国外计算机科学经典教材

Pattern Recognition

Concepts、Methods and Applications

模式识别

—原理、方法及应用

J.P.Marques de Sá 著

吴逸飞 译

清华大学出版社
http://www.tup.tsinghua.edu.cn

模式识别

—— 原理、方法及应用

J.P.Marques de sá 著

吴逸飞 译

清华大学出版社

(京) 新登字 158 号
北京市版权局著作权合同登记号：图字 01-2002-4079

内 容 简 介

本书对模式识别的原理和方法作了全面的阐述，并对其在现实生活中各个领域的具体应用加以例证。主要内容包括模式识别的基本概念、模式识别的主要研究方法，包括统计模式识别方法、神经网络方法和结构模式识别方法，同时介绍了该领域的最新研究方法和成果，如 VC 及 FS 维、支持向量机、Hopfield 网络中的松弛匹配等。

本书可作为计算机工程专业的教材，也可以为那些需要应用模式识别技术的专业人员提供指导和帮助。

图书在版编目(CIP)数据

模式识别——原理、方法及应用/(美)马奎斯德萨著；吴逸飞译.—北京：清华大学出版社，2002
书名原文：Pattern Recognition Concepts, Methods and Applications
ISBN 7-302-05994-2

Ⅰ.模… Ⅱ ①马…②吴… Ⅲ.模式识别 Ⅳ.O235
中国版本图书馆 CIP 数据核字(2002) 第 080193 号

出 版 者：清华大学出版社(北京清华大学学研大厦，邮编 100084)
http://www.tup.tsinghua.edu.cn
责任编辑：郭东青
印 刷 者：清华大学印刷厂
发 行 者：新华书店总店北京发行所
开　　本：787×960 1/16　**印张**：19.5　**字数**：391 千字
版　　次：2002 年 11 月第 1 版　2002 年 11 月第 1 次印刷
书　　号：ISBN 7-302-05994-2/TP・3575
印　　数：0001～4000
定　　价：45.00 元

前　言

模式识别中包含大量的方法，这些方法正在推动着不同领域内众多应用的发展。一般认为模式识别方法最大的实用性在于“智能”仿真，它在我们的日常生活中随处可见。例如，机器人辅助生产线、医疗诊断系统、经济发展预测系统、地球资源探测系统、卫星数据分析系统等都是它的应用领域。模式识别的普及促进了很多特定领域方法学的发展，丰富了与其他学科的联系。但由于理论分支太多，现在新的理论发展方向是把众多传统的模式识别方法结合在一起，这样，各种方法本身以及结合后的新方法都将得到更大的发展。

本书源于波尔图大学(Oporto University)电子和计算机工程系的一门关于模式识别的概论性课程，基于这门课的核心内容，本书给出了关于模式识别方法的一些通俗易懂、清晰连贯的讲解，并结合了现实生活中的一些例子和应用。本书主要面向工程和计算机科学领域学习模式识别课程的本科生和研究生。除了工程师和应用数学家外，本书也同样适用于其他领域的专业人员和研究人员，例如医师、生物学者、地质学者和经济学者，帮助他们学习和应用模式识别方法。本书给出了一些实际应用的示例，并在一定程度上提供了很多读者感兴趣的素材，尤其对于那些非专业技术人员的读者，当他们需要在自己的工作中应用模式识别方法，或者碰巧遇到涉及这个学科的项目时，本书可以对他们的具体应用提供一些帮助。

模式识别包含由特征和属性所描述的对象的数学模型，也涉及到一般意义上对象间的相似性的抽象概念。具体采用何种数学形式、模型和处理方法取决于所要解决问题的类型。从这个意义上讲，模式识别其实就是用数学解决实际问题。讲解模式识别时，如果得不到实际例子、应用的反馈和直观认识，其效果将是十分有限的。因此，读者可以通过一些数据练习书中介绍过的方法，或者简单验证一下讲解的实例。因此，从学习本书开始，读者就应该通过实际应用的指导来学习书中教授的方法，不需要做任何程序设计工作，而应集中精力学会如何才能正确应用所学到的概念。如有问题，可与 Tup.wk@263.net 联系。

本书的结构组织十分经典。第 1 章讲述模式识别的基本概念，包括 3 种主要方法(统计模式识别方法、神经网络方法、结构模式识别方法)和一些重要的实际问题。第 2 章通过特征空间的表示问题以及决策函数的概念来讨论模式识别。第 3 章讲述了一些数据聚类理论以及降维技巧。第 4 章讲解基于统计的模式识别方法，包括使用和不使用样本分布模

