国外计算机科学教材系列

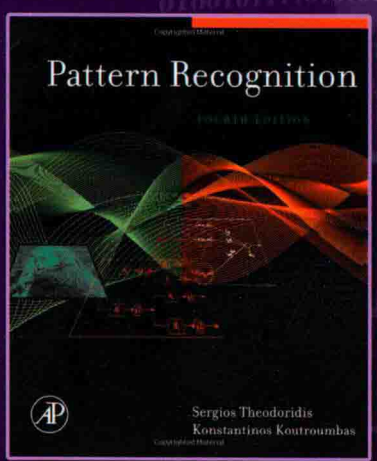
ELSEVIER

模式识别

（第四版）

Pattern Recognition

Fourth Edition



[希腊] Sergios Theodoridis
Konstantinos Koutroumbas 著

李晶皎 王爱侠 王 骄 等译

中国工信出版集团

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

国外计算机科学教材系列

模式识别

（第四版）

Pattern Recognition, Fourth Edition

[希腊] Sergios Theodoridis
Konstantinos Koutroumbas 著

李晶皎 王爱侠 王 骄 等译

電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

内 容 简 介

模式识别是信息科学和人工智能的重要组成部分，其主要应用领域有图像分析、光学字符识别、信道均衡、语言识别和音频分类等。本书全面阐述了模式识别的基础理论、方法和应用。全书在结合当前理论与实践的基础上，讨论了贝叶斯分类、贝叶斯网络、线性和非线性分类器设计、上下文相关分类、特征生成、特征选取技术、学习理论的基本概念及聚类概念与算法。与前一版相比，增加了大数据集和高维数据相关的最新算法；提供了最新的分类器和鲁棒回归的核方法；介绍了分类器组合技术，包括 Boosting 方法；新增了一些热点问题，如非线性降维、非负矩阵因数分解、关联性反馈、鲁棒回归、半监督学习、谱聚类和聚类组合技术。书中的每章均提供了习题与练习，并用 MATLAB 求解问题，给出了一些例题的多种求解方法；配套网站上提供了习题答案。

本书可作为高等院校自动化、计算机、电子和通信等专业研究生和高年级本科生的教材，也可作为计算机信息处理、自动控制等相关领域的工程技术人员的参考用书。

版权贸易合同登记号　图字：01-2009-1488

图书在版编目(CIP)数据

模式识别 /(希)西格尔斯·西奥多里蒂斯，(希)康斯坦提诺斯·库特龙巴斯著；李晶皎等译. —4 版
北京：电子工业出版社，2016.11
书名原文：Pattern Recognition, Fourth Edition
国外计算机科学教材系列
ISBN 978-7-121-30110-0
I. ①模… II. ①西… ②康… ③李… III. ①模式识别－高等学校－教材 IV. ①O235
中国版本图书馆 CIP 数据核字(2016)第 246780 号

策划编辑：谭海平
责任编辑：谭海平
印　　刷：三河市良远印务有限公司
装　　订：三河市良远印务有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编　100036
开　　本：787×1092　1/16　　印张：42　　字数：1129 千字
版　　次：2016 年 11 月第 1 版(原著第 4 版)
印　　次：2016 年 11 月第 1 次印刷
定　　价：89.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：(010)88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：(010)88254552，tan02@phei.com.cn。

译 者 序

模式识别诞生于20世纪20年代，随着20世纪40年代计算机的出现，20世纪50年代人工智能的兴起，模式识别在20世纪60年代初迅速发展成一门学科。模式识别研究的理论和方法在很多地方得到了成功的应用，从最初的光学字符识别（OCR），扩展到笔输入计算机、生物身份认证、DNA序列分析、化学气味识别、药物分子识别、图像理解、人脸辨识、表情识别、手势识别、语音识别、说话人识别、信息检索、数据挖掘和信号处理等。

尽管如此，与生物认知系统相比，模式识别系统的识别能力和鲁棒性还远不能让人满意。模式识别还有许多的基础理论和基本方法等待人们解决，新问题也层出不穷。为此，相关人员很需要一本关于这一领域的高水平学术著作，它既有基础知识的介绍，又有本领域研究现状的介绍，以及未来发展的展望等。本书正是这样一本经典著作。

