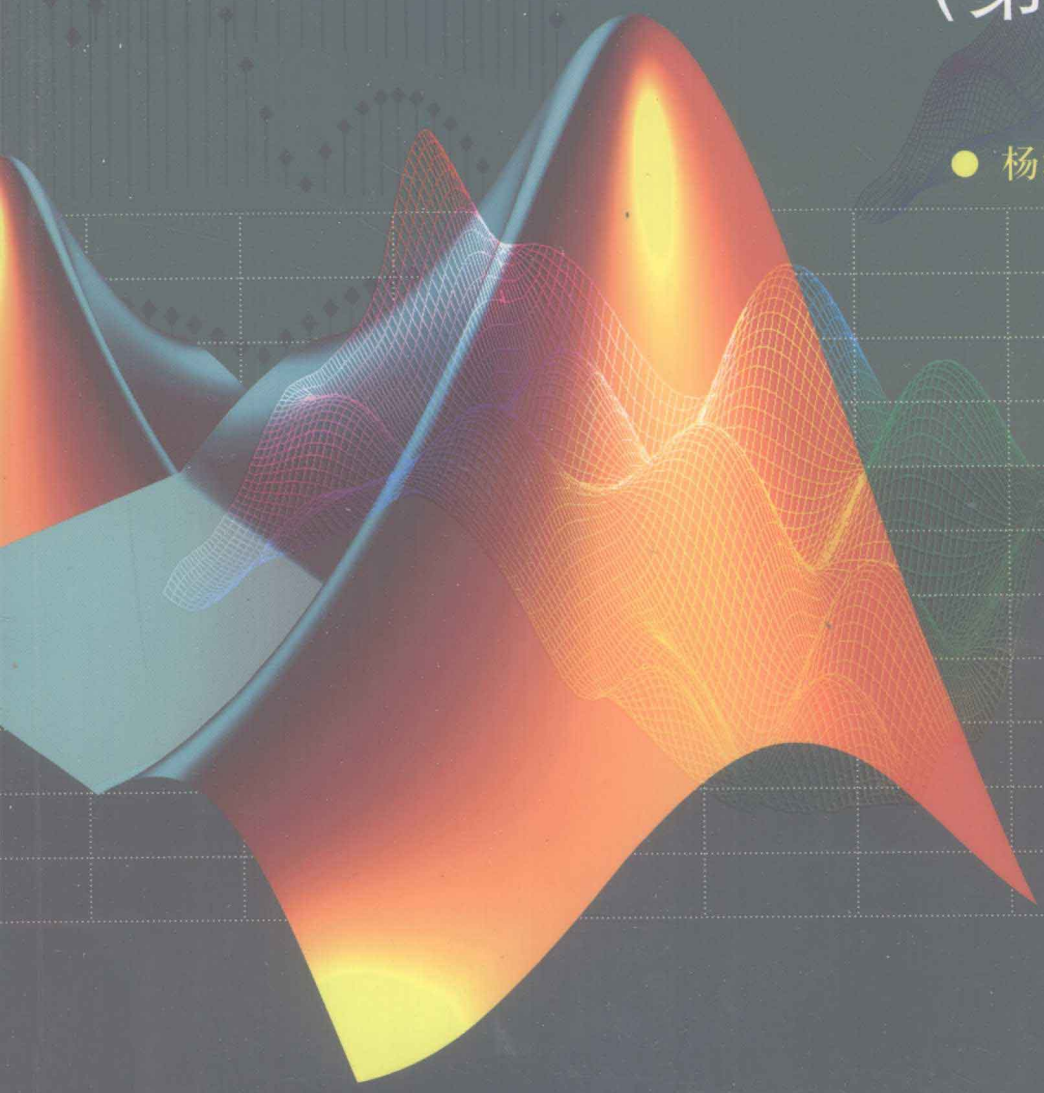
模式识别与智能计算

——Matlab技术实现

（第2版）

杨淑莹 著



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

模式识别与智能计算
——Matlab 技术实现
（第 2 版）

杨淑莹　著

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

内 容 简 介

本书广泛吸取统计学、神经网络、数据挖掘、机器学习、人工智能、群智能计算等学科的先进思想和理论，将其应用到模式识别领域中；以一种新的体系，系统、全面地介绍模式识别的理论、方法及应用。全书共分为14章，内容包括：模式识别概述，特征的选择与优化，模式相似性测度，基于概率统计的贝叶斯分类器设计，判别函数分类器设计，神经网络分类器设计（BP神经网络、径向基函数神经网络、自组织竞争神经网络、概率神经网络、对向传播神经网络、反馈型神经网络），决策树分类器设计，粗糙集分类器设计，聚类分析，模糊聚类分析，禁忌搜索算法聚类分析，遗传算法聚类分析，蚁群算法聚类分析，粒子群算法聚类分析。

本书内容新颖，实用性强，理论与实际应用密切结合，以手写数字识别为应用实例，介绍理论运用于实践的实现步骤及相应的Matlab代码，为广大研究工作者和工程技术人员对相关理论的应用提供借鉴。

本书可作为高等院校计算机工程、信息工程、生物医学工程、智能机器人学、工业自动化、模式识别等学科本科生、研究生的教材或教学参考书，也可供相关工程技术人员参考。

图书在版编目（CIP）数据

模式识别与智能计算：Matlab技术实现/杨淑莹著. —2版. —北京：电子工业出版社，2011.8
ISBN 978-7-121-14078-5

Ⅰ.①模… Ⅱ.①杨… Ⅲ.①模式识别-计算机辅助计算-软件包，MATLAB-高等学校-教材 ②人工智能-计算机辅助计算-软件包，MATLAB-高等学校-教材 Ⅳ.①O235-39 ②TP183

中国版本图书馆CIP数据核字（2011）第135589号

责任编辑：张 榕
印 刷：涿州市京南印刷厂
装 订：涿州市桃园装订有限公司
