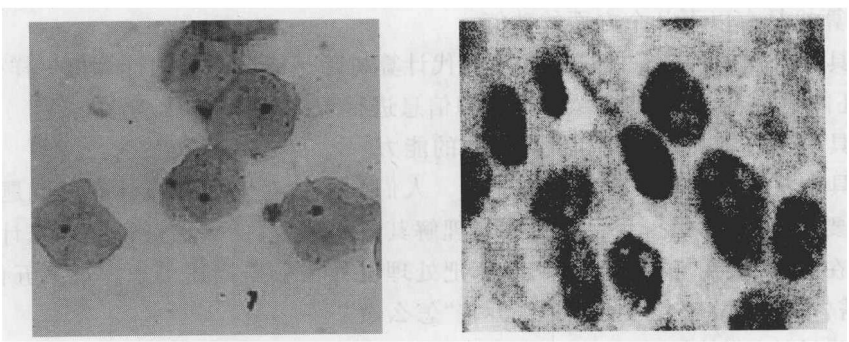
1.2 模式识别系统

1.2.1 简例

在讨论一般的模式识别系统之前,先以癌细胞识别为例子,来了解机器识别的全过程,以获得一个感性认识,从而建立模式识别的基本概念。

1. 信息输入与数据获取

首先,利用巴氏染色将从病人体内取出的待检物制成细胞涂片。然后,使用摄像头在显微镜目镜处进行拍摄,以便采集静态图像和动态的轨迹图像。摄像头连接图像采集卡,图像采集卡将接收到的模拟信号数字化后传入计算机,将图像存储在大容量的硬盘上。这样,显微细胞图像就转换成了数字化细胞图像,以满足计算机分析处理数字信息的要求。这个获得数据的过程实际上是一个抽样与量化的环节。这一过程也可以直接采用有足够高分辨率的数码摄像机或数码相机拍摄完成,只需将数码设备通过具有信号传输功能的器件与计算机连接,即可将拍摄到的图像传入计算机。图 1.1 所示为两个数字化的显微细胞图像。



彩色图像　　灰度图像

图 1.1　数字化显微细胞图像

通常所获取的医学细胞图像是经过染色处理的彩色图像,而计算机在进行图像处理过程中往往需要灰度图像。灰度图像是指只含亮度信息,不含色彩信息的图像,其中像素的亮度值可以取 0～255,共分为 256 个级别,分别反映原细胞图像中相应位置的光密度大小。灰度和 RBG 颜色的对应关系为:亮度 $Y=0.299R+0.587G+0.114B$。实际操作中,应根据具体情况和要求选择采用哪种图像。

数字化细胞图像是计算机进行分析的原始数据基础。

2. 数字化细胞图像的预处理与区域划分

数字图像预处理的目的在于：

① 去除在数据获取时引入的噪声与干扰；

② 去除所有夹杂在背景上的次要图像，以便突出主要的待识别的细胞图像，供计算机进行分析时使用。预处理过程中采用的是“平滑”、“边界增强”等数字图像处理技术。

区域划分的目的在于找出边界，划分出不同区域，为特征抽取做准备。换句话说就是要检测出细胞与背景、胞核与胞浆之间的两条边界线，从而将三个区域分割开来，以便进行细胞特征的抽取。这里将数字化图像划分为背景 B、胞浆 C、胞核 N 三个区域，如图 1.2 所示。