本书是第四版，于2009年由模式识别领域的两位顶级专家合著，他们是希腊雅典大学信息与通信系的Sergios Theodoridis教授和希腊雅典国家天文台空间应用与遥感研究院的Konstantinos Koutroumbas博士。第四版的特点是：大部分章节增加了MATLAB编程和练习，新增了一些模式识别最新研究成果，如非线性降维、非负矩阵因数分解、关联性反馈、鲁棒回归、半监督学习、谱聚类和聚类组合技术。

为了适用于电力电子工程、计算机工程、计算机科学和信息以及自动控制等专业的研究生，以及高年级本科生各种不同知识背景的学生，本书内容安排既全面，又相对独立。在各个章节中需要的一些数学工具，如概率、统计和约束优化等知识，在本书的4个附录中做了简单的讲解。本书可以面向大学生和研究生，可以作为一学期或两个学期的课程。本书也可以作为自学教材，或供研究人员和工程技术人员参考。

负责本书初译的人员有：东北大学信息学院的王骄、闫爱云、张瑶、王亮、李亮、薛长江、李鹏飞、宋光杰。

负责本书译校的人员有：东北大学信息学院的王爱侠、李贞妮。

东北大学李晶皎教授完成最后译校。

在翻译过程中，我们力求忠实、准确地把握原著，同时保留原著的风格。但由于译者水平有限，书中难免有错误和不准确之处，恳请广大读者批评指正。

前　言

本书是作者在20年来给研究生和本科生教学的基础上编写的，该课程面向很多专业的学生，例如电力电子工程、计算机工程、计算机科学和信息以及自动控制等专业的研究生。这些经验使我们得以把本书内容编写得既全面又相对独立，并且适用于各种不同知识背景的学生。读者需要具备的知识包括微积分学基础、初等线性代数和概率论基础。在各个章节中需要的一些数学工具，如概率、统计和约束优化等知识，在本书的4个附录中做了简单的讲解。本书面向大学生和研究生，可以作为一个学期或两个学期的课程。本书也可以作为自学教材，或供研究人员和工程技术人员参考。我们编写本书的动力之一是，使这本书适合于所有从事模式识别相关研究的人员。

范围和方法

本书采用统一的方式讲述各种模式识别方法。模式识别是多个应用领域的核心，包括图像分析、语音和声音识别、生物统计学、生物信息学、数据挖掘和信息检索等。尽管这些领域有很多不同点，但也有共同之处，对它们的研究也有统一的方法，例如数据分类、隐藏模式等。本书的重点在于讲述现在常用的方法。读者可以从本书获得并理解相关的基础知识，进而研究更多的与应用相关的方法。

本书的每一章都采用循序渐进的讲解方式，从基础开始过渡到比较高深的课题，最后对最新技术发表评论。我们尽量保持数学描述和直接叙述之间的平衡，这不是一件容易的任务。然而，我们坚信对于模式识别，如果试图回避数学，将使读者很难理解算法的本质并丧失研究新算法的潜能；本书会使得读者能够很容易地解决遇到的问题。在模式识别中，最终采用的合适技术和算法在很大程度上依赖于所要解决的问题。根据我们的经验，讲解模式识别是一个使学生复习数学基础知识的好方法。

新增内容

第四版新增内容如下：

- 大部分章节的最后新增了MATLAB代码和计算机实验。
- 更多的实例和新增的图增强了本书的可读性。
- 有关当前热门问题的新增章节如下：
  - 非线性降维
  - 非负矩阵分解
  - 关联性反馈
  - 鲁棒回归
  - 半监督学习
  - 谱聚类

— 聚类组合技术

部分章节被重写，增加了更多的当前应用方面的内容。

补充内容

MATLAB 文档可从与本书相关的网站下载，网址是 www.elsevierdirect.com/9781597492720。电子文档的图和每章后习题和练习的解答都可从网站上下载。读者还可以下载相关问题的详细证明和本书所有章节的课件。