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：23.25 字数：600千字
印 次：2011年8月第1次印刷
印 数：4 000册 定价：49.00元

凡所购买电子工业出版社的图书，如有缺损问题，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。

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。

服务热线：（010）88258888。

再版前言

模式识别已经成为当代高科技研究的重要领域之一，它已发展成为一门独立的新学科。模式识别技术迅速扩展，已经应用在人工智能、机器人、系统控制、遥感数据分析、生物医学工程、军事目标识别等领域，几乎遍及各个学科领域，在国民经济、国防建设、社会发展的各个方面得到广泛应用，产生了深远的影响。

再版新书广泛吸取了统计学、神经网络、数据挖掘、机器学习、人工智能、群智能计算等学科的先进思想和理论，将其扩充到模式识别体系中。以一种新的体系，系统、全面地介绍模式识别的理论、方法及应用。全书共分为三部分，第一部分基础篇，内容包括模式识别的基本概念，特征的选择与提取，模式相似性测度。这一部分介绍模式识别的基本概念和基本方法。第二部分分类器设计篇，内容包括：贝叶斯(Bayes)分类器设计，判别函数设计，神经网络分类器设计，决策树分类器设计，粗糙集分类器设计。这一部分利用手写数字分类识别的具体实例把模式识别方法结合起来，为广大研究工作者和工程技术人员对相关理论的应用起到借鉴作用。第三部分聚类分析，内容包括基本聚类算法，模拟退火聚类分析，模糊聚类分析，禁忌搜索算法聚类分析，遗传算法聚类分析，群体智能聚类算法(蚁群算法聚类分析，粒子算法群聚类分析)。这一部分采用一幅含有需要聚类分析的图像形象生动地说明各种聚类算法。

国内外论述模式识别技术的书籍不少，但由于这一领域涉及深奥的数学理论，往往使实际工作者感到困难，而大部分书是罗列模式识别的各种算法，见不到算法的实际效果和各种算法对比的结果，而这正是学习者和实际工作者所需要了解和掌握的内容。目前还确实缺少一本关于模式识别技术在实际应用方面，具有系统性、可比性和实用性的参考书。