型两种情况。第 5 章讲解神经网络方法并给出了一些典型范例，在研究分类和回归问题时，都应特别注意网络的性能评价和复杂度问题。第 6 章讲述结构分析方法，包括句法和非句法的方法。本书附录 A、B 分别是对书中数据集和软件工具的说明。

每一章中或几章之间出现的多个主题间的联系，都会在适当的时候加以说明，包括近期的一些论题，如支持向量机、数据挖掘和在结构匹配中神经网络的用法等。同时，如果主题有很重要的实际意义，例如维度比率问题，则给出一个详细的说明，并附上近期研究成果作为参考。

本书讲述每一种模式识别方法时，都是首先解释所涉及到的概念，并配有简单的例子和适当的图解。在讲述概念和方法中涉及到数学问题时，为了不引起混乱，自始至终都使用统一的符号。一旦方法已经被充分、详细地描述过，就将它们应用于实际数据中，以便读者能在重要的实际问题中掌握它。

从第 2 章起，每一章的最后都配有一套练习题，大部分练习需要使用本书提供的数据集，并作为模式识别设计工作中典型的计算机实践。其余的练习用来拓宽读者对书中例子的认识，测试读者对于它们的理解水平。

为了能够充分理解所讲述的内容，读者需要一定的概率论、统计学、线性代数和离散数学的背景知识。特别是统计学，读者需要熟悉统计推论中的主要概念和方法。

每一章后面都有一个参考书目的列表，包括对于书中全部内容的理论支持，在某些情况下它还可以指导读者作进一步的阅读。同时还包括对于一些背景知识的介绍，例如统计学领域的一些原理等。

数据集和软件工具需要在 Windows (95 或更高版本)环境中运行。这些数据集和软件工具中的许多都是通过 Microsoft Excel 生成的，应该可以在任何一种 Windows 版本中运行。其他一些软件工具需要通过正常程序进行安装。在书的附录 B 中可以找到对于这些软件工具的说明。根据这些说明以及书中包含的示例，读者在使用它们时，应该不会遇到太大的困难。

目　录

第1章 基本概念

1.1 对象识别

生物每天都在进行各种情况下的对象识别——如寻找食物、迁移、辨别敌害、辨认配偶等等，这是生物与生俱来的应付周围环境所必需的能力，而且其识别效率很高。这里，识别对象被认为是一种广泛的认知能力。当然，它可能只是很简单的本能，如当微生物来到pH值不适合的环境中时就会逃走；也可能需要高等的能力，如一个人需要从橱柜下面倒数第二个抽屉里取出剪刀。

随着建立智能自动化系统的需要，模仿各种形式的对象识别能力的方法得到了发展，同时，也带动了工业及其他领域技术的发展主流。在这些系统中，对象被表示成适当的形式，以便对它们进行处理，这种表示形式就称为模式(pattern)。本书中，对象(样本)和模式这两个词语可通用，它们表示相近的意思。

模式识别是一门研究对象描述和分类方法的学科。从计算的早期起，人们就发现设计和执行算法来模仿人类对物体的描绘和分类能力是一项有趣而富有挑战性的任务。因此，和多种学科有着紧密联系的模式识别系统和技术是科学研究的热门领域，吸引了许多来自不同领域的专家。

模式识别系统和技术有着极其广泛的应用，我们仅仅列举一些例子，其中涉及到几个专业领域：

农业：

- 产量分析
- 土壤评估

天文学：

- 天文望远镜图像分析
- 自动光谱学

生物学：

- 自动细胞学
- 染色体特性研究

遗传研究

市政管理：

交通状况分析和控制

城市增长评估

经济：

股票交易预测

企业行为分析

工程：

加工产品缺陷检测

特征识别

语音识别

自动导航系统

污染分析

地理：

岩石分类

矿产资源评估

使用卫星图像分析地理资源

地震分析

医学：

心电图分析

脑电图分析

医疗图像分析

军事：

航空摄像分析

雷达和声纳信号检测与分类

自动目标识别

安全：

指纹鉴定

监视和警报系统

从以上例子中可以推断，能进行分析和识别的模式可以是信号(例如心电图信号)、图像(例如航空摄像)或者普通的数值(例如股票交易率)表。

1.2 模式相似度和模式识别任务

模式识别中的一个基本概念是相似度(similarity)，这和已知的其他一些学科方法都不相关。一般认为两个对象相似是因为它们具有相似的特征。相似度经常被描述成更加抽象的概念，它并不是在几个对象之间衡量，而是在一个对象和一个目标概念(concept)之间进行衡量。例如，我们辨别出一个对象是苹果，因为它的特征符合理想化的苹果的图像、概念或者说典型模式(prototype)，所以我们认为它是苹果。也就是说，这个对象和苹果的概念相似，而和其他的概念，例如桔子不相似。