我们定期在网站上增加和更新 MATLAB 示例，欢迎读者多提建议。尽管网站上的内容经过多次仔细检查，但有些地方还是不可避免地存在错误，欢迎读者批评指正。

致谢

本书的出版离不开广大师生多年来的支持和帮助。特别感谢 Kostas Berberidis、Velissaris Gezerlis、Xaris Georgion、Kristina Georgoulakis、Leyteris Kofidis、Thanassis Liavas、Michalis Mavroforakis、Aggelos Pikrakis、Thanassis Rontogiannis、Margaritis Sdralis、Kostas Slavakis 和 Theodoros Yiannakoponlos。Yannis Kopsinis 和 Kostas Thernelis 自始至终都给予了莫大的支持和帮助。对本书的再版仔细阅读、提出大量批评和建议的有：Alexandros Bölnn、Dionissis Cavouras、Vassilis Digalakis、Vassilis Drakopoulos、Nikos Galatsanos、George Glentis、Spiros Hatzispyros、Evagelos Karkaletsis、Elias Koutsoupias、Aristides Likas、Gerassimos Mileounis、George Monstakides、George Paliouras、Stavros Perantonis、Takis Stamatoponlos、Nikos Vassilas、Manolis Zervakis 和 Vassilis Zissimopoulos。

本书的再版还要感谢读者的批评和建议，提出批评和建议的读者有：Tulay Adali, University of Maryland; Mehniet Celenk, Ohio University; Rama Chellappa, University of Maryland; Mark Clements, Georgia Institute of Technology; Robert Duin, Delft University of Technology; Miguel Figneroa, Villanueva University of Puerto Rico; Dimitris Gunopoulos, University of Athens; Mathias Kolsch, Naval Postgraduate School; Adam Krzyzak, Concordia University; Baoxiu Li, Arizona State University; David Miller, Pennsylvania State University; Bernhard Schölkopf, Max Planck Institute; Hari Sundaram, Arizona State University; Harry Wechsler, George Mason University 和 Alexander Zien，Max Planck Institute。

我们由衷地感谢这些同事所给予的批评和建议。非常感谢 N. Kalouptsidis 教授，长期以来我们的合作和友谊是本书灵感的来源。

最后，K. Koutroumbas 要感谢 Sophia、Dimitris-Marios 和 Valentini-Theodora 的耐心与支持。同时，S. Theodoridis 要感谢 Despina、Eva 和 Eleni，她们是快乐和动力的源泉。

目　录

第1章 导 论

1.1 模式识别的重要性

模式识别是一门以应用为基础的学科，目的是将对象进行分类。这些对象与应用领域有关，它们可以是图像、信号波形或者任何可测量且需要分类的对象。可以用专用术语“模式”（Pattern）来称呼这些对象。模式识别（Pattern Recognition）具有悠久的历史。在20世纪60年代以前，模式识别主要是统计学领域中的理论研究。同其他事物一样，计算机的出现提高了对模式识别实际应用的需求，而这反过来又对理论发展提出了更高的要求。就像我们的社会从工业化到后工业化阶段一样，在工业生产中，对自动化以及信息处理和检索的需求变得越来越重要，这种趋势把模式识别推向今天的工程应用和研究的高级阶段。在大多数机器智能系统中，模式识别是用于决策的主要部分。

在机器视觉中，模式识别是非常重要的。机器视觉系统通过照相机捕捉图像，然后通过分析，生成图像的描述信息。典型的机器视觉系统主要应用在制造业中，用于自动视觉检验或自动装配线。例如，在自动视觉检验应用中，生产的产品通过传送带移动到检验站，检验站的照相机确定产品是否合格。因此，必须在线分析图像，模式识别系统将产品分为“合格”和“不合格”两种。然后，根据分类结果采取相应的动作，比如丢弃不合格的产品。在装配线上，必须对不同的对象进行定位和识别，也就是说，将对象分类到已知类别的某一类中，如螺丝刀类、德国钥匙类以及任何工具制造单元，然后机器手把这些对象放置在正确的位置。