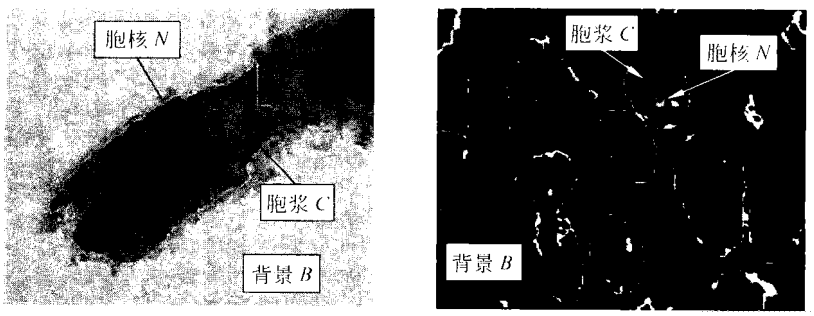


图 1.2 区域划分

在数字图像处理技术中，“区域划分”(或称“区域分割”、“边界检测”)的方法很多，在其相关课程中有专门的讨论。这里假设采用某种方法获得了如图 1.2 所示的结果，其中图 1.2(a)所示为疑似肿瘤细胞图像。

3. 细胞特征的抽取、选择和提取

细胞特征的抽取、选择和提取的目的是为了建立各种特征的数学模型，以利于分类，基本思路如图 1.3 所示。

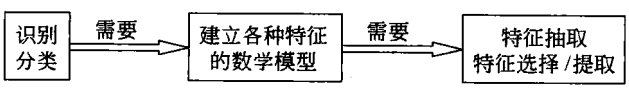


图 1.3 特征抽取、选择和提取的目的

抽取、选择和提取三个概念的含义是有差别的。特征抽取是指对数据的最初采集，细胞特征的抽取是识别分类的依据。处于人体不同部位或病变不同阶段的细胞反映出不同的形状和结构特征，因此在识别分类之前必须首先建立各种特征的数学模型，尽可能多地抽取特征，以供计算机进行定量分析时使用。这里我们共抽取 33 个特征：胞核面

积、胞核面积占整个细胞面积的百分比、胞核的总密度、胞浆面积、胞浆的平均光密度……通过特征抽取，可以建立一个 33 维的空间 X，每一维表示一个特征，每一个细胞在每一个特征（每一维）上都有一个度量值 x_i，因为每个细胞的度量值不同，所以这个值是一个随机变量。由于一个细胞可以通过 33 个特征表示，即由 33 个随机变量 x_i 表示，因此每个细胞可以用一个 33 维的随机向量表示，记为 $\boldsymbol{X}=[x_1,x_2,\cdots,x_{33}]^{\mathrm{T}}$，上标 T 是转置符号。这样，就完成了统计模式识别的第一项重要工作，即把一个物理实体“细胞”变成了一个数学模型“33 维的随机向量”，也即 33 维空间中的一点。

通常通过特征抽取所得到的原始特征数较多，如果使用全部可测量到的特征去判别分类，会因为判别空间维数太高而使问题变得很复杂。而事实上，由于某些特征之间往往存在一定的相关性，因此有必要也有可能在原始特征数据的基础上选择一些主要特征作为用于判别的特征，这就是特征选择。有时是采用某种变换技术，得出数目上比原来少的综合性特征用于分类，这称为特征维数压缩，习惯上称为特征提取。

例如，有五个特征 x_1,x_2,x_3,x_4,x_5，以及变换 $f(\cdot)$ 和 $g(\cdot)$，则可有

$$y_1=f(x_1,x_2,x_3,x_4,x_5),\quad y_2=g(x_1,x_2,x_3,x_4,x_5)$$

结果，X 空间中的特征向量 $\boldsymbol{X}=[x_1,x_2,x_3,x_4,x_5]^{\mathrm{T}}$ 变成 Y 空间中的特征向量 $\boldsymbol{Y}=[y_1,y_2]^{\mathrm{T}}$，也就是说通过特征提取，降低了原始空间的维数，特征向量从五维降成了二维。

特征向量组成的空间是识别分类赖以进行的空间，称为特征空间，本书中用大写斜体字母表示。特征向量就是特征空间中的一点，本书中用大写斜体加粗字母表示，其分量用相应的小写斜体字母带下标表示。一个特征向量代表一个研究对象，人们通常所称的模式、样本或模式样本等，实际上就是对特征向量而言的。