估定模式相似度的具体方法和要进行的模式识别任务紧密相关，接下来我们将进一步给予说明。

1.2.1 分类决策

估定几个对象之间的相似度时，需要借助于对象本身特有的特性或特征。假设要设计一个区分绿苹果和桔子的系统，图 1-1 中给出了绿苹果和桔子的典型模式图。在这个辨别任务中，我们可以使用显而易见的特征，即颜色和形状进行区分，如图 1-2 所示。

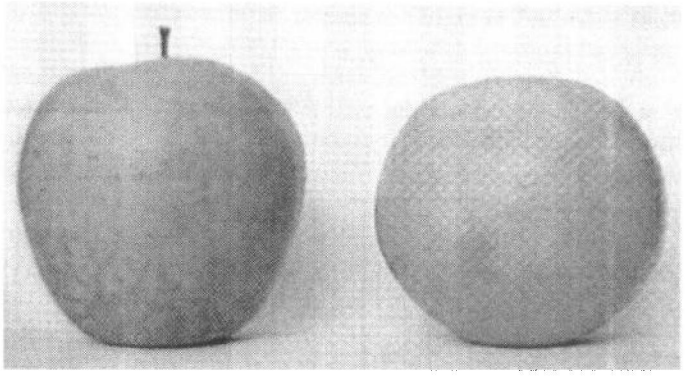

图 1-1 绿苹果和桔子的典型模式图

为了将颜色特征表达成数字形式，可以将图像分成红、绿、蓝三种基色来研究。下一步，我们从图像中选取一个感兴趣的中心区域，对该区域计算分布的红色素和绿色素在各自分布范围内(通常光强直方图的峰值为[0，255]；0 表示无色，255 表示满色)之比。图 1-3 显示的灰度图对应于苹果图像的绿色素，右边是所选矩形区域的绿色光强直方图。可以看出，绿色光强度直方图的峰值是 186，这表明最有可能的绿色光强度值为 186。同样，

我们研究红色素，可以得到另一个值 150。两个数值之比为 1.24，表明相对于红色而言，绿色占有优势。

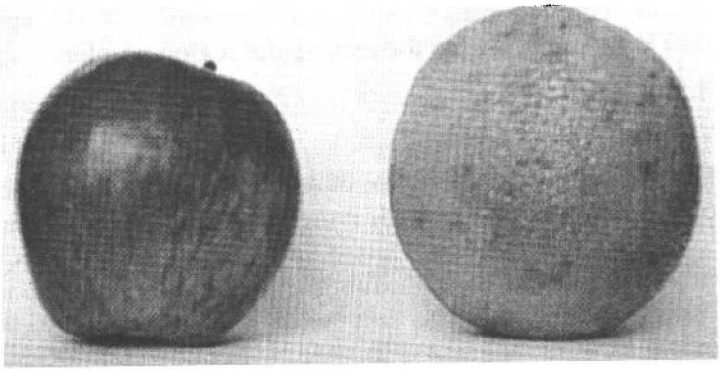

图 1-2　依照颜色和形状特征描绘的红苹果和浅绿色桔子

我们也可以将形状特征表达成数字形式。例如，可以测量图像顶部到最宽处的垂直距离，然后计算这个距离和图像高度的比值，即 x/h，如图 1-3 所示，x/h=0.37。注意到，这些都基于一个假设：物体是处于标准的直立位置。

如果我们准确选取了典型模式，则希望正常的绿苹果和桔子在二维特征空间所对应的点都将落在它们的典型模式所对应的点附近，即落在图 1-4a 中曲线封闭的区域之中。也就是说，如果我们选择了适当的特征，那么前面提到的这两种物体对应的点的集合将会在图中明显地分开，这样就可以对这两类水果进行区分了。

