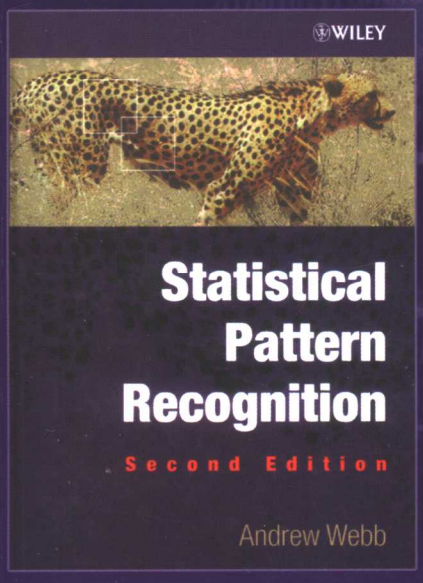
国外计算机科学教材系列

统计模式识别

（第二版）

Statistical Pattern Recognition

Second Edition



[英] Andrew R. Webb 著

王 萍 杨培龙 罗颖昕 译

WILEY

電子工業出版社
Publishing House of Electronics Industry
http://www.phei.com.cn

内容简介

本书对统计模式识别的基本理论和技术做了全面且详尽的介绍。包括用于分类器设计的重要方法和用于数据分析和预处理的关键技术。前者有基于概率密度函数估计的参数法和非参数法，基于判别函数构建的线性模型、径向基函数网络、支持向量机、投影方法（神经网络）和判别分析决策树等；后者涉及特征选择和特征提取以及聚类分析。

此外，本书还就分类器的特性测评和利用分类器的组合技术改进分类器特性等进行了较充分的讨论。并且，对模型选择、不可靠分类、缺值数据、离群值检测、连续变量与离散变量的混合等问题进行了探讨。

本书论述简明清楚、概念明确，应用实例涉及广泛、启发性强，是从事模式识别研究和应用工作的重要参考用书，也可以作为信息类研究生课程的教材。

Andrew R. Webb: **Statistical Pattern Recognition, Second Edition.**

ISBN 0-470-84513-9

版权贸易合同登记号　图字：01-2003-2429

图书在版编目（CIP）数据

统计模式识别（第二版）/（英）韦布（Webb, A. R.）著；王萍等译. - 北京：电子工业出版社，2004.10
（国外计算机科学教材系列）
书名原文：Statistical Pattern Recognition, Second Edition
ISBN 7-121-00432-1

Ⅰ.统… Ⅱ.①韦… ②王… Ⅲ.统计模式识别 - 教材 Ⅳ.0235

中国版本图书馆CIP数据核字（2004）第102640号

责任编辑：史　平
印　　刷：北京兴华印刷厂
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
经　　销：各地新华书店
开　　本：787 × 1092　1/16　　印张：25　字数：608千字
印　　次：2004年10月第1次印刷
定　　价：45.00元

出 版 说 明

21世纪初的5至10年是我国国民经济和社会发展的重要时期，也是信息产业快速发展的关键时期。在我国加入WTO后的今天，培养一支适应国际化竞争的一流IT人才队伍是我国高等教育的重要任务之一。信息科学和技术方面人才的优劣与多寡，是我国面对国际竞争时成败的关键因素。

当前，正值我国高等教育特别是信息科学领域的教育调整、变革的重大时期，为使我国教育体制与国际化接轨，有条件的高等院校正在为某些信息学科和技术课程使用国外优秀教材和优秀原版教材，以使我国在计算机教学上尽快赶上国际先进水平。

电子工业出版社秉承多年来引进国外优秀图书的经验，翻译出版了“国外计算机科学教材系列”丛书，这套教材覆盖学科范围广、领域宽、层次多，既有本科专业课程教材，也有研究生课程教材，以适应不同院系、不同专业、不同层次的师生对教材的需求，广大师生可自由选择和自由组合使用。这些教材涉及的学科方向包括网络与通信、操作系统、计算机组织与结构、算法与数据结构、数据库与信息处理、编程语言、图形图像与多媒体、软件工程等。同时，我们也适当引进了一些优秀英文原版教材，本着翻译版本和英文原版并重的原则，对重点图书既提供英文原版又提供相应的翻译版本。

