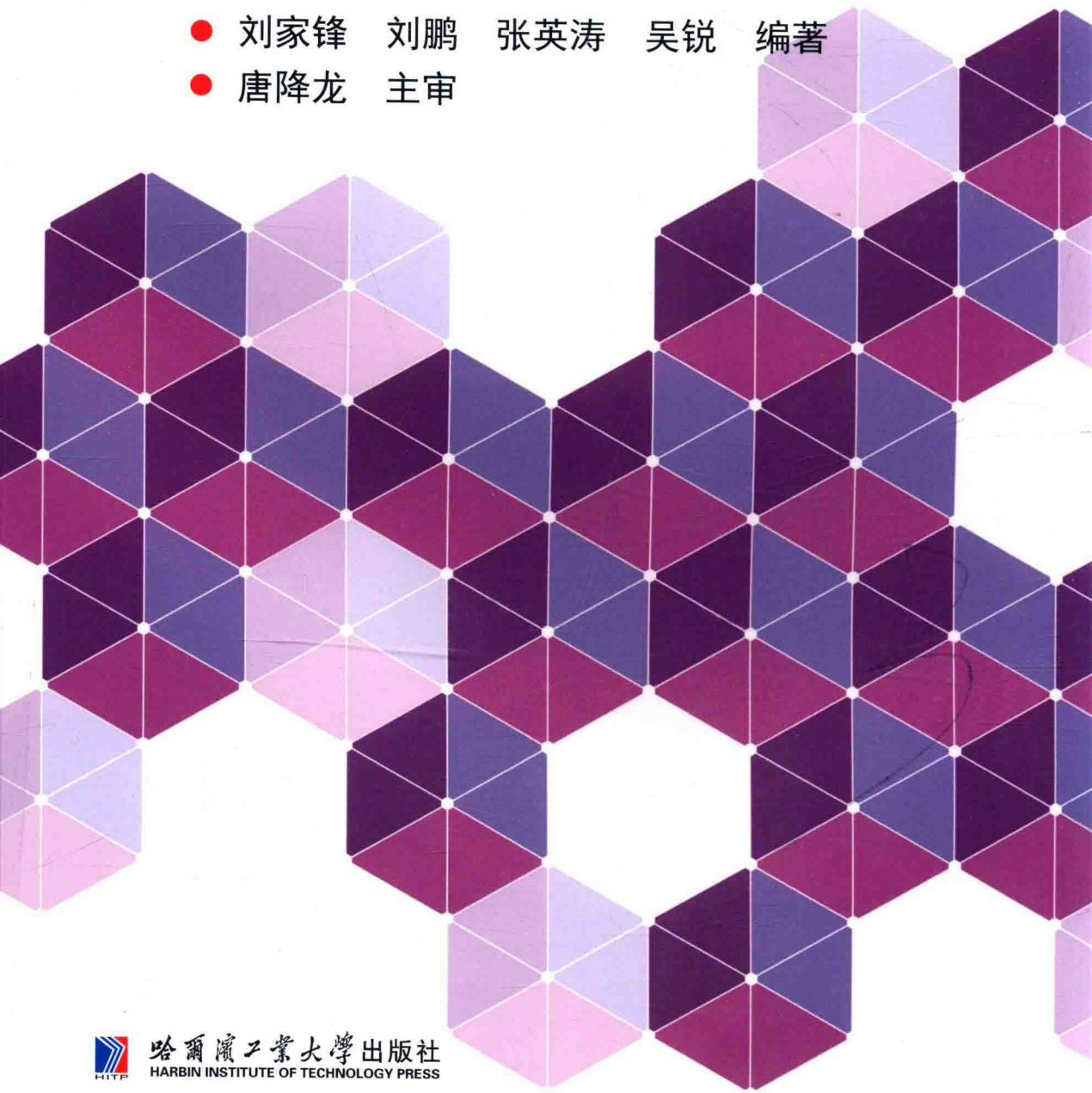
工业和信息化部“十二五”规划教材
“十二五”国家重点图书出版规划项目

模式识别（第2版）

Pattern Recognition

- 刘家锋　刘鹏　张英涛　吴锐　编著
- 唐降龙　主审



哈爾濱工業大學出版社
HARBIN INSTITUTE OF TECHNOLOGY PRESS

工业和信息化部“十二五”规划教材

“十二五”国家

模式识别（第2版）

Pattern Recognition

- 刘家锋　刘鹏　张英涛　吴锐　编著
- 唐降龙　主审

哈爾濱工業大學出版社

HARBIN INSTITUTE OF TECHNOLOGY PRESS

内容摘要

本书重点介绍模式识别的基本概念和基本方法，在保证理论完整性的前提下，详细讨论具体算法的基本思想、实现方法、优缺点以及适用领域，使读者在了解模式识别基本理论的同时能够掌握分类器设计方法，通过具体的应用实例和实践环节，帮助读者尽快做到理论与实践的结合，掌握模式识别方法用以解决在具体应用中所遇到的问题。