本书特点如下：

1. 选用新技术。除了介绍许多重要经典的内容以外，书中还包括了最近十几年来才刚刚发展起来的并被实践证明有用的新技术、新理论，比如支持向量机、BP 神经网络、RBF 神经网络、PNN 神经网络、CPN 神经网络、SORNN 神经网络、决策树、粗糙集理论、模糊集理论、模拟退火、遗传算法、蚁群算法、粒子群算法等，并将这些新技术应用于模式识别当中，提供这些新技术的实现方法和源代码。

2. 实用性强，针对实例介绍理论和技术，使理论和实践相结合，避免了空洞的理论说教。书中实例取材于手写数字模式识别，对于数字识别属于多类问题，在实际应用中具有广泛的代表性，读者对程序稍加改进，就可以应用到不同的场合，如文字识别、字符识别、图形识别等。

3. 针对每一种模式识别技术，书中分为理论基础、实现步骤、编程代码三部分。在掌握了基本理论之后，按照实现步骤的指导，可以了解算法的实现思路和方法，再进一步体会短小精悍的核心代码，学习者可以很快掌握模式识别技术，经过应用本书提供的实例程序，立刻会见到算法的实际效果。书中所有算法都用 Matlab 编程实现，便于读者学习和应用。

本书内容基本涵盖了目前“模式识别”重要的理论和方法，但并没有简单地将各种理论方法堆砌起来，而是将作者自身的研究成果和实践经验传授给读者，在介绍各种理论和方法时，

将不同算法应用于实际中，内容包括需要应用模式识别技术解决的问题，模式识别理论的讲解和推理，将理论转化为编程的步骤，计算机能够运行的源代码，计算机运行模式识别算法程序后的效果，以及不同算法应用于同一个问题的效果对比。使读者面对如此丰富的理论和方法不至于无所适从，而是有所学就会有所用。

由于至今还没有统一的、有效的可应用于所有的模式识别的理论，当前的一种普遍看法是，不存在对所有的模式识别问题都适用的单一模型和解决识别问题的单一技术，我们所要做的是把模式识别方法与具体问题结合起来，把模式识别与统计学、神经网络、数据挖掘、机器学习、人工智能、群智能计算等学科的先进思想和理论结合起来，为读者提供一个多种理论的测试平台，并在此基础上，深入掌握各种理论的效能和应用的可能性，互相取长补短，开创模式识别应用的新局面。

本书可作为高等院校计算机工程、信息工程、生物医学工程、智能机器人学、工业自动化、模式识别等学科研究生、本科生的教材或教学参考书，也可供有关工程技术人员参考。

参加本书编写的还有：邓飞、张成、王立群、任翠池、冯帆、王博凯、牛廷伟、王丽贤、王光彪、贾紫鹃等，他们在作者指导下的研究工作中付出了辛苦的劳动，取得了有益的研究成果，正是在他们的努力下本书得以顺利完成，在此表示衷心的感谢。同时，对张桦教授、徐伯夏研究员、李兰友教授给予的帮助和支持表示衷心的感谢。本书的出版得到天津理工大学出版基金的资助。由于编者业务水平和实践经验有限，书中缺点与错误在所难免，欢迎读者予以指正！

作者将不辜负广大读者的期望，努力工作，不断充实新的内容。为方便广大读者，提供了技术支持电子邮箱：ysying1262@126.com。读者可通过该邮箱及时与作者取得联系，获得技术支持。