在图书选题上，我们大都选择国外著名出版公司出版的高校教材，如Pearson Education培生教育出版集团、麦格劳－希尔教育出版集团、麻省理工学院出版社、剑桥大学出版社等。撰写教材的许多作者都是蜚声世界的教授、学者，如道格拉斯·科默（Douglas E. Comer）、威廉·斯托林斯（William Stallings）、哈维·戴特尔（Harvey M. Deitel）、尤利斯·布莱克（Uyless Black）等。

为确保教材的选题质量和翻译质量，我们约请了清华大学、北京大学、北京航空航天大学、复旦大学、上海交通大学、南京大学、浙江大学、哈尔滨工业大学、华中科技大学、西安交通大学、国防科学技术大学、解放军理工大学等著名高校的教授和骨干教师参与了本系列教材的选题、翻译和审校工作。他们中既有讲授同类教材的骨干教师、博士，也有积累了几十年教学经验的老教授和博士生导师。

在该系列教材的选题、翻译和编辑加工过程中，为提高教材质量，我们做了大量细致的工作，包括对所选教材进行全面论证；选择编辑时力求达到专业对口；对排版、印制质量进行严格把关。对于英文教材中出现的错误，我们通过与作者联络和网上下载勘误表等方式，逐一进行了修订。

此外，我们还将与国外著名出版公司合作，提供一些教材的教学支持资料，希望能为授课老师提供帮助。今后，我们将继续加强与各高校教师的密切联系，为广大师生引进更多的国外优秀教材和参考书，为我国计算机科学教学体系与国际教学体系的接轨做出努力。

电子工业出版社

教材出版委员会

译 者 序

信息时代，无处不存在对模式识别的需求。概括地讲，模式识别是一门以应用数学为理论基础，利用计算机应用技术，解决实际分类及识别问题的学问。按照研究问题的特点及解决问题的手段特征，通常有统计模式识别和结构模式识别之分。前者以多元统计理论为数学基础，以数据特征的形式对问题进行描述；而后者则以形式语言为数学基础，以结构图元的形式对问题进行描述。它们都致力于将隐含在大量样本中的类间差异的规律归纳出来，并综合成适当的分类、识别乃至预测模型。

从发展的角度来看，在传统的、较成熟的分类和识别的基础上，模糊数学思想方法的介入、人工神经网络对统计模型类型的丰富、进化算法等一批优秀算法的出现、支持向量机等一些新方法的提出，等等，使统计模式识别的研究和应用充满活力。

英国著名学者 Andrew R. Webb 所著《统计模式识别》一书对统计模式识别领域进行了全面介绍，并在以下方面具有鲜明特点：

编写体系：本书以“分类与识别”为主线，在“基本概念 - 理论分析 - 方法讲解 - 应用实例 - 更多的研究”的框架下，介绍统计模式识别的每一个具体方法；再以应用研究、建议、参考文献等，对由若干方法形成的一类问题进行综述。其中，“更多的研究”能够使读者从知识点伸展到面，进一步了解相关问题的研究动态及人们普遍关注的问题。而“应用研究”则将模式识别技术与广泛的实际问题紧密相连，颇具启迪性。“总结”及“建议”凝结了作者的体会和经验，很有指导性。“参考文献”给出了所列文献与书中内容的联系及其特色。这样的组织格局使读者从局部到全局、从理论到方法、从方法到应用、从研究动态到问题展望，一览无余。

将最新研究方法融入统计模式识别框架：作者在“分类与识别”主线下带出对统计模式识别新概念、新方法（例如人工神经网、模糊思想用于聚类、支持向量机、新的非参数方法等）较详尽的介绍。使人们能够更深层次地理解它们的构成内涵和其识别行为属性，从而为根据具体问题特点灵活、合理地选用它们提供帮助。

内容前后呼应：作者在保持各章节内容相对独立的前提下，特别加强了“谈此及彼”的特色，使读者能够对一种重要方法进行多角度的理解和消化。

辩证评述和比较性研究：模式识别问题本身决定了目前使用的模式识别方法和技术没有绝对的好与坏。相信读者会从本书的字里行间领略到作者科学严禁的理论分析及辩证客观的方法评述，并从中受益。另外，本书特别强调并略加笔墨的“比较性研究”近年来受到模式识别学者和专家的重视，值得读者关注。