本书主要面向初学者和有一定自学能力的科研工作者，致力于对模式识别理论框架下概念的准确把握和算法实际应用能力的培养，算法与实用并重。本书可作为计算机科学与技术、电子科学与技术、控制科学与工程等专业本科生、研究生的教材或教学参考书，也可为其他相关专业人员了解模式识别方法提供入门参考。

图书在版编目(CIP)数据

模式识别/刘家锋等编著. —2 版. —哈尔滨：哈尔滨工业大学出版社，2017.6

ISBN 978－7－5603－6713－2

Ⅰ.①模… Ⅱ.①刘… Ⅲ.①模式识别 Ⅳ.①O235

中国版本图书馆 CIP 数据核字(2017)第 125675 号

策划编辑 王桂芝 李子江
责任编辑 李广鑫 刘 威
出版发行 哈尔滨工业大学出版社
社 址 哈尔滨市南岗区复华四道街 10 号 邮编 150006
传 真 0451－86414749
网 址 http://hitpress.hit.edu.cn
印 刷 哈尔滨市工大节能印刷厂
开 本 787mm×1092mm 1/16 印张 16.25 字数 391 千字
版 次 2014 年 8 月第 1 版 2017 年 6 月第 2 版
2017 年 6 月第 1 次印刷
书 号 ISBN 978－7－5603－6713－2
定 价 35.00 元

再版前言

在这个信息化的时代里，计算机已经无处不在。随之而来的是人们对计算机智能化程度要求的不断提高，希望计算机能够自动感知和适应周围环境，或通过认知和学习能够理解和发现信息与数据背后隐藏的事物的内在规律。模式识别正是研究解决此类问题的学科。几十年来，这个领域的发展非常迅速，取得了丰富的研究成果，模式识别方法已被成功地应用于字符识别、生物特征身份认证、语音识别、图像理解与计算机视觉、信息检索、数据挖掘等多个领域。“模式识别”作为一门课程也在高等院校的教学过程中越来越受到了重视。本书是作者在哈尔滨工业大学模式识别研究中心多年讲授“模式识别”课程所积累的教学经验和课程讲义基础上整理形成的。

模式识别领域有丰富的研究内容，为了解决相关问题，研究者不断提出的模式识别方法及其各具特色的变体往往会使初学者陷入无所适从、难以抉择的困境；同时，作为一门既有严格理论框架，又有广泛应用背景的学科，模式识别课程的教学方式多种多样。本书从课程教学的角度出发，没有将最新的研究成果呈现在读者面前，而是侧重于介绍具体模式识别方法的思想来源、工作原理及其实现过程，重点剖析该领域中最经典和最常用的几种模式识别系统设计和实现方法，提示读者每个算法在实现过程中需要注意的诸多细节，并给出相应的Matlab代码供读者参考，使初学者尽快掌握和运用模式识别的理论和方法来解决实际问题。

本着由浅入深、初学者易于掌握的原则，本书第2章和第3章主要是从模式的相似性和距离度量入手，分别介绍了两种简单的分类方法和三种聚类算法，并且讨论了如何评价一个识别系统的性能以及聚类结果的有效性；第4章和第6章分别讨论了线性和非线性判别函数分类器的设计方法，重点介绍了线性网络、多层感知器网络和支持向量机的原理和算法；第5章的内容与其他章节略有不同，在这一章里介绍的算法不是对模式进行分类，而是对描述模式的特征进行选择和提取，降低特征的维数，以简化分类器的设计；第7章在分析了贝叶斯判别这一模式识别理论基础的同时，重点介绍了高斯模型、高斯混合模型以及隐含马尔科夫模型的分类与学习算法；第8章以图像识别、字符识别的实例形式给出了模式识别算法的实际应用方法。

此次再版由刘家锋、刘鹏、张英涛、吴锐、赵巍、金野共同负责，具体分工如下：第1、4、5章由刘家锋负责，第2章由张英涛负责，第3章由赵巍负责，第6章由刘鹏负责，第7章由金野负责，第8章由吴锐负责。全书由刘家锋、赵巍统稿，唐降龙教授审阅了全书的内容。

