统计模式识别

STATISTICAL PATTERN RECOGNITION

陈季镐 著 邱焕章 邱华 译 姜崇熙 校译

北京邮电学院出版社

统计模式识别

〔美〕陈季镐 著
邱焕章 邱 华 译
姜崇熙 校译

北京邮电学院出版社

内 容 简 介

本书主要讨论统计模式识别的理论和方法，包括线性和非线性分类理论、特征选择和抽取、监督和非监督参数估计、非参数方法和复合判决理论、聚类方法、序贯模式识别系统、有限存贮的反馈识别系统、上下文分析在模式识别中的应用以及模式识别和通信理论的关系等内容，并在书中作了一些计算机识别的举例。

本书除每章后有小结、习题及参考文献和书籍外，在书的最后尚有习题解答，供学习者参考使用。

本书除可作为大专院校研究生和本科生有关专业的教材之外，尚可供计算机信息处理、生物医学工程、通信、自动控制等科技领域中从事有关工作的科技人员参考使用。

统计模式识别

作　　者　陈 季 镐　（美）
译　　者　邱 焕 章　邱　华
校　　译　姜 崇 熙
责任编辑　时 友 芬

*

北京邮电学院出版社出版
新华书店北京发行所发行　　各地新华书店经售
北京密云华都印刷厂印装

*

850×1168毫米　1/32　印张 10.25　字数 261.5千字
1989年5月第一版　1989年3月第一次印刷
印数：1—2000册
ISBN 7-5635-0008-1/TN·2　定价：2.45元

译　序

我们翻译的这本书是著者在美国大学中为研究生讲授统计模式识别时所编著的教材。有关这方面的书籍，目前国内尚不够丰富。特别是带有习题解答的更不多见。

本书在为新跨入统计模式识别领域的人，阐述了基本概念和背景后，着重系统地介绍了若干识别方法，并讨论了它们与计算机的关系，体现了本领域内一些关键方法和成果，以及它们的发展。本书对序贯识别方法讲的尤为详细，特别是对模式识别与通信关系的阐述，更具有特色。

我们相信，本书可作为高等院校的专业教材，也可供有关专业的科研、工程技术人员和自学人员参考使用。

本书在翻译的过程中，有幸得到了目前正在美国东南麻省大学任教的原书著者陈季镐教授的支持和帮助，他为译者寄来了原书的习题解答和勘误表，为提高译本的质量和充实本书的内容提供了条件，对此表示衷心地感谢。

本书在翻译过程中，还得到了从事高教工作四十余年的天津大学姜荣熙付教授的热情帮助和悉心校译，对此表示由衷地感谢。

由于译者水平有限，译本中难免有错误和不妥的地方，望读者批评指正。

译　者

一九八八、七、廿、

原　序

为在模式识别中使用统计方法，过去十五年间已作了大量的工作，作者相信，有关这方面的成果，尤其最近六年来取得的成就，将不仅可以更好地设计识别机，而且在统计数据处理、通信和控制系统以及一些与计算机有关的方面，都有广泛的应用前景。本书力求将这些结果组织在一个紧凑的、逻辑的结构之中，以便对学生或研究人员等有所帮助。所作的各种尝试，都是为了以合理的深度和广度、有条理地向读者提供各种主要的统计模式识别方法。

为了不致使本书太厚，著者不想汇集或一一陈述所有的成果，而只选择了一些课题，用以向读者介绍主要的发展和一些主要的成果。读者在理解了某个课题以后，如有兴趣，可以再去阅读每章后面推荐的参考文献和书目，以便深入学习。本书没有涉及随机自动机的使用和识别用的语言方法等课题。

选择识别特征是模式识别中最重要的课题之一。它与数理统计的“试验设计”这一基本问题有关。虽然有关特征选择的成就还不够多，但其它识别方面的努力，对这方面的欠缺当有所补充，例如非线性判决界的应用、用适当方法表达模式、上下文分析、以及反馈等等方面的各种应用，都创造了改善性能的条件，而这些改善性能的条件，可望来自更为完好的特征集。