参加本书翻译工作的有王萍、杨培龙、罗颖昕，并由王萍统稿。由于译者水平有限，译文中难免有疏漏和不妥之处，恳请读者不吝赐教。

前　言

本书介绍统计模式识别的基本理论和技术，其中大部分内容涉及到识别和分类问题，取材于工程学、统计学、计算机科学和社会学等领域的相关文献。在这些文献中，反映了许多当今最有用的模式识别技术和最新的非参数识别方法，本书一并对它们做出简明的介绍，并对各项技术附以应用研究实例来说明。至于书中涉及的模式识别的应用、对比研究法及理论推导的细节可以在书后各类文献中找到。

统计模式识别是一个非常活跃的研究领域，它在近年来的许多进展得益于计算机不断增长的计算能力，并使得一些技术的应用范围得到拓展。本书的大部分章节简述了这些较宽范围的实际应用和较深层次的推理技术。

本书为模式识别领域（多学科的技术综合可以成为领域）的工作人员及研究者编写，部分内容也可以作为信息类研究生课程的教材。读者应具备概率论和线性代数的基本知识及一些基本数学方法（例如，解决具有等式约束和不等式约束的拉格朗日数乘法），附录中给出了它们的最基本内容。每章后面所附习题从一目了然的问题到较为复杂的工程性问题都有所涉及。

第 1 章作为统计模式识别的绪论，给出了一些名词术语的定义，并介绍了监督型分类和非监督型分类。就监督型分类而言，有两种研究方法，其一是基于概率密度函数的估计，其二则是基于判别函数的构建。在这一章的最后对模式识别的完整过程进行了概括，细节问题则安排在后续章节中讨论。第 2 章和第 3 章讨论了识别问题的密度函数法。其中，第 2 章讲解密度函数的参数估计，第 3 章推导非参数分类器。

第 4 章到第 7 章研究监督型分类问题的判别函数的构建方法。第 4 章集中讨论线性判别函数，其中涉及的大多数判别法（包括优化、正则化和支持向量机）也适用于非线性研究。第 5 章探讨基于核函数的方法。特别地，径向基函数网络和支持向量机这些用于判别和回归的技术近年来受到普遍关注。第 6 章介绍非线性模型（基于投影的方法）。第 7 章讨论一种判别分析决策树，涉及分类回归树法（CART）和多元自适应回归样条函数（MARS）。

第 8 章讨论分类器特性，包括分类器的特性测评和利用分类器的组合技术改进分类器的特性等。

第 9 章和第 10 章探讨数据分析和预处理技术（这些工作通常先于第 2 章到第 7 章介绍的有监督分类工作）。第 9 章讲述特征选择和特征提取。通过特征选择和特征提取，可以降低描述原始数据特征的维数。这项工作通常是分类器整体设计工作的一部分，只是人为地将它们（特征提取和模式分类）划分为相对独立的过程。特征提取可以帮助我们深入地了解数据的结构，以及分类器需要选用的类型，因此该项研究备受关注。第 10 章讲述非监督分类——聚类，即在样本群中找到不同的结构并借此将其划分成独立部分的过程。

最后一章即第 11 章，介绍其他重要的研究课题。附录覆盖了更大范围的背景材料。如果选用本书作为教科书，则应先研读关于相异测度、估计、线性代数、数据分析和基础概率的相关课程。

本书相关网站 www.statistical-pattern-recognition.net 中包含在技术和应用方面更深入信息的参考资料和链接。

在编写本书第二版的过程中，得到了很多人的帮助。很感谢我的同事和朋友们，他们对原稿的不同部分给予了许多宝贵意见。这里，我要特别感谢 Mark Briers，Keith Copsey，Stephen Luttrell，John O'Loghlen 和 Kevin Weekes（尤其感谢 Keith 提供的第 2 章中的例子）；感谢 Wiley 出版社为原稿的出版所做的努力；更要特别感谢 Rosemary 的关心与支持。