字符（字母或数字）识别是模式识别应用的另一个重要领域，主要用于自动化和信息处理。光学字符识别（Optical Character Recognition, OCR）系统已经开始在商业中应用，我们或多或少都对其有所了解。OCR系统有一个前端设备，它由光源、扫描镜头、文档传送机和检测器组成。在光敏检测器的输出端，光的强度变化转换成数字信号，并形成图像阵列。然后，用一系列的图像处理技术完成线和字符的分段，模式识别软件完成字符识别的任务，也就是将每一个符号分到相应的“字符、数字、标点符号”类中。与存储扫描图像相比，存储经识别处理的文档的好处是：更容易进行文字处理；存储ASCII字符比存储文档的图像效率更高。除了印刷体字符识别系统外，现在更多的研究集中于手写体识别。这种系统的典型商业应用是银行支票的机器识别，机器必须能够识别数字的个数和阿拉伯数字，并进行匹配，而且能够检查收款人相应的支出信用是否相符。哪怕只有一半的支票识别正确，这样的机器也可以将人力从枯燥的工作中解脱出来。另一个应用是邮局的自动邮政系统，它进行邮政编码识别。在线手写体识别系统是具有巨大商业利益的另一应用领域，此系统将用于笔输入计算机。在这种计算机中，数据的输入不是通过键盘而是通过手写，这顺应了开发具有人类技能接口的机器这一发展趋势。

计算机辅助诊断（Computer-aided diagnosis）是模式识别的另一个重要的应用，目的是帮助医生做诊断决定，当然最终的诊断由医生来完成。计算机辅助诊断已经应用于实际，主要研究各种医疗数据，如X射线、计算机断层图、超声波图、心电图（ECG）和脑电图（EEG）。计算机辅助诊断的系统需求来源于如下事实：医疗数据较难解释并且解释结果多依赖于医生的经验。我

们以检查乳腺癌的乳腺 X 射线照相术为例，尽管乳腺 X 射线照相术是检测乳腺癌的最好方法，但是 10%~30%的患病妇女在乳腺 X 射线照相术中可能得到相反的乳腺 X 射线照片。在这种情况下，大约 2/3 的放射线医师不能检测出癌变，很明显这是错误的。这可能是因为图像质量不好、放射线医师眼睛疲劳或病情等原因造成的。通过另一个放射线医师再次查看照片可以提高正确分类的百分比。因此，可以发明一种模式识别系统通过提出第二种观点来帮助放射线医师进行诊断。基于乳腺 X 射线照片日益准确的诊断反过来会减少乳腺癌疑似病例的数量，使这些人免于承受外科胸部活组织检查的痛苦。

语音识别（Speech recognition）是模式识别的另一个研究领域，在这个领域中已经进行了大量的研究。语音是人类最自然的沟通和交换信息的方式。因此，长期以来，建立能够识别语音信息的智能机器已成为科学家和科幻小说家的目标。这种机器的潜在应用是广泛的。例如，可以用来有效改善制造业的环境，可以远程控制危险环境中的机器，以及可以通过对话控制机器来帮助残疾人。经过努力，另外一个已经取得一定成功的重要应用是用麦克风向计算机进行语音输入，语音识别系统的软件能够识别语音文本，翻译成 ASCII 码，并可以显示在显示器上和存储在存储器中。计算机语音输入的速度是熟练打字员输入的两倍，而且有助于增强我们和聋哑人的交流能力。

数据库中的数据挖掘和知识探索（Data mining and knowledge discovery）是模式识别的另一个重要应用领域。数据挖掘广泛应用于医学和生物学、市场和财务分析、企业管理、科学探索、图像和音乐检索。它之所以受欢迎，是因为信息时代和知识社会里不断增强的信息检索，以及将其转化成知识的需求。由于信息存在于各种形式的海量数据中，如文字、图像、音频和视频，它们存储在全世界的不同地方。在数据库中，查找信息的最传统方法是基于模型描述，对象检索是基于关键词描述和部分字匹配。然而，这种搜索类型的前提是，已存储信息已经进行了人工标注；这是一项很费时的工作，尽管当存储的信息量有限时是可能完成的，但是当可用信息量变大时，这是不可能完成的。此外，当存储信息广为分布、由不同种类站点和用户共享时，人工标准就成问题了。基于内容的检索系统越来越受欢迎，这种系统是根据提交到系统中的对象与全世界各网站上的对象间的“相似性”来查询信息的。在基于内容的图像检索（Content-Based Image Retrieval, CBIR）系统中，图像传送到输入设备中（如扫描仪）；该系统返回基于可测量信号判断“相似”图像。信号是可编码的，如颜色、纹理和形状的相关信息。在基于内容的音乐检索系统中，例如（从音乐作品中摘录），用麦克风输入设备输入，系统返回“相似”音乐作品。在这种情况下，相似性取决于描述音乐作品的某些可测量特征，比如音乐韵律、音乐节奏和某些重复段落的位置。