第一章介绍的是统计模式识别的基本概念。大多数识别系统的主要目的在于取得最大正确识别率。为此，必须首先建立最佳判决界，这是第二章介绍的内容。第三章叙述的是一些描述模式的最佳方式。这个问题并不比第四章所述的特征选择次要，但往往被忽视。几乎对所有的模式，信息统计值都是间接选择特征集的良好准则。第五章和第六章所述的具有递归和非递归算法的监

督或非监督学习（或估计），都是统计模式识别中具有丰硕成果的领域。这些算法在控制系统参数估计中也特别有用。第六章中的随机近似和有关的技术并不需用全部的参数统计量。但在求取统计模式识别问题的实际解答时，需要使用第七章描述的非参数（无分布的）统计方法，因为参数统计量的假设并没有被证明。

第八章叙述的是以分析数据结构的方法确定最佳聚类和模式簇。这是模式识别的另一重要发展。其目的在于确定模式的类别并减少数据量，且使误差减到最小。虽然在大多数识别技术中处理的特征数量是固定的，但并非真正需要使用所有的特征。如果使用第九章所述的序贯判决理论，则从平均数来看，有可能选用较少数量的特征，即可达到既定的性能。序贯方法的实际优点是处理经过适当排序后的特征。这些特征可以由预定的方法或信息反馈方法加以排序。第十章中具有存贮限制的识别系统，是一个应受到更多重视的基本问题。几乎在所有的识别应用中，相继模式间都存在相互依存关系。第十一章概述的正是如何利用上下文分析以设计最佳识别系统的问题。最后，第十二章叙述了学习算法在自适应信号检测和通信接收机中的应用问题。

本书除第八、十和十一章以外，都是我1968年秋在波士顿东北大学为研究生讲课的笔记原稿。虽然是研究生水平的教材，但也可供大学生研讨班、短期讲座或自学之用。每章后面都附有参考文献和书目。书中给出的习题，一部分是补充教材中所讨论的问题，一部分供计算机识别试验之用。本书多以直观的解释代替繁长的数学推导。书中介绍了一些计算机识别的结果，这些结果与采用的数据集密切有关。读者如有兴趣，可以用自己生成的数据对给出的某些技术进行实验。有幸的是在过去的几年里，有些研究人员如数据趋势公司(Data Trends,Inc.) 的W.H.海莱曼 (Highleyman)博士,哈尼韦尔公司(Honeywell,Inc.) 的A.L.诺尔(Knoll)博士,斯坦福研究院(Stanford Reasearch Institute) 的J.蒙桑 (Muson) 博士等为字符识别的研究社

团提供有他们的字符识别数据集。利用公用数据集，可以对不同识别技术的优劣进行比较。

统计模式识别与数理统计中几乎每个基本问题都密切相关。因此无可置疑，统计模式识别的进展将大大得益于统计判决理论的最新发展，反之亦然。在这一领域中，理论上的发展比实际应用要快得多。但不论是在统计模式识别的理论方面还是在实用方面，在未来的岁月中，都将是大有发展前景的。

陈季镐

目　录

第一章 绪 论

1 模式的描述

什么是模式，从字面理解，可以是一个模型、标志或作某些事情的方案。模式可能是具体的，也可能是抽象的。几乎凡人们五官所能及的任何东西——一个字符、一张照片、生物波形，语音模式、气味、味道等等，都可以认为是一种模式。

一种模式类别则是一组具有确定特性的模式。例如，识别心电图时，就可以分为心脏正常与心脏不正常两类模式。在语音识别中，可以根据各种对话的言词或不同的讲话者进行模式分类。