目　录

符　号

下面列出了一些常用的符号，并使用国际惯例进行标注。例如，对于变量和变量上的观测值，倾向于使用同一符号，它们的不同意思可以从上下文中明显看出来。而且，将 x 的密度函数记为 $p(x)$，将 y 的密度函数记为 $p(y)$，即使这两个函数并不相同。用粗体的小写字符表示向量，而用粗体的大写字符表示矩阵。

p	变量个数
C	类别数
n	观测值个数
n_i	类 i 中的观测值个数
ω_i	类 i 的标记
$X_1, \cdots, X_p$	p 个随机变量
$x_1, \cdots, x_p$	变量 $X_1, \cdots, X_p$ 上的观测值
$\boldsymbol{x} = (x_1, \cdots, x_p)^T$	观测值向量
$\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]^T$	$n \times p$ 阶数据矩阵
$\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$	
$P(\boldsymbol{x}) = \text{prob}(X_1 \leqslant x_1, \cdots, X_p \leqslant x_p)$	
$p(\boldsymbol{x}) = \partial P / \partial \boldsymbol{x}$	
$p(\omega_i)$	类 i 的先验概率
$\boldsymbol{\mu} = \int \boldsymbol{x} p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$	总体均值
$\boldsymbol{\mu}_i = \int \boldsymbol{x} p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$	类 i 的均值，$i = 1, \cdots, C$
$\boldsymbol{m} = (1/n) \sum_{r=1}^{n} \boldsymbol{x}_r$	样本均值
$\boldsymbol{m}_i = (1/n_i) \sum_{r=1}^{n} z_{ir} \boldsymbol{x}_r$	类 i 的样本均值，$i = 1, \cdots, C$ 若 $\boldsymbol{x}_r \in \omega_i$，则 $z_{ir} = 1$，否则为 0 n_i 为 $\omega_i = \sum_{r=1}^{n} z_{ir}$ 中的模式数量
$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{r=1}^{n} (\boldsymbol{x}_r - \boldsymbol{m})(\boldsymbol{x}_r - \boldsymbol{m})^T$	样本协方差矩阵(极大似然估计)
$n/(n-1)\hat{\boldsymbol{\Sigma}}$	样本协方差矩阵(无偏估计)
$\hat{\boldsymbol{\Sigma}}_i = (1/n_i) \sum_{j=1}^{n} z_{ij} (\boldsymbol{x}_j - \boldsymbol{m}_i)(\boldsymbol{x}_j - \boldsymbol{m}_i)^T$	类 i 的样本协方差矩阵(极大似然估计)
$\boldsymbol{S}_i = \frac{n_i}{n_i - 1} \hat{\boldsymbol{\Sigma}}_i$	类 i 的样本协方差矩阵(无偏估计)

$\boldsymbol{S}_W = \sum_{i=1}^{C} \frac{n_i}{n} \hat{\boldsymbol{\Sigma}}_i$	合并类内样本协方差矩阵
$\boldsymbol{S} = \frac{n}{n-C}\boldsymbol{S}_W$	合并类内样本协方差矩阵(无偏估计)
$\boldsymbol{S}_B = \sum_{i=1}^{C} \frac{n_i}{n}(\boldsymbol{m}_i - \boldsymbol{m})(\boldsymbol{m}_i - \boldsymbol{m})^T$	样本类间矩阵
$\boldsymbol{S}_B + \boldsymbol{S}_W = \hat{\boldsymbol{\Sigma}}$	
$\| \boldsymbol{A} \|^2 \sum_{ij} A_{ij}^2$	
$N(\boldsymbol{m}, \boldsymbol{\Sigma})$	均值为 $\boldsymbol{m}$、协方差矩阵为 $\boldsymbol{\Sigma}$ 的正态分布
$\mathrm{E}[Y \mid X]$	给定 X 时 Y 的期望
$\mathrm{I}(\theta)$	若 θ 为真,其值为 1,否则为 0

附录 E 给出了特定概率密度函数的符号。

第1章　统计模式识别概论

统计模式识别一词概括了从问题描述、数据采集到识别分类、结果评价及解释的各个阶段。本章介绍其中的基本术语和两种互补的识别方法。

1.1　统计模式识别

1.1.1　引言

本书论述了模式识别的基本方法及实际应用技术，重点阐述用于识别的统计理论，同时关注聚类问题。因此，本书的主题可以概括为“分类”，其中包括有监督分类和无监督分类。前者利用分类信息设计分类器(识别)，而后者则是在未知分类信息前提下的分类(聚类)。