鉴于作者水平有限，书中难免存在疏漏和不妥之处，敬请读者指正。联系人：赵巍 zhaowei@hit.edu.cn。

作者

哈尔滨工业大学计算机科学与技术学院

模式识别研究中心

2017年5月

目　录

第1章 绪 论

人能够很容易地分辨出不同的物体，认出熟悉的人，听懂别人说的话。辨识能力是人类最基本的智能行为，是人感知和理解周围环境，与外部世界进行交流的基础。识别能力对人类来说是极为平常的，甚至动物对不同的对象也有一定的分辨能力，例如可以区分不同的食物，发现敌害以避免受到攻击等。

生物体(包括人)是如何识别对象的？是如何具有识别能力的？这类问题属于认知科学的范畴，是心理学、哲学、生物学和神经科学的研究内容；而模式识别则是从工程的角度考虑，针对给定的任务和应用，研究如何使计算机具有识别能力的理论和方法。

什么是模式？粗略地说，存在于外部世界中每一个要识别的对象都可以称作是一个模式；更准确地说，模式并不是指识别对象本身，外部世界的事物只有通过人的视觉、听觉、嗅觉、触觉器官的感知才能够为人所认识，而模式则是指计算机通过对信号的采样、量化和处理之后得到的关于识别对象描述的一组属性集合，例如视觉识别对象的颜色、大小、形状，听觉识别对象的声音在各个频率上的能量分布等。在特定的任务和应用中，不同的模式可能属于同一个类别，例如同属于桌子类别的对象可以有不同的大小、形状和颜色。模式识别有时也被称为模式分类，所要研究的就是如何根据模式判断不同的识别对象是否属于相同类别。

1.1 模式识别的应用

统计学领域对于人类决策和分类行为的理论研究有着很长的历史。到了20世纪60年代，随着计算机的发明以及之后在多个领域的广泛应用，自动化技术和人工智能系统对模式识别提出了迫切的需求，这极大地推动了这个领域理论、方法和应用的研究。下面通过几个实际的应用场景来说明模式识别的过程和方法。

如何准确鉴定一个人的身份、保证信息安全是金融、电子商务、重要场所的安全检查、刑侦等领域需要解决的重要问题。原有的身份认证手段——卡(ID卡)、密钥、口令等极易伪造和丢失，无法满足信息化时代的要求。生物特征鉴别就是利用人体特有的生物特征进行身份认证的技术，这些特征包括指纹、视网膜、虹膜、面部图像、指静脉等生理特征，也包括笔迹、声纹、步态等行为特征。以人脸识别为例，首先需要由照相机或摄像机拍摄包含人脸的图片，数字化采样成为数字图像输入计算机；然后使用图像处理技术检测出人脸所在的区域，校正图像的亮度和位置、方向；最后由识别系统与保存在人脸数据库中的图像进行比对，确定其身份。

对生产线上的产品进行缺损检测是保证产品质量的重要手段，在快速的生产过程中由人力来检测产品的缺陷是一件很繁重的工作，往往会由于人的疲劳出现过多的漏检和误检，采用计算机视觉和模式识别技术自动检测产品质量是现代化生产线上的一个重要环节。当

产品在生产线上运行到一定的位置时触发相应的成像设备拍摄该产品的图像，然后由识别系统在线地将其分类为“合格产品”或“有缺陷产品”，并将分类结果发送到相应的执行机构，由执行机构进行不同的处理，例如将缺陷产品剔除，将合格产品装箱等。

字符识别和语音识别技术现在已经被普遍地应用到了办公自动化领域和日常的移动智能设备上。语音识别首先由麦克风采集人说话的声音，转换成数字波形信号输入计算机；然后使用数字信号处理技术对输入的声音信号进行处理，例如滤除噪声、分析信号的频谱等；识别系统根据信号处理的结果进行分类，实现声音信息到文字信息的转换；最后将识别的结果交由其他环节使用，如在屏幕上显示文字信息的内容，将文字信息由一种语言翻译成另一种语言，或者作为智能问答系统的输入。

字符识别技术根据识别的内容可以分为手写字符识别和印刷体字符识别，根据字符的输入形式可以分为在线识别和光学字符识别。人在移动智能设备的触摸屏或数位板上书写，计算机根据书写的轨迹识别字符的方式称为在线的手写字符识别；由扫描仪或照相机将印刷或手写在纸上的字符转换成数字图像，然后由识别系统将其转换为相应的汉字、字母、数字或标点符号的过程称为光学字符识别。在光学字符识别中首先需要采用一系列的图像处理技术对输入的手写或印刷字符图像进行处理，校正图像的方向、切分文本行、分割出单个字符，然后由识别系统对包含单个字符的图像进行分类，将其转换为相应的字符编码，如ASCII 码、汉字编码、Unicode 码等。