自 20 世纪 90 年代中期开始，生物医学的采集和 DNA 数据分析得到了爆炸性的增长。所有的 DNA 序列包含有 4 个基本要素，核苷酸：腺嘌呤(A)、胞嘧啶(C)、鸟嘌呤(G)和胸腺嘧啶(T)；就像字母表中的字母和音乐中的 7 个音符一样，这 4 种核苷酸结合形成 S 形扭曲梯形长序列。通常情况下，基因由成百上千的核苷酸以特定顺序排列而成。特定的基因序列模型与特殊的疾病有关，在医学方面发挥着重要的作用。为此，模式识别是一个关键的领域，它为相似性搜索和 DNA 序列的对比提供了大量的开发工具。在医学领域，识别健康组织和病变组织之间的差异是非常重要的。

前面提到的只是众多可能应用中的 4 个例子。典型的应用包括指纹识别、签名认证、文本检索、表情和手势识别。目前的应用研究吸引了很多研究者，其目的是使人机互动更简单，并增强计算机在办公自动化等环境中的作用。为了激发想象力，值得一提的是 MPEG-7 标准，它包含对数字图书馆中录像带的视频信息检索：在数字图书馆中查找所有显示某人“X”微笑的视频场景。

当然，要在所有这些应用中达到最终目标，模式识别还依赖于其他一些学科的发展，如语言学、计算机图形学和计算机视觉等。

为了唤起读者对模式识别的好奇心，下面简要介绍模式识别的基本结构和方法，其中包括已经研究出来的各种各样的模式识别方法。

1.2　特征、特征向量和分类器

首先模拟一个简单的例子“mimicking”，这是一个医疗图像分类任务。图 1.1 给出了两幅图像，每幅图像中都有一块突出区域，两个区域彼此有明显的不同之处。我们可以认为图 1.1(a)所示的图像是良性的，属于 A 类；而图 1.1(b)所示的图像是恶性的（癌），属于 B 类。进一步假设有效样本（图像）不止这些，我们可以访问图像数据库，其中有一系列样本，一些样本是 A 类，一些是 B 类。

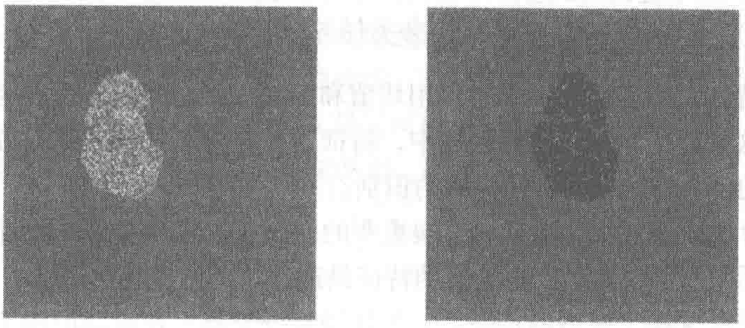


图 1.1　图像兴趣区举例：(a)A 类；(b)B 类

第一步是确定可测量值，用来区别两个图像区域。图 1.2 给出了每一个区域中的强度均值与其标准偏差的关系。每一点代表着已知数据库中一幅不同的图像，这表明 A 类样本和 B 类样本分布于不同的区域，中间的直线正好将这两类分开。假设现在有一幅新的图像，不知道它属于哪一类，我们计算兴趣区的均值强度和标准偏差，并画出相应的点，在图 1.2 中用*号表示，可以判定未知类型的样本更接近于 A 类。