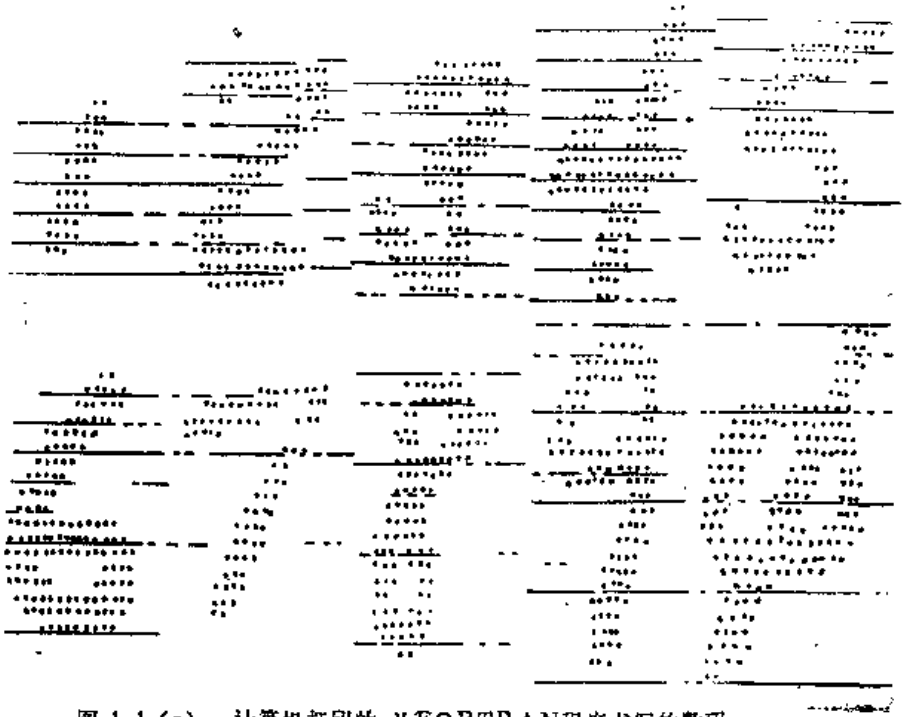

图 1.1 (a) 计算机打印的，为FORTRAN程序书写的数码。
（样本由斯坦福研究院J.蒙桑博士提供）

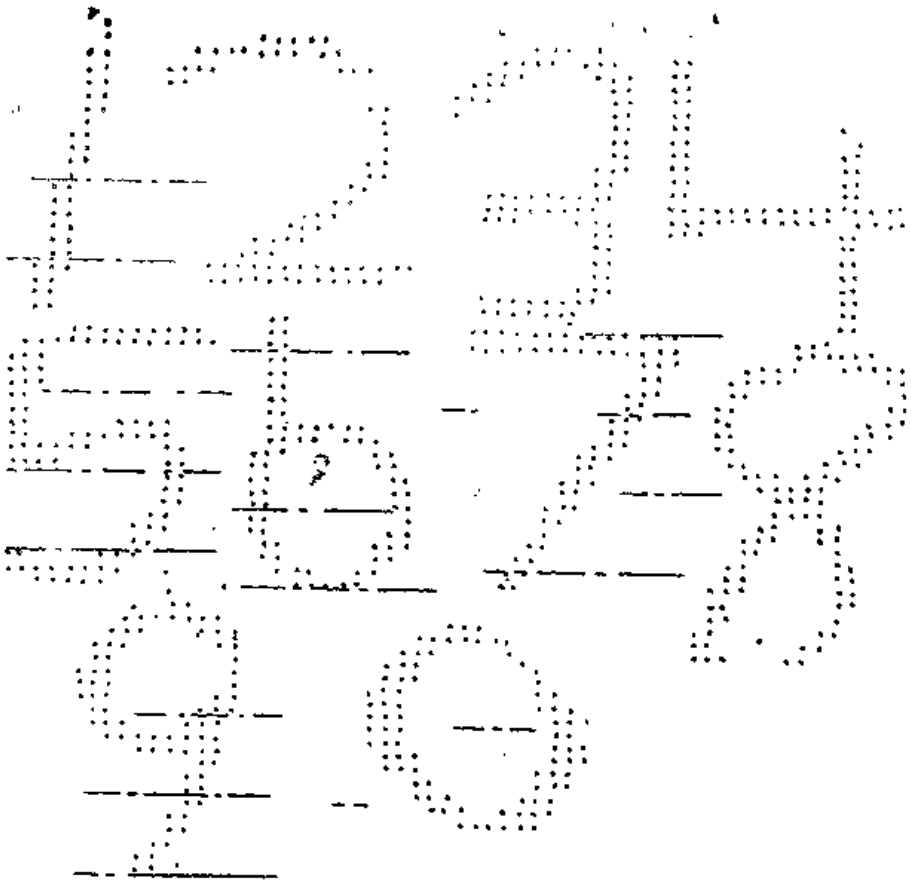

图 1.1 (b)　手写印刷体数码的典型样本
（由哈尼韦尔的A.L.诺尔[注]博士提供）

所谓模式识别，则是在某些一定的度量或观察基础上，将某个模式划分到某种模式类别中去。这些一定的度量或观察是主观决定的。所以不同型式的分类，使用的度量，也不尽相同。

图1.1示出了一组典型模式，它们是一些字符样本模式和一些时间系列模式。由于观点不同，模式类别成员的变化在性质上可以是确定性的(非随机的)，也可以是随机的。如根据已知铅字特性识别字符"*A*"时，一个"*A*"字的数字模式将会由于打字带印油的变化或在数字化过程中掺进了噪声等原因而有所改变。这些导致模式变化的原因，其性质可能是确定性的，也可能是统计性

[注]　原书作*A*·诺尔——译者

的。如果使用统计模式的识别方法，可以认为这些来自存贮参考的模式（理想的或平均的模式）变化是随机的。于是这种变化就必须用概率量来描述。最近20年在模式识别发展过程中，不论哪个学派都取得了丰硕的成果。本书主要论述在模式识别领域中，统计的或与统计相关的方法。