计算机辅助诊断是模式识别技术应用的另一个重要领域。X 光、CT、B 超、核磁共振是现代医学影像检查的重要手段，但是医学图像普遍存在图像质量比较差、不易被人直观理解、病灶存在于一些细微之处等缺点，是否能够使用医学影像的手段准确诊断疾病很大程度上依赖于医生的经验。在实践中人们也发现，当医生需要做出判断时如果能够听取另外一位更有经验的影像科医生查看同一幅图像之后的意见，就可以有效提高诊断的准确率。计算机辅助诊断系统通过对已有病例的学习获得医学图像中与疾病相关的知识，能够对输入图像进行分析、判断，可以对医生的诊断起到辅助作用。

从以上所举的几个例子可以看出，人是运用自身的经验和知识完成了对外部事物、景物的感知和辨识，而模式识别则是希望能够将这些知识和经验传授给计算机，使得计算机也能够具有自动识别和感知周围事物的能力。当计算机具有自动识别的能力之后，就可以替代或辅助人类完成许多繁重、危险的工作。除了上述应用领域之外，模式识别技术也被广泛应用于机器人、自动车辆导航、考古、地质勘探、航天、军事等领域。

1.2 模式识别系统

从上一节所举的应用实例可以看出，模式识别系统所完成的工作是从外部世界获取一个所要识别对象的数据，经过分析和处理之后辨识其类别属性。完整的识别系统一般需要包括识别和训练两个过程，如图 1.1 所示。

1. 数据采集及预处理

数据采集是将外部世界需要识别的对象数字化为波形、图像、文本等计算机可以处理的形式输入识别系统。在数据采集过程中难免会有噪声或者其他与识别对象不相关的信息混入，在预处理过程中需要滤波去除噪声，将识别对象从背景中分离出来。

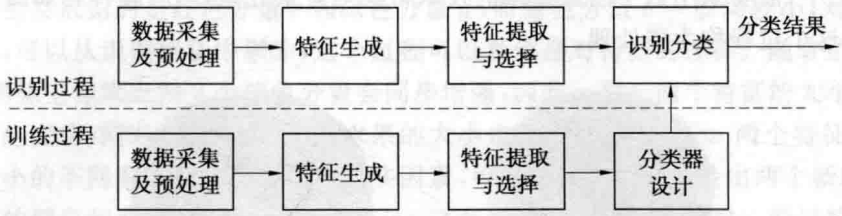


图 1.1 模式识别系统框图

2. 特征生成

经过数据采集得到的数据量一般比较大，很难直接由分类器进行识别，需要对原始信息进行处理，找出描述不同类别对象之间差异的“特征”，分类器再根据这些特征来判别识别对象的类别属性。

3. 特征提取与选择

采用什么样的特征对对象进行识别是模式识别系统设计的一个关键问题，但是特征往往是与实际应用相关的，不同的分类问题需要不同的特征。对于识别系统的设计者来说，一般很难预先确切地知道所面对的问题到底需要哪些特征，哪些特征能够很好地区分所要识别的对象。通常的做法是尽量多地从原始数据中生成与问题相关的特征，然后选择出最有效的特征，或对这些特征进行组合得到一组更有效的特征，这个过程一般称作特征的提取与选择。

4. 识别分类

经过特征生成和特征提取与选择之后，每个被识别的对象都被描述为一组特征，这组特征一般被称为“特征矢量”。每个对象所对应的特征矢量可以看作是“特征空间”中的一个点，识别分类环节根据特征矢量判别对象所属的类别，这个过程可以看作是采用一定数学方法实现了从“特征空间”到“类别空间”的映射。

5. 分类器设计

对于一些简单问题，可以采用人工的方式设计分类器，决定什么样的特征矢量应该映射为哪一个类别。然而，对于一个复杂的实际问题来说，人工方式设计分类器往往是低效的，特别是当特征的数量很多、特征空间的维数很高时，很难通过人的直观感觉设计出一个合适的分类器。人对周围不同物体和对象的分辨能力也不是与生俱来的，而是通过后天不断地学习和训练，以及对经验的总结逐渐形成的，识别系统中的分类器一般也需要一个训练和学习的过程。在训练过程中设计者需要提供大量不同识别对象的实例(这些实例一般被称作“训练样本”)，而识别系统则采用一定的训练和学习算法在这些训练样本的基础之上自动完成分类器的设计。