著　　者

再版说明

《模式识别与智能计算——Matlab 技术实现》出版至今已三年多了，期间经过多次印刷，现已所剩无几。近来，应广大读者的学习要求，我决定修订再版。

这次修订，增加了局部搜索算法，即禁忌搜索算法，使本书的寻优算法涵盖了基本聚类算法、基本启发式的局部搜索和基于群体智能的全局搜索三大类算法，较第一版内容更加全面。基本聚类算法当中比较典型的有：层次聚类算法，K 均值算法和迭代自组织的数据分析、模糊聚类算法，它们采用点对点计算方式。基于模拟退火思想改进的 K 均值聚类算法和禁忌搜索算法属于启发式方法，是对局部邻域搜索扩展后的一种全局逐步寻优算法，其中模拟退火算法从单个解出发，通过扰动产生一个新的候选解，禁忌搜索算法从单个解出发产生多个新的候选解。群体智能搜索算法有遗传算法、蚁群算法和粒子群算法等，它们采用全局分布随机产生多个候选解，属于全局搜索算法。这些算法各有不同的特点，随着读者对这些算法的了解和深入研究，将它们结合起来，形成混合算法，将会避免单一算法的缺点，保证算法的收敛性，从而提高解的质量。

此外，还将全书内容进行归纳整合，将特征的选择与优化内容进行精简，压缩部分复杂内容；改进了一些分析、论断和文字表述，同时改进了部分编程代码，力求使之更为准确。

著　者

目　录

第 1 章　模式识别概述

本章要点：

☑ 模式识别的基本概念
☑ 特征空间优化设计问题
☑ 分类器设计
☑ 聚类设计
☑ 模式识别的应用

1.1　模式识别的基本概念

模式识别(Pattern Recognition) 就是机器识别、计算机识别或机器自动识别，目的在于让机器自动识别事物。例如，手写数字的识别，结果就是将手写的数字分到具体的数字类别中；智能交通管理系统的识别，就是判断是否有汽车闯红灯，闯红灯的汽车车牌号码；还有文字识别、语音识别、图像中物体识别，等等。该学科研究的内容是使机器能做以前只能由人类才能做的事，具备人所具有的对各种事物与现象进行分析、描述与判断的部分能力。模式识别是直观的、无处不在的，实际上人类在日常生活的每个环节，都从事着模式识别的活动。人和动物较容易做到模式识别，但对机器来说却是非常困难的。让机器能识别、分类，就需要研究识别的方法，这就是这门学科的任务。

模式识别研究的目的是利用计算机对物理对象进行分类，在错误概率最小的条件下，使识别的结果尽量与客观物体相符合。机器辨别事物最基本的方法是计算，原则上讲是对计算机要分析的事物与标准模板的相似程度进行计算。例如，要识别一个手写的数字，就要将它与从 0 ~ 9 的模板做比较，看跟哪个模板最相似，或最接近。因此首先要能从度量中看出不同事物之间的差异，才能分辨当前要识别的事物，因此最关键的是找到有效地度量不同类事物的差异的方法。

在模式识别学科中，就“模式”与“模式类”而言，模式类是一类事物的代表，而“模式”则是某一事物的具体体现，例如，数字 0、1、2、3、4、5、6、7、8、9 是模式类，而用户任意手写的一个数字或任意一个印刷数字则是“模式"，是数字的具体化。

1. 模式的描述方法

在模式识别技术中，被观测的每个对象称为样品，例如，在手写数字识别中，每个手写数字可以作为一个样品，如果共写了 N 个数字，我们把这 N 个数字叫做 N 个样品($X_1, X_2, \cdots, X_j, \cdots, X_N$)，其中 0 表示有 N_0 个样品，1 表示有 N_1 个样品，2 表示有 N_2 个样品，3 表示有 N_3 个样品，……，一共有 $\omega_1, \omega_2, \cdots, \omega_M (M = 10)$ 个不同的类别。

对于一个样品来说，必须确定一些与识别有关的因素，作为研究的根据，每一个因素称为一个特征。模式就是样品所具有特征的描述。模式的特征集又可用处于同一个特征空间的特

征向量表示。特征向量的每个元素称为特征,该向量也因此称为特征向量,一般我们用小写英文字母 x,y,z 来表示特征。如果一个样品 $\boldsymbol{X}$ 有 n 个特征,则可把 $\boldsymbol{X}$ 看做一个 n 维列向量,该向量 $\boldsymbol{X}$ 称为特征向量,记做

$$\boldsymbol{X}=\begin{pmatrix}x_1\\x_2\\\vdots\\x_n\end{pmatrix}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$$