模式识别作为一个研究领域，迅速发展于20世纪60年代。它是一个多领域的交叉学科，该学科涉及到统计学、工程学、人工智能、计算机科学、心理学和生理学等。许多人为了解决实际问题进入该领域。其中包括字符自动识别、医疗诊断等经典问题和个人信用评分、商品销售分析、信用卡交易分析等关于数据挖掘的新问题。如此广泛的模式识别应用，吸引了众多的研究力量，产生出许多新的方法，推动着该学科的进一步发展。而能在一定程度上仿效人类行为的智能机器的发展，激发了另外一些人对人工智能的研究兴趣。在人工智能的研究中，出现过一些过于乐观的观点和不切实际的言论，并于20世纪70年代和20世纪80年代，在一定程度上，先后出现了基于知识的系统(knowledge-based system)和神经网络(neural network)竞相发展的局面。

在上述领域，尤其是在和概率与统计相交叠的领域已取得重大进展的前提下，近年来又出现了许多令人振奋的方法学和应用两个方面的新进展。这些(如核函数方法和贝叶斯计算方法)均得益于早期研究所形成的牢固基础和如今能够容易得到且日益强大的计算方法。

机器学习是研究如何使机器适应环境和通过事例进行学习的一门学科。本书中的论题可以归于机器学习的范畴。尽管机器学习更多地把重点放在计算的精深方法而不是统计方法上，但两者(统计模式识别和机器学习)还是有着许多共同的研究领域。

1.1.2　基本模型

鉴于模式识别的许多技术涵盖了多个学科的发展，自然会出现不同学科对相同术语不同的甚至相反界定的情形。在此，我们将“模式(样本)”表示为 p 维数据向量 $\boldsymbol{x}=(x_1,\cdots,x_p)^T$。其中，$x$ 表示被分类对象，p 表示用于分类的特征变量的数量，T 表示向量的转置，x_i 表示第 i 个特征变量的观测值。若识别问题含有 C 个类，记为 $\omega_1,\cdots,\omega_C$，则关于每一个模式 $\boldsymbol{x}$ 的分类变量记为 z，z 表示 x 的类别，即若 $z=i$，则模式 $\boldsymbol{x}$ 属于 ω_i，$i\in\{1,\cdots,C\}$。

在语音识别中对声波的测量结果、为确定疾病类型对病人进行的检测结果(诊断)、为预测可能的病情发展对病人进行的检测结果(预后)、对气候参数的测量(天气预报)以及字符识别

中用到的数字化图像等都是上面所说的模式(样本)。可以看出,“模式”一词的技术意义不一定涉及到图像中的结构。

本书内容围绕“分类器设计”、“识别模式(样本)”以及“制定分类规则”等几个主题展开。图 1.1 给出了模式分类器的示意图。一旦分类器的参数确定下来,分类器便能对给定样本产生某种意义下的最佳响应,该响应通常是对样本所属类别的估计。一组类别属性已知的样本 $\{(\boldsymbol{x}_i, z_i), i=1,\cdots,n\}$ 可以作为分类器的训练集或设计集,分类器的设计就是用训练集确定分类器的内部参数。由此形成的分类器可用于估计未知样本 $\boldsymbol{x}$ 的类别属性。

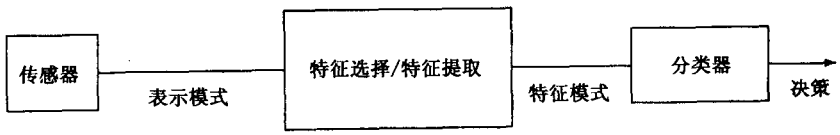


图 1.1 模式分类器

分类器的形式取决于多种因素。其一为训练数据的分布以及对数据分布所做的假设。其二是错分代价——做出错误决策的代价。在很多实际应用中,错分代价包括金钱、时间和其他主观性判断等多个方面,是很难量化的。例如,在医疗诊断中,每一种疗法都有相应的代价,这时的错分代价会涉及到不同药物的费用、每个疗程中病人所受的痛苦以及产生并发症的可能性等方面。