下面通过一个简单的应用来详细介绍识别系统的各个环节。一个食品加工厂需要桃子和橘子两种水果，假设进厂时两种水果是混在一起的，需要通过一个传送带将其分开进行加工。

可以在传送带的特定位置安装一个 CCD 摄像机，当一个水果到达镜头下方时自动拍摄一幅数字图像输入计算机，这个过程完成的是对识别对象的数据采集。输入图像中的水果

处于背景之中，需要采用数字图像处理的技术将其与背景分离，然后旋转到如图 1.2 所示的正立位置，这个过程称为预处理。

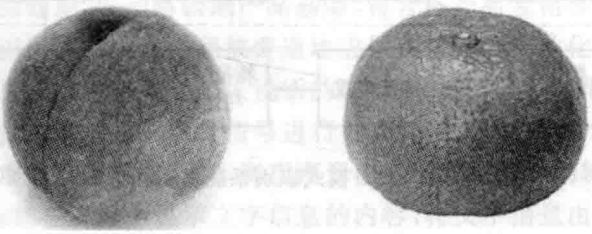


图 1.2 桃子和橘子的图像

下一步需要根据预处理之后的水果图像生成用于识别的特征。通过观察可以发现，桃子和橘子最明显的差别在于颜色的不同，因此首先选择颜色作为识别特征。数字图像可以看作是一个像素点的矩阵，每个像素点的颜色由红、绿、蓝 3 种基色的强度决定，一般来说每个基色的强度由一个字节表示，0 表示最弱，255 表示最强。经过预处理之后的图像中只包含一个水果，但是由很多个像素点所组成，使用某个像素的颜色代表整个水果的颜色不是一个合理的选择，可以以图 1.3 中水果图像中心矩形内所有像素点的颜色平均值作为生成的颜色特征 (r,g,b)。

成熟的橘子通常是橙色的，而桃子偏红色，没有完全成熟的橘子和桃子都会有较多的绿色成分，因此仅仅依靠颜色特征有时不能很好地区分两种水果，需要有其他特征的辅助。观察图 1.2 会发现，一般来说桃子的形状偏圆，而橘子偏扁，因此由水果的图像还可以生成出形状特征。然而，准确地描述一个真实物体的形状是很困难的。特征生成的最终目的是为识别服务，可以采用一种简单的方式获取区分桃子和橘子形状的特征。对于已经旋转到正立位置的水果图像，可以很方便地计算出图像前景区域的高度 h 和宽度 w，这样就生成了形状的粗略描述特征 (h,w)，如图 1.4 所示。

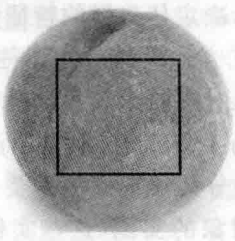


图 1.3 颜色特征的生成

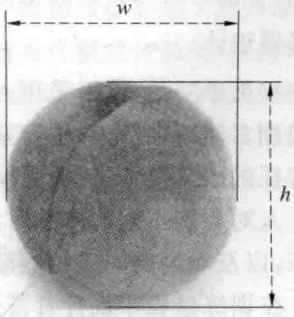


图 1.4 形状特征的生成

将颜色特征和形状特征结合在一起就得到了描述水果图像的 5 个特征，在模式识别中一般将这些特征写成列矢量的形式，称为特征矢量：$\boldsymbol{y}=(r,g,b,h,w)^{\mathrm{T}}$。按照这样的方式，每一个识别对象经过特征生成之后都会得到一个 5 维的特征矢量，对应着特征空间中的一个点。

如果再来考察一下这 5 个识别特征就会发现，桃子和橘子的颜色分别是红色和橙色，因

此区分两者主要依据的是红色分量 r 和绿色分量 g，而蓝色分量 b 一般均较小，对识别起的作用也很小，可以从识别特征中剔除，这个过程可以看作是对特征的选择。随着拍摄光照的增强，数字图像上像素点的 3 个颜色分量会同步增强，因此 r 和 g 两个特征的大小会随着光照的强弱发生变化；同时，待识别的不同水果的大小也是不同的，h 和 w 两个特征也会随着水果个体大小的不同发生变化。考虑到这些因素，可以由原始特征构造出两个新的特征，分别描述水果的颜色和形状：$x_1=g/r$，$x_2=h/w$，这个过程称为特征的提取。经过特征的选择和提取之后，每个识别对象被描述为一个 2 维的特征矢量：$\boldsymbol{x}=(x_1,x_2)^{\mathrm{T}}$，以特征矢量形式描述的识别对象往往也被称为“样本”。