2 模式识别的概率表达方法

一旦取得一个模式的度量，则模式分类器，也即一次识别，就会把它分类到某一模式类别中去。识别机的框图见图1.2。接受器对模式作预处理，以便选择并抽取描述模式的特征。归类器判定这一模式属于哪一类，即参与类别划分问题。迄今为止，利用一个较好的分类方案来改善正确识别的百分率，曾做过较多的工作。但相对地说，对特征选择问题作的工作则较少。

假定模式是随机的，它的特性就必须用某些统计值来区分开来。现以x代表所作的每次度量，也可以认为x是N维度量（样本）空间的一个点。用ω_i代表具有一定性质状态的每个模式类别，其先验概率$P_i = P(\omega_i)$，$i = 1, 2, \cdots, m$，m是类别总数。如果根据度量x识别机把实际上是第i类的模式，误判为第j类，就会造成一个损失。令$L(\omega_i, d_j)$为与以上判决相关的损失，并令$P = \{P(\omega_i)\}$为先验概率的集合，以及$p(x/\omega_i)$为第i类的概率密度函数，则平均风险函数将是

$$R(P, d_j) = \int_{\Omega_x} \sum_{i=1}^{m} L(\omega_i, d_j) P_i p(x/\omega_i) \mathrm{d}x \qquad (1.1)$$