图 1.1 大致勾画出了模式分类的过程。其间,需要经过几个独立的数据变换阶段。这些变换有时也叫做预处理、特征选择或特征提取,变换的结果通常导致模式维数的降低(减少特征数)、多余的和不相关信息的剔除,并将数据模式转换成更适合于后续分类工作所需的形式。本征维数(intrinsic dimensionality)指的是捕获数据结构所需的最少变量数。以上面提到的语音识别为例,首先是将语音波形变换到频率域,再找到共振峰(频谱中的峰值),并进一步形成特征。这是一个特征提取的过程(用原始变量的非线性组合形成新的变量)。而特征选择是在给定的一组变量中选择一个子集的过程。

不同的作者可能会使用不同的术语。在此,表示模式(representation pattern)一词指的是由传感器(如摄像机、雷达)测得的向量,而特征模式(feature pattern)则用来表示测得的原始向量经变换(特征选择或特征提取)后形成的数量较少的一组变量。对某些问题来讲,特征向量可以经测量直接得到,这时,不再需要特征的自动选择。而特征选择也是由研究人员来完成的,研究人员通过经验、已掌握的知识,以及对问题域的认识来决定哪些变量对分类有用。然而,在大多数情况下,通常需要对测得的数据进行一次或多次变换。

有些模式分类器需要执行以上提及的每一步,这些步骤的操作相互独立。而有些模式分类器却不完全如此。另外,某些分类器需要针对特殊问题(如语音识别问题)对数据进行预处理。本书中所讨论的特征选择和特征提取并不针对某个特殊应用,但这并不是说,对任何应用都适合,实际上是指特殊应用的预处理工作应该留给研究人员来完成。

1.2 解决模式识别问题的步骤

以下列出了模式识别研究工作所包括的若干个步骤(附录 D 中给出了更多的细节),但并

不是全部。其中,有些步骤被合并致使各步之间的操作差异不甚明显,而有些步骤因仅适用于特殊应用的数据处理而未被列出。下面几点是相当典型的:

1. 问题表述:准确理解研究的目的,并对下一步工作做出计划。
2. 数据采集:对相关变量进行测量,并详细记录数据采集的过程(基本事实)。
3. 数据的初始检查:核对数据,计算总体统计量并绘出曲线图以获得对数据结构的感性认识。
4. 特征选择或特征提取:从测量集中选出最有利于分类的一组变量。这些新的变量可以通过对原始变量集的线性或非线性变换(特征提取)得到。从某种意义上说,特征提取和分类的区别是人为所致。
5. 非监督模式分类或聚类:可以看做是探索性的数据分析,由此可以为研究科目提供有用的结论。另一方面,它也可以是有监督分类过程预处理数据的一种方法。
6. 应用适当的识别或回归方法:用训练样本集设计分类器。
7. 结果评估:包括将训练出的分类器用于独立的有标签样本的考试集。
8. 解释说明。

以上步骤是一个反复的过程:分析结果有可能提出更多的新假设,而验证这些假设又需要进一步的数据采集。而且,这一循环可能终止于不同的步骤:所提问题或许仅靠对数据的初次测试便告解决,或许后来又发现这些数据根本不能解决原问题而需要对其进行重新描述。

本书重点介绍步骤 4,5,6 中的相关技术。

1.3　问题讨论

本书的主题是关于分类器的设计,即给定模式类别已知的训练集,设计一个分类器,使这个分类器对期望的工作条件(测试条件)来说是最优的。

以上所述看似简单明了,实际内含若干要点。首先,用于设计分类器的训练集样本数有限。这时,若选取内含若干自由参数的,过于复杂的分类器形式,则分类器可能会因对设计集中的噪声模拟而引发过度拟合。如果分类器不够复杂,则又无法捕获数据中的结构。比如,我们用多项式曲线来拟合一组数据点。如果多项式的次数过高,尽管相关曲线通过或靠近数据点,从而获得较低的拟合错误,但噪声却使拟合曲线极不稳定且模型易波动。如果多项式次数过低,则拟合错误较大,因而不能模拟曲线内在的变化规律。