图 1.5 显示了由颜色和形状所描述的桃子和橘子在 2 维特征空间中的大致分布，不同种类的水果分布在不同的区域，考虑到有些桃子的颜色并不是红色的而是青色的，因此可能分布在两个区域。分类器可以根据待识别的样本所处的区域对其类别进行判别。这样，由数据采集到识别分类就构成了一个完整的模式识别过程。

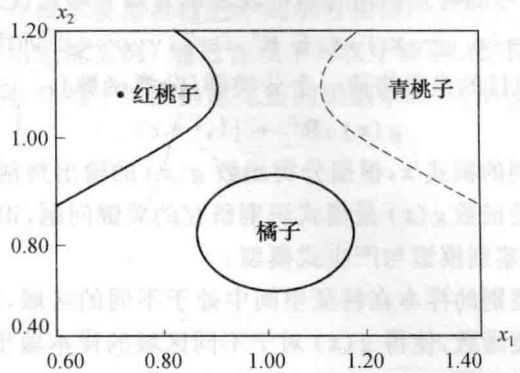


图 1.5　特征空间中桃子和橘子的大致分布

在这个水果识别的例子中，由于经过特征生成、提取与选择之后得到的是一个 2 维的特征矢量，因此可以很容易地分析出不同类别识别对象分布的大概区域。而对于一个复杂的识别问题来说，往往会提取出更多维的特征，每个识别样本就会处于一个高维的特征空间之中，完全依靠人的观察很难确定不同类别的分布区域。通常的做法是由计算机自动地完成不同类别区域的划分，这就是一个分类器的学习和设计的过程。例如在水果识别问题中，可以预先采集一些桃子和橘子放到传送带上，由识别系统采集数字图像并提取成相应的特征矢量，这些样本所构成的集合一般被称为“训练样本集”。然后计算机采用一定的算法，根据输入的训练样本集自动地完成分类器的设计工作，经过训练和学习的分类器可以用于对未知类别的样本进行分类。识别过程和训练过程构成了一个完整的模式识别系统。

1.3　模式识别方法

在模式识别系统中，数据采集、预处理和特征生成的过程是与问题相关的，不同的应用需要采用不同的采集手段，生成不同的特征，而模式识别主要研究的是如何根据生成的特征集合选择和提取出更加有效的识别特征，以及分类器的训练、学习与识别方法。

1. 有监督学习与无监督学习

模式识别方法根据训练样本的不同可以分为有监督学习方法和无监督学习方法。有监督学习又被称为有教师学习，需要知道训练样本集合中的每个样本具体属于哪一个类别；无监督学习又被称为无教师学习，只知道训练样本集合中的每个训练样本，而不知道每个训练样本所属的类别，甚至在有些情况下所属类别的数量也是未知的。在水果识别的例子中，如果在分类器的训练过程中，首先采用人工的方式将一批水果分成桃子和橘子两类，分别输入到系统中，那么就可以采用有监督学习方法进行分类器的设计和学习；如果预先没有进行人工分类，直接将混合在一起的水果输入到系统中，那么就需要采用无监督学习的方法来设计和学习分类器，在分类器工作时会将需要识别的水果分成两类，由人来确定哪一类属于桃子，哪一类属于橘子。

2. 鉴别模型与产生式模型

一般来说，有监督学习的模式识别问题可以表示为如下形式：已知由 n 个 d 维特征矢量组成的训练样本集合 $D=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}$，$\boldsymbol{x}_i\in\mathbf{R}^d$，$i=1,\cdots,n$，$n$ 个训练样本分别属于 c 个不同的类别 $\omega_1,\cdots,\omega_c$；学习的目的是要构造一个分类器（分类函数）：

$$g(\boldsymbol{x}):\mathbf{R}^d\rightarrow\{1,\cdots,c\}$$

对于每一个需要识别的模式 $\boldsymbol{x}$，根据分类函数 $g(\boldsymbol{x})$ 的输出判别其类别的属性。

如何构造和学习分类函数 $g(\boldsymbol{x})$ 是模式识别研究的关键问题，识别方法根据设计思想的不同大致可以分为两类：鉴别模型与产生式模型。

鉴别模型认为不同类别的样本在特征空间中处于不同的区域，这类方法在训练过程中通过训练样本集学习分类函数，使得 $g(\boldsymbol{x})$ 对于不同区域的样本输出不同值，依据判别函数的输出值来判断待识别样本处于特征空间中哪个类别的区域。根据分类函数 $g(\boldsymbol{x})$ 复杂程度的不同，鉴别模型又可以分为线性和非线性两种。