若有一批样品共有 N 个,每个样品有 n 个特征,这些数值可以构成一个 n 行 N 列的矩阵,称为原始资料矩阵,如表 1-1 所示。

表 1-1　原始资料矩阵

样品 / 特征	$\boldsymbol{X}_1$	$\boldsymbol{X}_2$	…	$\boldsymbol{X}_j$	…	$\boldsymbol{X}_N$
x_1	x_{11}	x_{21}	…	x_{j1}	…	x_{N1}
x_2	x_{12}	x_{22}	…	x_{j2}	…	x_{N2}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_i	x_{1i}	x_{2i}	…	x_{ji}	…	x_{Ni}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_n	x_{1n}	x_{2n}	…	x_{jn}	…	x_{Nn}

模式识别问题就是根据 $\boldsymbol{X}$ 的 n 个特征来判别模式 $\boldsymbol{X}$ 属于 $\omega_1,\omega_2,\cdots,\omega_M$ 类中的哪一类。待识别的不同模式都在同一特征空间中考察,不同模式类由于性质上的不同,它们在各特征取值范围内有所不同,因而会在特征空间的不同区域中出现。要记住向量的运算是建立在各个分量基础之上的。

因此,模式识别系统的目标是在特征空间和解释空间之间找到一种映射关系。特征空间由从模式得到的对分类有用的度量、属性或基元构成的空间。解释空间由 M 个所属类别的集合构成。

如果一个对象的特征观测值为 $\{x_1,x_2,\cdots,x_n\}$,它可构成一个 n 维的特征向量值 $\boldsymbol{X}$,即

$$\boldsymbol{X}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$$

式中,$x_1,x_2,\cdots,x_n$ 为特征向量 $\boldsymbol{X}$ 的各个分量。

一个模式可以看做 n 维空间中的向量或点,此空间称为模式的特征空间 R^n。在模式识别过程中,要对许多具体对象进行测量,以获得许多观测值,其中有均值、方差、协方差与协方差矩阵等。

2. 模式识别系统

一个典型的模式识别系统如图 1-1 所示,由数据获取、预处理、特征提取、分类决策及分类器设计五部分组成。一般分为上下两部分:上半部分完成未知类别模式的分类;下半部分属于分类器设计的训练过程,利用样品进行训练,确定分类器的具体参数,完成分类器的设计。而分类决策在识别过程中起作用,对待识别的样品进行分类决策。