可以看出，通过特征抽取取得的数据是原始的第一手资料，是进行特征选择或特征提取的依据。如何进行特征选择或特征提取是模式识别研究的主要课题之一，也是非常重要的一个方面。特征选择或特征提取的方法对后续识别分类方法的选择以及分类效果都有很大的影响。

4. 判别分类

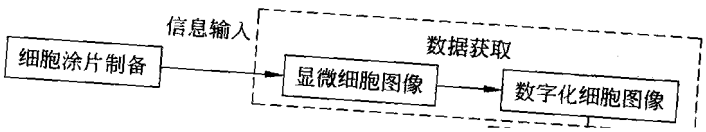
判别分类是模式识别研究的另一个主要内容，有多种多样的理论和方法，本书将在后续章节中着重讨论。这里我们假定针对癌细胞的识别应用了某种方法，得到的结果为：

（1）气管细胞 97 个，识别错误率为 7.2%。

（2）肺细胞 166 个，识别错误率为 18%。

在完成识别的同时，提供了错误率，可见判别的好坏是通过错误率给出的。识别错误率包括将正常细胞误判为癌细胞和将癌细胞误判为正常细胞两种情况，两种错误的代价和风险是不同的。本书后面的内容还将对此进行较深入的分析。

整个识别过程如图 1.4 所示。这个例子虽很粗略，但它比较典型地给出了模式识别的一般步骤。


信息输入
细胞涂片制备
数据获取
显微细胞图像
数字化细胞图像

1.3 模式识别概况

1.3.1 模式识别发展简介

模式识别诞生于20世纪20年代，1929年G. Tauschek发明阅读机，能够阅读0～9的数字。30年代Fisher提出统计分类理论，奠定了统计模式识别的基础，在60—70年代统计模式识别得到快速发展，成为模式识别的主要理论。50年代Noam Chemsky提出形式语言理论，美籍华人付京荪提出句法模式识别。60年代L. A. Zadeh提出了模糊集理论，目前模糊模式识别理论已经得到了较广泛的应用。80年代Hopfield提出神经元网络模型理论，近年来人工神经元网络在模式识别和人工智能方面也得到了较为广泛的应用。90年代以后小样本学习理论、支持向量机(Support Vector Machine, SVM)也受到了很大的重视。

从上面这个简单的时间表中可以看出，模式识别基本上是20世纪五六十年代开始快速发展，20世纪70年代初奠定理论基础，从而建立了独立的学科体系。传统的用于模式识别的方法，局限于统计模式识别与句法模式两大类。随着模糊数学的迅速发展，传统模式识别方法也深入到了模式识别的许多环节，出现了模糊模式识别。接着又出现了基于神经元模型的人工神经网络模式识别方法。这四种方法共同构成支持模式识别学科的四大支柱。需要说明的是，模式识别是一门处于迅速发展中的学科，有生命力的新理论经过一段时间的研究发展到比较成熟的阶段后，必然会充实到模式识别的理论体系中，同时，发展缓慢的理论也会自然地被逐渐淘汰掉。

1.3.2 模式识别分类

1. 按理论分类

按照在模式的识别过程中所依据的理论方法的不同，可将模式识别分为统计模式识别、句法模式识别、模糊模式识别法和神经网络模式识别法。

1) 统计模式识别

统计模式识别是定量描述的识别方法。以模式集在特征空间中分布的类概率密度函数为基础，对总体特征进行研究，包括判别函数法和聚类分析法。对于分类结果的好坏，同样用概率统计中的概念进行评价，如距离方差等。

统计模式识别的历史最长，与其他几种理论相比发展得最为成熟，是模式分类的经典性和基础性技术，目前仍是模式识别的主要理论，也是本书介绍的主要内容。

2) 句法模式识别

句法模式识别也称结构模式识别，是根据识别对象的结构特征，以形式语言理论为基础的一种模式识别方法。其出发点是识别对象的结构描述和自然语言存在一定的对