产生式模型将模式看作是分布在特征空间中的一个随机矢量，每一个类别的模式可能出现在空间的任意一点，只不过在某些区域出现的概率大一些，某些区域出现的概率小一些（甚至为 0）。针对出现在特征空间中某一点的待识别模式，产生式模型根据该点属于哪个类别的概率更大来判别其类别属性。

一种模式识别方法属于鉴别模型还是产生式模型是相对的，并不是绝对的，鉴别模型方法换一种方式理解也可以看作是产生式模型方法，而产生式模型也可以做鉴别学习。

1.4 内 容 安 排

为了便于初学者的理解，本书的内容安排遵循由浅入深，由简入繁的原则。

第 2 章（距离分类器）介绍了几种简单直观的模式识别方法，这些方法主要依据的是模式之间的距离（或相似程度）来构造分类器，这些方法有时也被称为“模板匹配”。距离分类器的特点是算法简单，分类器学习和识别的效率很高，目前仍然是解决识别问题的主要方法之一，特别是某些类别数量比较多、识别速度要求比较快的应用场景。

第 3 章（聚类分析）主要介绍了几种无监督学习中的聚类分析方法，这些方法仍然采用距离来度量模式之间的相似程度，然后根据无监督样本集合中样本的相似程度将其划分成

不同的子集，实现对无监督样本的聚类。

第 4 章（线性判别函数分类器）介绍了一种最简单的鉴别模型分类器 —— 线性判别函数分类器的学习与识别方法，线性分类器采用线性函数构造出特征空间中的一系列（超）平面，由这些（超）平面将空间划分成不同的区域，每个区域对应不同的类别。

由于后两章介绍的识别方法比较复杂，当识别特征的数量比较多时往往很难取得好的效果，为了使读者能够在学习过程中很方便地使用所学习的识别算法在相应的样本集上进行实验，本书的第 5 章（特征选择与特征提取）介绍了几种常用的降低特征矢量维数的方法，方便后续章节算法的实验。

第 6 章（非线性判别函数分类器）介绍的算法同线性分类器一脉相承，通过不同的方式将线性的判别方法转化成为非线性的方法，使用更加复杂的（超）曲面对特征空间进行划分。

第 7 章（统计分类器及其学习）内容的核心是贝叶斯分类器，在此基础上介绍了几种常用的产生式概率模型，以及概率模型参数的不同学习和估计方法。

第 8 章（模式识别应用系统实例）通过在线手写汉字识别、纸币图像识别和乳腺超声图像识别 3 个具体应用实例，介绍了如何构建完整的识别系统，应用模式识别方法解决具体的分类问题。

第 2 章　距离分类器

能够识别不同对象似乎是人类一种与生俱来的能力，当问一个人为什么认为一个对象属于这一个类别，而不是那一个类别时，最可能得到的回答是目标与这个类别更像。例如当遇见一个人的时候，会在脑海中与以前见过的人进行比对，如果发现他(她) 同某人长得非常相似，则很有可能判断遇见的就是这个人。

识别对象与某个类别是否相似是人在做出判断时的一个基本依据，根据这个思路，也可以利用相似性来构造用于计算机识别的分类器，这就是本章将要介绍的“距离分类器”。只要能够判断样本与类别之间的相似程度或者样本与样本之间的相似程度，就可以构造出一个距离分类器，所以说这是一种最简单的分类方法。

2.1　距离分类器

2.1.1　距离分类器的一般形式

距离分类器的目的是将需要识别的样本 $\boldsymbol{x}$ 分类到与其最相似的类别中，因此如果能够度量 $\boldsymbol{x}$ 与每一个类别的相似程度 $s(\boldsymbol{x},\omega_i)$，$i=1,\cdots,c$，那么就可以采用如下的方式进行分类：

$$\text{如果 } j=\underset{1\leqslant i\leqslant c}{\arg\max}\, s(\boldsymbol{x},\omega_i)\text{，则判别 } \boldsymbol{x}\in\omega_j \tag{2.1}$$

这是一种常用的数学化表示方式，含义是如果 j 是在所有 i 的可能取值中使得 $s(\boldsymbol{x},\omega_i)$ 最大者，则判别 $\boldsymbol{x}$ 属于 ω_j 类。距离分类器可以用一个简单的过程实现：