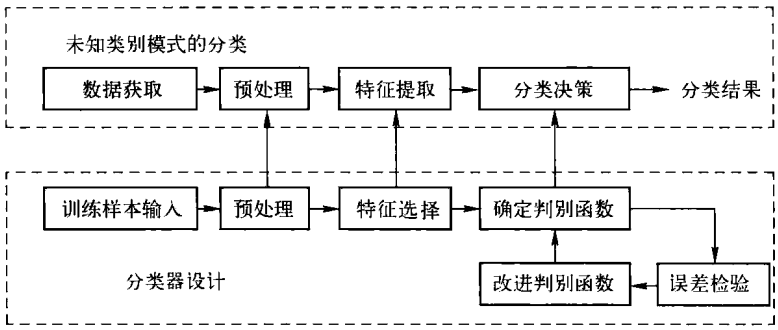


图 1-1　模式识别系统及识别过程

模式识别系统组成单元功能如下。

(1) 数据获取

用计算机可以运算的符号来表示所研究的对象,一般获取的数据类型有以下几种。

① 二维图像:文字、指纹、地图、照片等。

② 一维波形:脑电图、心电图、季节震动波形等。

③ 物理参量和逻辑值:体温、化验数据、参量正常与否的描述。

(2) 预处理

对输入测量仪器或其他因素所造成的退化现象进行复原、去噪声,提取有用信息。

(3) 特征提取和选择

对原始数据进行变换,得到最能反映分类本质的特征。将维数较高的测量空间(原始数据组成的空间)转变为维数较低的特征空间(分类识别赖以进行的空间)。

(4) 分类决策

在特征空间中用模式识别方法把被识别对象归为某一类别。

(5) 分类器设计

基本做法是在样品训练集基础上确定判别函数,改进判别函数和误差检验。

研究模式识别的主要目的是利用计算机进行模式识别,并对样本进行分类。执行模式识别的计算机系统称为模式识别系统。设计人员按需要设计模式识别系统,而该系统被用来执行模式分类的具体任务。

3. 统计模式识别研究的主要问题

统计模式识别主要研究的问题有:特征的选择与优化、分类判别、聚类判别。

(1) 特征的选择与优化

如何确定合适的特征空间是设计模式识别系统一个十分重要的问题,对特征空间进行优化有两种基本方法。一种是特征选择,如果所选用的特征空间能使同类物体分布具有紧致性,可以为分类器设计成功提供良好的基础;反之,如果不同类别的样品在该特征空间中混杂在一起,再好的设计方法也无法提高分类器的准确性。另一种是特征的组合优化,通过一种映射变换改造原特征空间,构造一个新的精简的特征空间。

(2) 分类判别

已知若干个样品的类别以及特征,例如,手写阿拉伯数字的判别是 10 个类的分类问题,机器首先要知道每个手写数字的形状特征,对同一个数字,不同的人有不同的写法,甚至同一个人对同一个数字也有多种写法,就必须让机器知道它属于哪一类。因此对分类问题需要建立样品库。根据这些样品库建立判别分类函数,这一过程由机器来实现,称为学习过程,然后对一个未知的新对象分析它的特征,决定它属于哪一类。这是一种监督学习的方法。

(3) 聚类判别

已知若干对象和它们的特征,但不知道每个对象属于哪一个类,而且事先并不知道究竟分成多少类,用某种相似性度量的方法,即"物以类聚,人以群分",把特征相同的归为一类。例如,手写了若干个阿拉伯数字,把相同的数字归为一类。这是一种非监督学习的方法。

机器识别也往往借鉴人的思维活动,像人类一样找出待识别物的外形或颜色等特征,进行分析、判断,然后加以分门别类,即识别它们。模式识别的方法很多,很难将其全部概括,也很难说哪种方法最佳,常常需要根据实际情况运用多种方法进行实验,然后选择最佳的分类方法。

1.2 特征空间优化设计问题

如何确定合适的特征空间是设计模式识别系统中一个十分重要,甚至更为关键的问题。如果所选用的特征空间能使同类物体分布具有紧致性,即各类样本能分布在该特征空间中彼此分割开的区域内,这就为分类器设计成功提供良好的基础。反之,如果不同类别的样本在该特征空间中混杂在一起,再好的设计方法也无法提高分类器的准确性。

在已有了特征的描述方法之后,也就是已有了一个初始的特征空间,需要对它进行改造,改造目的在于提高其某方面的性能,因此又称特征的优化问题。一般来说,对初始的特征空间进行优化就是为了降维,即初始的特征空间维数较高,能否改成一个维数较低的空间。优化后的特征空间应该更有利于后续的分类计算。对特征空间进行优化有两种基本方法:一种是特征选择,另一种是特征的组合优化。特征选择就是对原特征空间进行筛选,筛选掉一些次要的特征,构造出一个新的精简的特征空间,涉及对要识别的事物用什么方法进行描述和分析的问题;而特征的组合优化通过一种映射变换改造原特征空间,也就是说,新的每一个特征是原有特征的一个函数,使用变换的手段,在这里主要限定在线性变换的方法上,通过变换来实现降维。