于是,不一定追求训练集的最小错误准则下的最优性能。也就是说,在分类问题中,实现对设计集 100%的分类准确率是可能的,但其一般化性能,即在真实工作环境下,数据所表现出的预期性能(在无限的考试集上表现出来的性能)比经过仔细设计得到的性能要差。因此,学习如何选择合适的模型是重要的。

实际上,数据中所蕴含的结构及噪声通常是未知的。不应该把训练分类器(决定其参数的过程)看做是与模型选择相独立的问题,尽管人们经常这样认为。

其次是最优分类器设计中的"最优"问题。用于计量分类器性能的方法有多种,最常见的办法是计量分类器的错误率,这种方法有一些局限性。另一种办法是计算类的估计概率与类的实际概率的接近度,这种方法更适用于大多数场合。然而,由于期望的标准很难直接达到最

优,因此人们更多地选择另外的可供代替的评价标准优化分类器设计。例如,用平方误差最优法训练分类器,而用误差率评价分类器。

最后,通常假定训练数据能够代表测试环境。否则,或者环境受到了噪声的干扰而训练数据没能表现出来,或者抽取数据的总体发生了变化(总体训练),这些必须在设计分类器时加以考虑。

1.4 有监督分类和无监督分类

有监督分类(或识别)和无监督分类(在某些统计学文献中有时将它们简单地叫做分类和聚类)是两种主要的分类划分。

在有监督分类中,有一组带有类别标签的数据样本,其中每一个数据样本实际上就是对一组变量的测量值,这些就是设计分类器时要用到的样本。

为什么需要设计一个能够对未知的数据进行自动分类的方法呢?能将标定设计集类别的方法等同地运用于考试集吗?在某些情况下回答是肯定的。但即便如此,我们仍然希望研制出一种自动的方法以减少劳力密集性过程。而在另外一些情况下,人不可能参与分类。前一种情况的例子是工业检测,操作人员将标识图仔细地贴到生产线上的被检物上,分检机靠标识图工作。在实际应用中,我们希望将人类从繁重枯燥的劳动中解放出来,同时希望分检机的工作更可靠。后一种情况的例子是雷达探测目标中识别物的自动分类。雷达的机械装置定位于转盘上,通过各方向角的测量获得数据。这时,人们没有能力从雷达图中可靠地识别目标,也不能进行远距离处理。

在无监督分类中,数据的类别标识未知。我们试图找到数据所属的类别,以及类与类相区别的特征。第 10 章讲到的聚类技术也可用做有监督分类的一部分。即将聚类方案独立地应用于每个类,再把类中每个聚类的代表性样本(如聚类均值)作为类的原型。

1.5 研究统计模式识别问题的方法

本书讲述的主要问题是模式分类。给定一组以模式向量 $\boldsymbol{x}$ 表示的观测值,希望将其归于 C 个可能类的某一类 $\omega_i, i=1,\cdots,C$ 中,决策规则将测量空间划分成 C 个区域 $\Omega_i, i=1,\cdots,C$。如果观测向量位于 Ω_i,则假定它属于类 ω_i。每个区域都有可能是多连通的,也就是说,每个区域都可能由几个分离的部分组成。区域 Ω_i 之间的分界是决策边界或决策面。一般来讲,靠近区域边界处是最易发生分类错误的地方,这时,除非获得了进一步的分类信息,否则都不对该模式做出决策。这一做法称做拒绝选择,因此,C 类问题存在 $C+1$ 个决策结果,其中拒绝域记为 ω_0。

本节介绍了两种主要的识别方法,这两种方法还会在以后的章节中进行深入研究。第一个方法是采用潜在的类条件概率密度函数(给定类的特征向量的概率密度函数)的知识进行识别。当然,这些概率密度在许多实际应用中是未知的,因而需要根据已获得正确分类的样本集(称为设计集或训练集)对其进行估计。第 2 章和第 3 章清晰地讲述了估计概率密度函数的若干技术。

第二种方法研究各种决策规则,这些规则直接使用数据估计决策边界,而不需要计算概率密度函数。第 4 章、第 5 章和第 6 章具体讨论了这种方法的相关技术。