距离分类器的一般算法

■ 输入：需要识别的样本 $\boldsymbol{x}$；
■ 计算 $\boldsymbol{x}$ 与所有类别的相似度 $s(\boldsymbol{x},\omega_i)$，$i=1,\cdots,c$；
■ 输出：相似度最大的类别 ω_j。

距离分类器的实现非常简单，需要解决的关键问题是如何度量样本 $\boldsymbol{x}$ 与类别 ω_i 的相似程度，下面介绍几种最常用的样本与类别之间相似度的度量方式。

2.1.2　模板匹配

先来看一种最简单的情况，假设关于每个类别的先验知识就是一个能够代表这个类别的样本。例如我们曾经遇到过某个人，记住了这个人的长相，或者见过某种动物或植物，当再次见到这个人或这种动植物时，自然就会将其与记忆中的形象进行比对。对于分类器来

说，输入的待识别样本是一个经过特征生成和提取之后的矢量，而代表第 i 个类别的样本也可以经过同样的过程表示为 $\boldsymbol{\mu}_i$。

在每个类别只有一个代表样本的情况下，最自然的方式就是用待识别样本 $\boldsymbol{x}$ 与类别代表样本之间的相似程度作为样本与类别相似程度的度量，即 $s(\boldsymbol{x},\omega_i)=s(\boldsymbol{x},\boldsymbol{\mu}_i)$。$\boldsymbol{x}$ 和 $\boldsymbol{\mu}_i$ 均为特征矢量，可以看作是 d 维特征空间中的两个点，所以很自然地可以用两者之间的“距离”来度量相似程度，距离越近相似程度越高，距离越远相似程度越低。由于使用“距离”度量样本之间以及样本与类别之间相似程度是一种最常用的方法，因此本章介绍的分类器被称为“距离分类器”。每个类别的代表样本有时也被称为“模板”，相应的分类方法称为“模板匹配”。

严格意义下“距离”的概念将在 2.3 节讨论，目前只是将其理解为一般意义的“距离”——欧氏距离。考虑到距离越大相似度越低的因素，可以按照如下方式计算样本 $\boldsymbol{x}$ 和 $\boldsymbol{\mu}$ 之间的相似程度：

$$s(\boldsymbol{x},\boldsymbol{\mu})=-d(\boldsymbol{x},\boldsymbol{\mu})=-\|\boldsymbol{x}-\boldsymbol{\mu}\|_2=-\sqrt{\sum_{i=1}^{d}(x_i-\mu_i)^2} \tag{2.2}$$

$\|\cdot\|_2$ 在数学上称作是矢量的“l_2 范数”，这里可以理解为矢量的长度，差矢量的 l_2 范数表示的就是这两个点之间的欧氏距离。按照这样的方式定义了样本与类别之间的相似度，相应的模板匹配过程可以表示为

$$\text{如果 } j=\arg\min_{1\leqslant i\leqslant c} d(\boldsymbol{x},\boldsymbol{\mu}_i)\text{，则判别 } \boldsymbol{x}\in\omega_j$$

计算过程如图 2.1 所示。

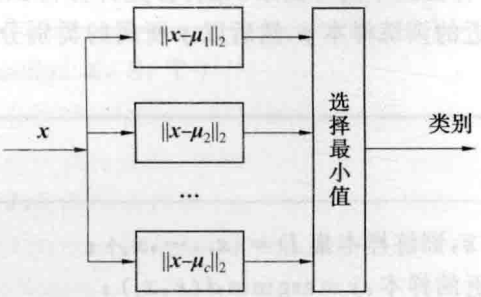


图 2.1　模板匹配的过程

以这样的方式进行识别，实际上是将特征空间划分成了 c 个区域，每个区域中的点距离该区域中的代表模板距离最近。如图 2.2，每个区域代表一个类别，如果待识别样本处于某个区域之内，则判别它属于相应的类别。两个区域的交界一般称为“判别界面”，在二维特征空间中是一条垂直平分两个类别代表样本连线的直线，在三维空间中是垂直平分的平面，而在高维空间中则被称为“超平面”。

2.1.3　最近邻分类

模板匹配用待识别样本与代表每个类别的一个样本之间的距离度量它与类别之间的相似程度。在大多数的模式识别问题中，每一个类别可以得到很多训练样本，例如在桃子和橘子的分类问题中，可以将预先手工分类好的多个水果经过特征生成和提取之后输入计算机。当每个类别存在多于一个训练样本时，如何来度量待识别样本与类别之间的相似程