1. 特征选择

在模式识别中特征选择是一个重要问题,直接从样品得到的数据量往往是相当大的。例如,从一个图像中可以有几十万个数据,而一个卫星云图的数据量更多。为了对样品进行准确的识别,需要进行特征选择或特征压缩。特征选择指对原始数据进行抽取,抽取那些对区别不同类别最为重要的特征,而舍去那些对分类并无多大贡献的特征,得到能反映分类本质的特征。如果把区别不同类别的特征都从输入数据中找到,这时自动模式识别问题就简化为匹配和查表,模式识别就简单多了。对一个模式类特征选择得好与坏,很难在事先完全预测,而只能针对从整个分类识别系统获得的分类结果给予评价。

对分类器设计来说，使用什么样的特征描述事物，也就是说，使用什么样的特征空间是个很重要的问题。颜色指标对区分红灯与绿灯很有效。因为前者是红色，后者是绿色，用这个指标上的差异很容易将红灯与绿灯区分开。但是如果用颜色指标区别人脸就会困难得多。换句话说，在这种情况下，这个指标就不太有效了。

特征的选择常常面临着保留哪些描述量，删除哪些描述量，通常要经过从多到少的过程，因为在设计识别方案的初期阶段，应该尽量多地列举出各种可能与分类有关的特征，这样可以充分利用各种有用的信息，改善分类效果。但大量的特征中肯定会包含许多彼此相关的因素，造成特征的重复和浪费，给计算带来困难。Kanal. L 曾经总结过经验：样品数 N 与特征数 n 之比应足够大，通常样本数 N 是特征数 n 的 5 ~ 10 倍。为了使特征数从多变少，需要进行特征选择，特征选择通常包括两方面内容：一方面是对单个特征的选择，即对每个特征分别进行评价，从中找出对识别作用最大的那些特征；另一方面是从大量的原有特征出发构造出少数的有效新特征，这种方法称为降维映射。

对一个具体问题来说，有以下两个不同的层次。

（1）物理量的获取与转换

这是指用什么样的传感器获取电信号，对从传感器中得到的信号，可以称为原始信息，因为它要经过加工、处理才能得到对模式分类更加有用的信号，如摄取景物要用到摄像机。文字与数字识别首先要用扫描仪等设备。手写体文字所用传感器与印刷体文字也很可能不同。这些都属于物理量的获取，并且已转换成电信号，为计算机分析打下基础。

（2）描述事物方法的选择与设计

在得到了原始信息之后，必须对原始信息进行加工，以获取对分类最有效的信息。设计所要信息的形式是十分关键的。例如，对数字的识别特征提取可以有多种方法，有的分析从框架的左边框到数字之间的距离变化反映了不同数字的不同形状，这可以用来作为数字分类的依据。另外一种方法是在每个数字图形上提取特征值，定义一个 $N \times N$ 模板，在本书实例程序中设定 $N = 5$，将每个样品的长度和宽度 5 等分，平均有 25 个等份，构成一个 5×5 模板。对每一份内的像素个数进行统计，再除以每一份的面积总数，即得特征初值，将特征初值大于 20% 所对应的模板置为 1，取得该数字对应的特征。

对事物的描述方法是充分发挥设计者智慧的过程，这个层次的工作往往因事物而异，与设计者本人的知识结构也有关。这是一个目前还无法自动进行的过程。这个层次的工作是最关键的，但因为太缺乏共性，也不是本书讨论的内容。

2. 特征优化

假设已有 D 维特征向量空间，$\boldsymbol{Y} = \{y_1, y_2, \cdots, y_D\}$，特征的组合优化问题涉及到特征选择和特征提取两部分。特征选择是指从原有的 D 维特征空间，删去一些特征描述量，从而得到精简后的特征空间。在这个特征空间中，样本由 n 维的特征向量描述：$\boldsymbol{X} = \{x_1, x_2, \cdots, x_n\}$，$n < D$。由于 $\boldsymbol{X}$ 只是 $\boldsymbol{Y}$ 的一个子集，因此每个分量 x_i 必然能在原特征集中找到其对应的描述量 $x_i = y_j$。