其中积分要在Ω_x表示的x空间上进行。识别机将选择使式(1.1)代表的平均风险最小的模式类别。划分类别问题，与统计判决理论完全相同。

因此数理统计的许多有效的成果，都可以在模式分类中应

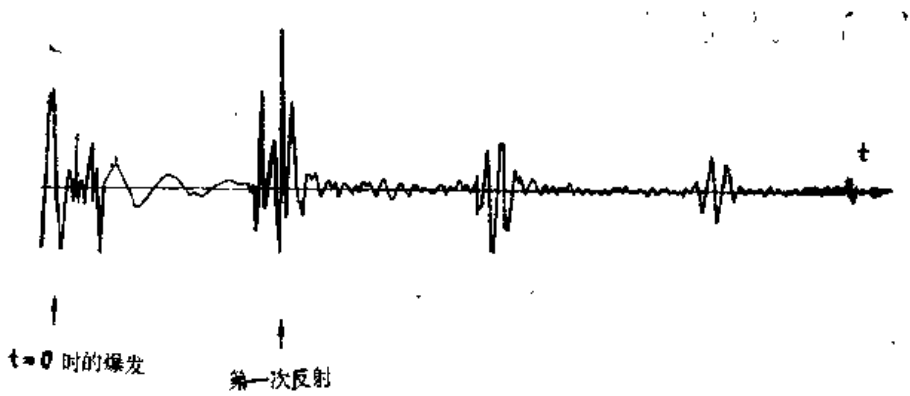

图 1.1 (c) 典型的海洋地震数据集(由伍兹霍尔(Woods Hole)海洋研究所的K.普拉达(Prada)博士提供) 也参见C.H.陈, On the APPlication of Pattern Recognition Techniques to Oceanographic Signal Processing, 1970 IEEE海洋环境工程国际会议, 巴拿马城, 佛罗里达州。