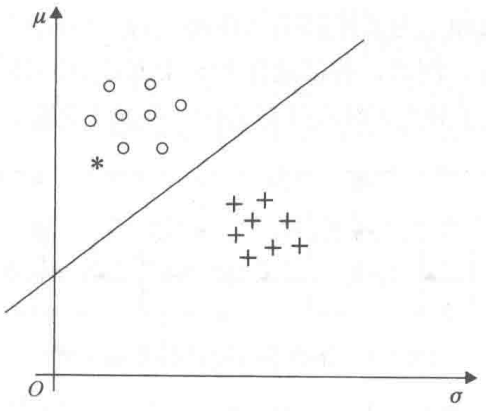


图 1.2　来自于 A 类（○）和 B 类（+）的一系列图像的相对于标准偏差的均值，这种情况下用一条直线将两类分开

前面所描述的人工分类的任务过程概述了大部分模式识别问题的基本原理。在这个例子中，

用来分类的测量方法——均值和标准偏差称为特征值。在一般的情况下使用 l 个特征 x_i, $i = 1, 2, \cdots, l$ 组成特征向量

$$\boldsymbol{x}=[x_1,x_2,\cdots,x_l]^{\mathrm{T}}$$

其中 T 表示转置。每一个特征向量表示一个样本（对象）。在本书中，特征和特征向量分别视为随机变量和向量，因为来源于不同样本的测量值是随机的数据。一方面是因为测量仪器的测量噪声，另一方面是因为每一种模式有截然不同的特点。例如，由于不同个体之间生理的不同，所以 X 射线成像有很大的不同，这也是图 1.1 中每一类的特征点发散的原因。

图 1.2 中的直线称为决策线，由它决定的分类器将特征空间划分为不同的类空间。如果对应于一个未知类别的样本特征向量 $\boldsymbol{x}$ 落在 A 类区域，则划为 A 类，否则划为 B 类。但这并不意味着决策是正确的。如果不正确，则出现了一个错误的分类。为了在图 1.2 中画出这条直线，我们需要知道图中每一个点的类别标签（A 类或 B 类），即用来设计分类器的样本（特征向量）的所属类是已知的，这些样本称为训练样本（训练特征向量）。

上面给出了定义和基本原理，下面给出分类任务中的基本问题。

- 怎样得到特征？在前面的例子中，用均值和标准偏差作为特征，因为我们知道应该从图像中提取这些特征。但在实际问题中，特征不是显而易见的。这是分类系统设计的特征提取阶段的任务，它完成已知样本的识别。
- 特征数 l 为多少最好？这也是一个很重要的问题，它在分类系统设计的特征选择阶段完成。在实际问题中，总是产生大量的特征供选择，要选择其中最好的使用。
- 对指定的任务选择了合适的特征后，怎样设计分类器？在前面的例子中，只是为了观察方便，根据经验画了一条直线。在实际问题中不可能这样，必须按照最优准则将线画在最优的位置。具有可接受性能的线性分类器（直线或 l 维空间的超平面）可能没有判定规则。一般来说，不同类别的区域划分是非线性的。在 l 维特征空间中，采用什么样的非线性分类器以及采用什么样的优化准则，这些问题在分类器设计阶段解决。
- 当分类器设计完毕后，如何评估分类器的性能？也就是说，分类误差率是多少？这是系统评估阶段的任务。

图 1.3 给出了分类系统设计的各个阶段，从这些反馈箭头可以看出，每一步都不是独立的。相反，它们相互关联、相互依赖。为了提高整体性能，每一阶段都有可能返回到前一阶段重新设计。而且有一些阶段可以合并，例如，特征选择和分类器设计阶段处于同一优化任务中。

虽然上面已经向读者描述了分类系统设计核心的一些基本问题，但是还有一些问题必须提到。

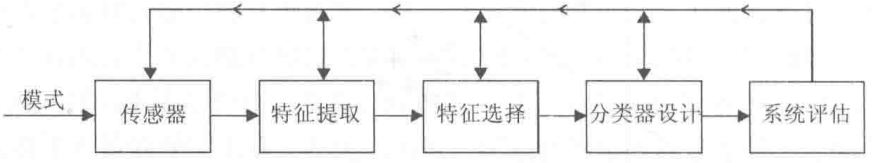


图 1.3 分类系统设计的基本步骤

1.3 有监督、无监督和半监督学习

在图 1.1 的例子中，假设有一个可用的训练数据集，并通过挖掘先验已知信息来设计分类器，这称为有监督模式识别（Supervised Pattern Recognition）。但是，并不总是这种情况，另外一种模