特征优化则是找到一个映射关系：

$$A: \boldsymbol{Y} \rightarrow \boldsymbol{X}$$

使样本新特征描述维数比原维数低。其中每个分量 x_i 是原特征向量各分量的函数，即

$$x_i = A(y_1, y_2, \cdots, y_D)$$

因此这两种降维的基本方法是不同的。在实际应用中可将两者结合起来使用,例如,先进行特征选择,指从原有的 D 维特征空间,删去一些特征描述量,从而得到精简后的特征空间,然后再进一步进行特征优化,或反过来操作。

要对原特征空间进行优化,就要对优化的结果进行评价,在实际应用中经常采用的评价方法是对系统性能进行测试。最主要的测试指标是识别正确率,其他指标还有识别计算速度、存储容量等。需要有定量分析比较的方法,判断所得到的特征维数及所使用特征是否对分类最有利,这种用以定量检验分类性能的准则称为类别可分离性判据。为此,人们设法从另一些更直观的方法出发,设计类别可分离性判据,用来检验不同的特征组合对分类性能好坏的影响,甚至用来导出特征选择与特征提取的方法。对特征空间进行优化是一种计算过程,它的基本方法仍然是模式识别的典型方法,即找到一种准则(或称判据),通常用一种式子表示,以及计算出一种优化方法,使这种计算准则达到一个极值。

总之,特征选择与特征优化的任务是求出一组对分类最有效的特征。有效是指在特征维数减少到同等水平时,其分类性能最佳。

1.3　分类器设计

模式识别分类问题是指根据待识别对象所呈现的观测值,将其分到某个类别中去。具体步骤是建立特征空间中的训练集,已知训练集里每个点所属的类别,从这些条件出发,寻求某种判别函数或判别准则,设计判决函数模型,然后根据训练集中的样品确定模型中的参数,便可将这模型用于判别,利用判别函数或判别准则去判别每个未知类别的点应该属于哪一个类。

如何做出合理的判决就是模式识别分类器要讨论的问题。在统计模式识别中,感兴趣的主要问题并不是决策正误,而在于如何使决策错误造成的分类误差在整个识别过程中的风险代价达到最小。模式识别算法的设计都是强调“最佳”与“最优”,即希望所设计的系统在性能上最优。这种最优是针对某一种设计原则讲的,这种原则称为准则,常用的准则有最小错误率准则、最小风险准则、近邻准则、Fisher 准则、均方误差最小准则、感知准则等。设计准则,并使该准则达到最优的条件是设计模式识别系统最基本的方法。模式识别中以确定准则函数来实现优化的计算框架。分类器设计使用哪种原则是关键,会影响到分类器的效果。不同的决策规则反映了分类器设计者的不同考虑,对决策结果有不同的影响。分类决策在识别过程中起作用,对待识别的样品进行分类决策。

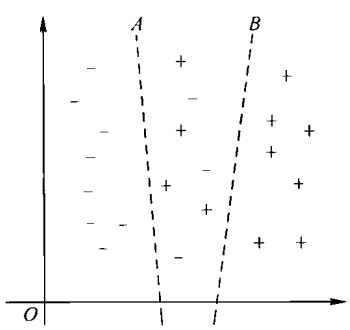


图 1-2　分界线示意图

一般来说,M 类不同的物体应该具有各不相同的属性值,在 n 维特征空间中,各自有不同的分布。当某一特征向量值 $\boldsymbol{X}$ 只为某一类物体所特有,对其做出决策是容易的,也不会出什么差错。问题在于常常会出现模棱两可的情况。由于属于不同类的待识别对象存在着呈现相同特征值的可能,即所观测到的某一样品的特征向量为 $\boldsymbol{X}$,而在 M 类中又有不止一类可能呈现这一 $\boldsymbol{X}$ 值,如图 1-2 所示,A、B 直线之间的样品属于不同类别,但是它们具有相同的特征值。例如,癌症病人初期症状与正常人的症状相同,这两个类别样品分别用“-”与“+”表示。从图中可见这两类样品在二维特征空间中相互穿插,很难用简