心电图实用特征度量

(1)心率
(2)P-R间隔
(3)QRS间隔
(4)Q-T间隔
(5)P-R间隔/心率比

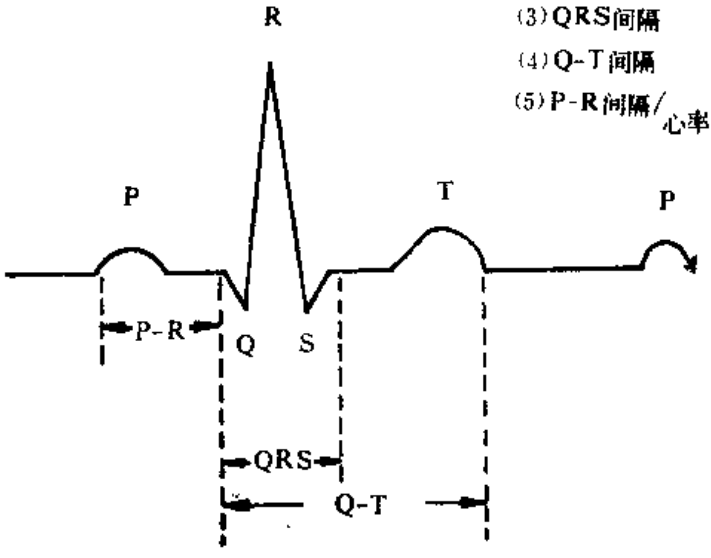

图 1.1 (d) 一段典型的心电图

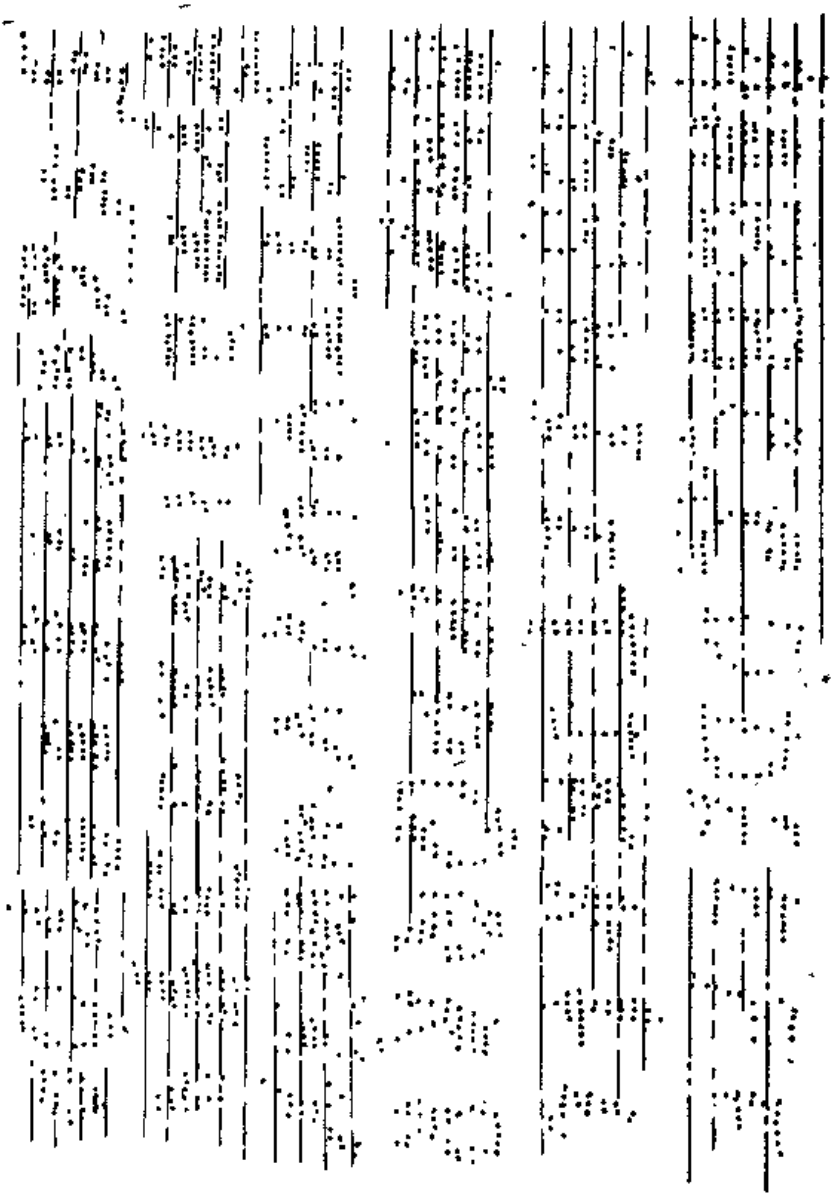

图 1.1 (*a*)　海莱曼数据集的典型样本

用。至于接受器则没有如此明显的等效性。但可以使信息度量（一个概率量）极大化，从而取得一个在所有类别中鉴别能力最大的

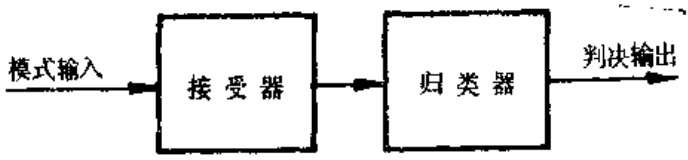

图 1.2 识别机框图

特征集。但在有些情况下†，统计方法既不理想也不可行。但无论如何每个问题都可以用概率来表达，而且至少在理论上可以得到统计解。本书除统计判决理论以外，还将讨论一些其它统计准则。

3 几何释义

用几何释义解释模式识别中的一些基本概念，是常用的既清晰而又方便的办法。现假定有一个具有类别A和B的二维度量集合，如图1.3所示。在图1.3(a)中，可以构造一条非线性判决界，把样本空间无误差地分成类别A和类别B两部分。两个类别互不相交的样本空间，称作可分的样本空间。如果判决界是线性的，如图1.3(a)中虚线所示，则两个类别是线性可分的。但是两个类别往往不能用单一判决面加以分离，如图1.3(b)所示。如果应用统计判决理论，则能以最小的误差分离样本空间。还应当注意到，用模式分类的概率表示法来表示典型样本空间是较为方便

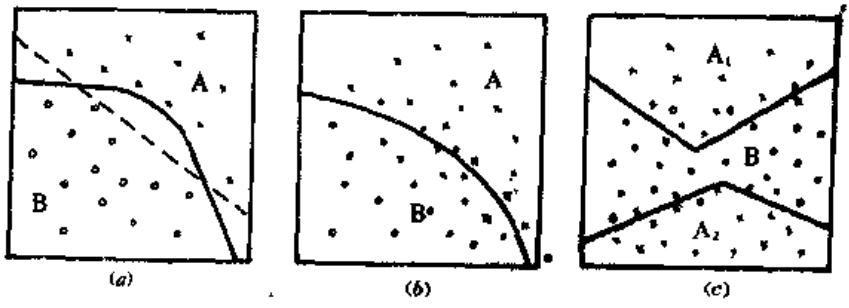

图 1.3 典型的两类别二维样本空间

† 例如在光学文字识别中，模式的维数太大，以致无法使用减少维数的统计方法。