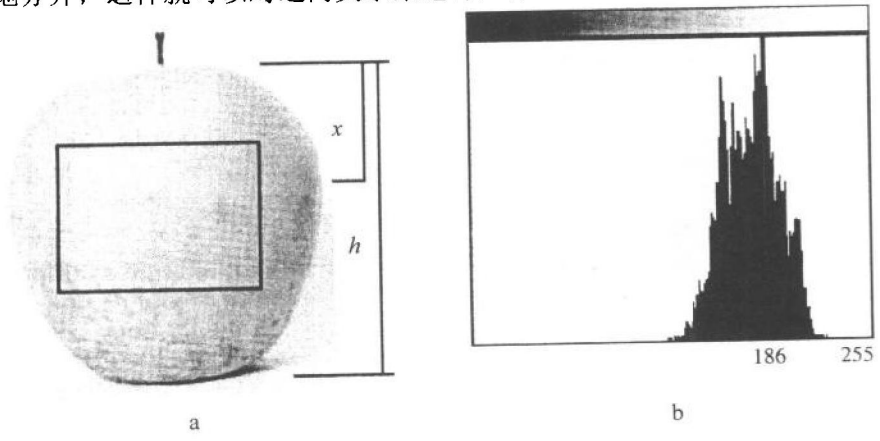

图 1-3　(a) 苹果图像的绿色素对应的灰度图；　(b) 图(a)中所示矩形区域的光强直方图

将某一对象归为某一类的模式识别任务其实就是进行分类。从数学观点上讲，在分类

时将模式表达成向量的形式很方便。下面是上例中的二维向量：

$$\boldsymbol{x}=\begin{bmatrix}x_1\\x_2\end{bmatrix}=\begin{bmatrix}\text{颜色}\\\text{形状}\end{bmatrix}$$

由绿苹果的典型模式我们可以得到：

$$\boldsymbol{x}_{\text{绿苹果}}=\begin{bmatrix}1.24\\0.37\end{bmatrix}$$

典型模式的特征向量所对应的点被扩展为一个方形和一个圆形区域，分别对应绿苹果和桔子，如图 1-4 所示。

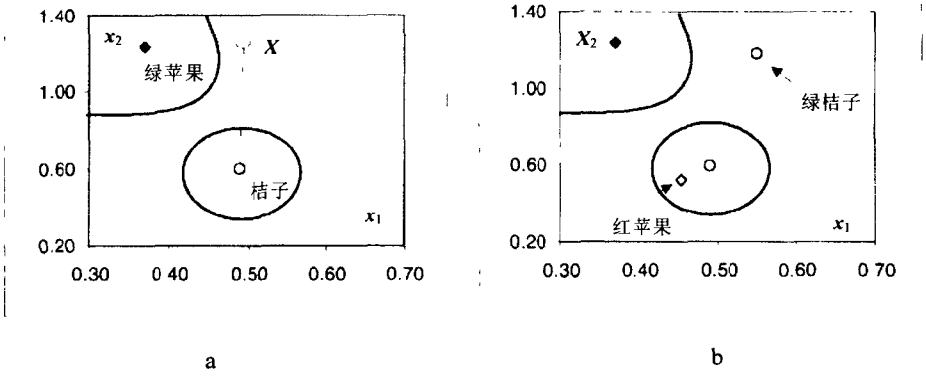

图 1-4 (a) 特征空间中的绿苹果和桔子； (b) 一个“类似”桔子的红苹果和一个可疑的浅绿色桔子

我们设想一个这样的机器，它可以通过所描述的特征分辨绿苹果和桔子。将一个水果提供给机器，通过计算它的特征得到一个向量，它和颜色-形状平面内的点 $\boldsymbol{x}$ 相对应见图 1-4a。然后机器使用该特征值作为输入，来判断它是绿苹果还是桔子。一种合理的方法是根据该点与典型模式所对应的点之间的欧氏距离来判别。也就是说，对机器来说，相似度的依据是欧氏距离，它将决定对象是否为绿苹果。在这种情况下，机器的输出是可取两个值的变量，例如，0 表示对象为绿苹果，1 则表示为桔子，这样的机器称之为分类器。

假设将图 1-2 中所示的红苹果和浅绿色桔子作为分类器的输入，那么它们的特征向量所对应的点如图 1-4(b)所示。红苹果被误认为桔子，因为它所对应的点距离桔子的典型模式点比距离苹果的典型模式点要近。出现这种结果并不奇怪，毕竟，这次辨别超出了分类器的范围。至于浅绿色桔子，它的特征向量距两个典型模式点的欧氏距离几乎相等，以至于无法对它进行分类。如果使用垂直偏差距离而非水平偏差距离代替欧氏距离作为判别依

据，那么浅绿色桔子也将被错误的分类。

总的来说，实际应用中的模式分类系统都存在缺陷，主要由以下原因引起：

- 使用的特征不适当或不充分。例如，通过补充表面粗糙度作为纹理特征，原先无法分类的浅绿色桔子将有可能被正确分类。
- 用来设计分类器的样本不是足够全面和具有代表性。例如，如果我们想区分苹果和桔子，那么我们将不得不考虑苹果样本中的每一种类，包括红苹果。
- 分类器的效率不是很高。例如，测量距离时的准确率不高，使用了不适当的典型模式。
- 类别之间存在交集以至于分类器无法区分。