式识别是没有已知类别标签的训练数据可用。在这种情况下，给定一组特征向量 $\boldsymbol{x}$ 来揭示潜在的相似性，并且将相似的特征向量分为一组，这就是无监督模式识别（Unsupervised Pattern Recognition）或聚类（Clustering）。在社会科学和工程中会出现这种情况，例如遥感、图像分段、图像和语音编码。下面来看两个这样的问题。

在多光谱遥感中，放在人造卫星、航天飞机或太空工作站的灵敏扫描器测得从地球表面发射的电磁能量。这个能量可能是反射的太阳能（被动），或者是媒介发射的部分能量，目的是为了探测地球表面的情况。扫描器对电磁辐射的部分波段敏感，地球表面情况的特征不同，对波段反射的能量就不同。例如，在可见红外波段内，矿物质、潮湿的土壤、水域和潮湿的植被都是反射能量的主要贡献者。在热红外区，主要反映地球表面和地表下的热容量和热特性。因此，每一个波段测量地球表面同一块地方不同的特性，用这种方法可以根据不同波段的反射能量分布来生成地球表面的图像。研究这些信息的目的是识别各种地面类型，如公路、农田、森林、火烧地面、水和患病的农作物等。最后，生成了地球表面每一个单元的特征向量 $\boldsymbol{x}$，向量中的元素 x_i, $i = 1, 2, \cdots, l$ 对应于各种光谱波段中像素的强度。实际上，光谱波段的数量是变化的。

聚类算法可以用来完成对 l 维特征向量的分组。对应于相同地面类型的点，如水，将其聚类在一起形成一组。一旦这样分组以后，分析人员就可以通过将每一组中的样本点和地面数据的参考信息（地图或观察结果）相联系来识别地面类型，图 1.4 说明了这个过程。

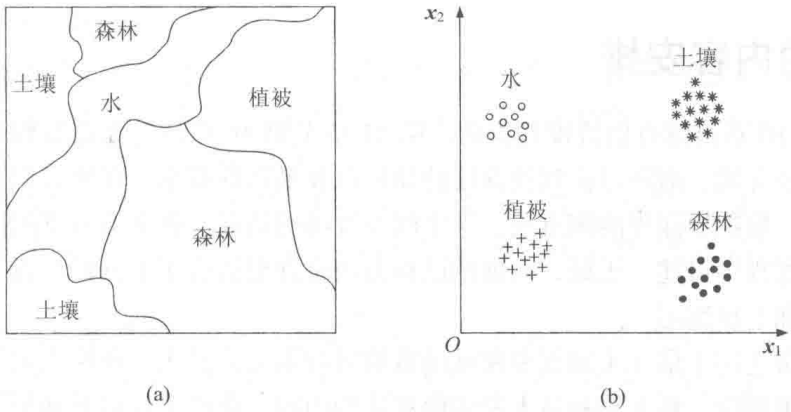


图 1.4　(a)各种地面覆盖物图示；(b)两波段多光谱图像聚类的各种特征

聚类也广泛地应用于社会科学，进行研究、调查、统计数据以及得到一些有用的结论来引导正确的行为。再来看一个简单的例子。假定我们既要研究一个国家的国民生产总值和人的文盲水平是否有关，又要研究国民生产总值和儿童的死亡率是否有关。在这个例子中，每个国家都用一个三维特征向量来表示，且特征向量的每一项与之对应。聚类算法将揭示低国民生产总值、高文盲和高儿童死亡率（以人口百分比表示）的这些国家的聚类相似性。

无监督模式识别主要用于确定两个特征向量之间的“相似度”以及合适的测度，并选择一个算法方案，基于选定的相似性测度对向量进行聚类（分组）。通常，不同的算法方案可能产生不同的结果，这一点必须由专家进行解释。

设计分类系统的半监督学习/模式识别（Semi-supervised learning/pattern recognition）与有监督模式识别有着同一目标。但是现在，设计者已经得出一系列原始未知类别的模式，外加已知类别的训练模式。通常我们称前者为标记数据，称后者为未标记数据。当系统设计者得到数量很有限的标记数据时，半监督模式识别就很重要了。在这些情况下，从未标记样本恢复附加信息，与现