在本书中，我们将重点放在选取适当的特征和设计高效的分类器上。最初的特征选取，与其说是门科学，倒不如说是门艺术。像任何一门艺术一样，通过实验和实践，特征的选取也会逐步得到改进。除了选取适当的特征和估计相似度外，人类还有其他手段来保证分类的高度准确，包括根据前因后果和高级知识结构，这些属于人工智能课程的范围，本书中将不作讨论。但人类在识别对象时也并非绝不出错，如果把浅绿色桔子混在一篮子柠檬中，那么根据生活经验也不见得就能将它从中区分开来。

1.2.2 回归问题

现在考虑另一类与认知推理过程紧密相关的问题。我们观察这样一个行为，动物依据气候的变化和自身生物周期的生理变化而迁徙。日常生活中，推理非常重要，因为它能较准确地引导人们作出决定。例如，众所周知的，路上行驶的汽车与前方车辆保持适当的距离，预报天气情况，根据经济形式的变化预测公司投资的盈利情况以及评估贷款的发放。

我们来考虑这样一个例子，预测 A 公司在股市中的股价情况，可作为依据的以往的资料有：A 公司和其他公司的股价、汇率和利率。在这种情况下，我们想依据过去一段时间同一股票的连续股价和其他变量的变化情况来预测股价变化。如图 1-5 所示，预测一天后股价的依据是 r_A，r_B，r_C，欧元对美元的汇率及 6 个月的利率。

我们看到的这个时间序列预测问题是个多类分类问题的例子，在数学上称为函数近似或回归问题。具有回归解法功能的系统进行预测时，通常预测值(图 1-5 中黑色圆点所示)与真实值(曲线)会有些偏差。预测值与真实值(也称为目标值)之间的差别产生了一个预测误差。我们的目标就是尽可能地减小误差。

事实上回归问题也能被转化为分类问题。我们对独立变量 r_A 的值域进行间隔足够小的分割，将回归解法转化成分类问题进行解决，这样，进行正确的分类就相当于使预测值

落在恰当的间隔之中，这里一个间隔等价于一个类别。从这个意义上讲，我们可以将这一系列变量看作一个特征向量，即[r_A r_B r_C 欧元 - 美元汇率 6 个月的利率]。此外，我们用距离来作为预期值和目标值之间相似性的度量尺度。对于一个粗略的回归问题：预测$r_a(t)$是否大于前一天的值 $r_a(t-1)$，此时等价于一个两类的分类问题，即对 $r_a(t)-r_a(t-1)$所得数值的符号(+或－)进行分类。

有时回归问题是分类的一部分。例如，在识别活体生物组织时，医学研究者经常使用一个称为品质因数的量，它取决于几个特征，如颜色、肌理、血管的光反射系数和密度。一个自动生物组织识别系统在作出分类判断之前，总是先试图回归到人类专家估定的品质因数。

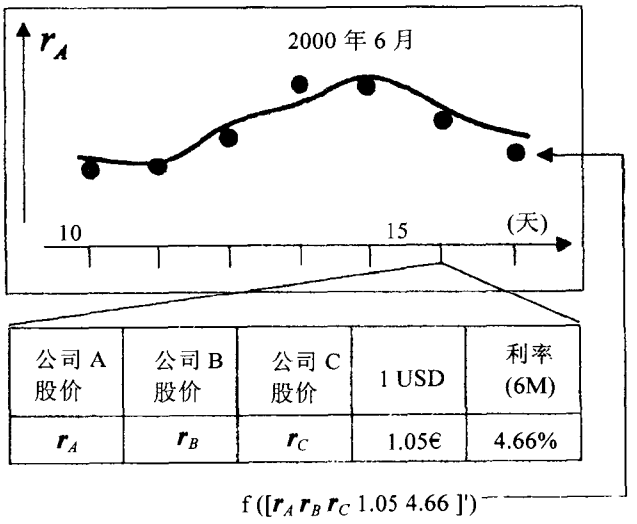

图 1-5 一天后的股价预测。r_A，r_B，r_C分别是 3 个公司的股价。根据 6 月 15 日的股价、欧元-美元的汇率和 6 个月的利率做出的 6 月 16 日 r_A 真实值(连续曲线)的函数近似(黑色圆点)

1.2.3 描述

在分类和回归问题中，相似度都是距离，因此可以用数值来度量。另一种类型的相似度与对象的特征结构有关。假设在一段时间内我们对胎儿的心跳频率进行跟踪记录，用仪器记录下胎儿瞬时的心跳频率(50 次/分钟到 200 次/分钟)，然后产科医生根据这些纪录来分析胎儿的健康状况。图 1-6 给出了一个这样的记录。