模式识别

理论、方法和应用

王碧泉　陈祖荫　著

地震出版社

地震科学联合基金会资助项目

4 模式识别

理论、方法和应用

王碧泉　陈祖荫　著

地震出版社

内 容 提 要

本书系统总结了统计模式识别的理论、方法与应用。全书共九章，全面介绍了特征选择、聚类和判别等方面的常用模型和算法，模式识别在地震学、数字图象处理和决策管理等领域中的应用，以及作者近年来的其他有关研究成果。本书的著述兼顾了提高与普及两个方面，并特别注重实用的需要，各类方法都给出了详细算法和应用实例，既适合于从事模式识别与图象处理研究的专业人员阅读，又可供从事应用数字、自动化技术、计算机科学、遥感技术、地震、地质、生物医学、管理科学等的科技人员以及大专院校有关师生参考。

模 式 识 别

理论、方法和应用

王碧泉　陈祖荫　著

责任编辑：陈非比

*

地 震 出 版 社 出版

北京复兴路 63 号

国防大学第一印刷厂印刷

新华书店北京发行所发行

全国各地新华书店经售

*

850×1168　1/32　12 印张　325 千字

1989 年 11 月第一版　　1989 年 11 月第一次印刷

印数　0001—2400

ISBN　7-5028-0232-0/O·7

(619)　定价：7.50 元

序

一门方兴未艾、前景灿烂的边缘学科必定包含有理论、方法与应用三方面内容。它的理论应当是深刻的、比较系统的并且不断发展的；它的方法应当是有效的、多样化的并且不断创新的；它的应用应当是领域广泛的、富有成果的并且不断开拓的。近30年来建立起来的模式识别算得上这样的学科。把这一学科介绍给广大科学工作者与工程技术人员，是一项意义重大的工作。

王碧泉与陈祖荫同志研究模式识别多年，得到了很多出色的成果。他们合写的这本专著《模式识别——理论、方法和应用》，与国外已经出版的许多模式识别著作不同，侧重于阐述这一学科中比较常用的方法，侧重于介绍这一学科在若干领域中的应用。我曾有机会看到作者的手稿，深有“不虚此读”、“先睹为快”之感。

在模式识别的方法方面，作者经过细致筛选，论述了特征选择、聚类、分类判别、检验以及模糊聚类等方法。从实用与有效的角度出发，明确说明了方法所依据的基本原理但又不拘泥于数学论证细节；清楚地阐述了不同方法的使用范围与特征，并配有很好的实例，使读者易于根据自己面临的问题选择与使用——这些都会给人留下深刻的印象。

在模式识别的应用方面，书中分别讨论了这一学科在地震学、数字图象处理与决策管理中的应用。两位作者对地震模式识别有深入的研究，对地震区划及潜在震源的判定，强震发生时间

的预测以及地震前兆的综合分析等方面，都做了很好的工作。在后两个领域，本书作者及其合作者也做了不少工作。这就使得本书对模式识别应用的论述具有与众不同的特色。

毫无疑问，对模式识别有兴趣的读者能从本书获得很大的启发。

沈 永 欢

1987 年于北京

前　言

模式识别是一门新兴的边缘学科，它的应用几乎遍及各个科学领域。有人曾经断言，在如自动控制这样的领域中，唯一有希望解决最高一级问题的途径，是应用模式识别和人工智能的综合成果。这样一种思想也有望用于复杂的社会经济系统、生物医学、环保系统以及运输系统等。

无论国内外现在都已有一批关于模式识别的专著问世，但是这一领域中深奥的数学理论往往使得实际工作者感到困难。例如，为了阐明贝叶斯分类方法，通常需要讲述大量的统计理论，从假设检验一直讲到密度函数的估计。以致使得布雷默曼(Bremermann，H.J.)发出这样的感叹："贝叶斯山太拥挤了"。

与大量的专著不同，本书是一本介绍实用的模式识别方法的著作。在本书中，我们企图叙述一批常用和有效的模式识别方法和算法，而对复杂的数学理论则只介绍其结果，尽量不做深奥的证明。我们希望，各个实际领域中的科技工作者能够通过本书比较容易地掌握一些模式识别方法，并将其付诸实施。

要在一本书中罗列所有的模式识别方法显然是不可能的。我们选择方法的标准是：(1) 常用而有效的方法；(2) 近年来新提出的一些比较"漂亮"的方法；(3) 我们自己的某些工作。对于一些已为人们熟知或者理论价值胜于实用价值的方法，则予以忽略。

对于模式识别的应用，我们着重介绍了三个方面，即地震学、图象处理和管理科学。我们希望读者通过以上三个方面了解模式识别的实际应用。关于模式识别在地震学中的应用，是由苏联学者开始的，继之苏美又联合进行了许多研究；在我国近几年来也有很大进展。关于这一方面的内容在其他专著中还未见过报

道。

本书是在国家地震局地球物理研究所和北京工业大学应用数学系多年合作的基础上编写的，实际上反映了这一合作集体这几年的研究成果。吕宏伯同志撰写了第二章第四节的初稿，聂金宗同志撰写了第三章第三节和第六章第四、五节的初稿，上述两同志还对第一至第三章，第六章和第八章提出了宝贵的意见。王玉秀同志对第三、四、五、七章，马秀芳同志对第四、七章，陈锦标同志对第二章都曾提出了宝贵的意见。王春珍同志协助作者进行了大量的图件绘制及清稿工作。以上诸位同志对本书的出版都做了重要的贡献，在此一并表示感谢。最后作者特别感谢沈永欢同志对本书的关心并给本书写了序言。

作者恳切地希望得到读者的批评和指正。

作　者

1987 年于北京

目　录

绪　论

模式识别(pattern recognition，亦可译作模式辨认、图象识别、图形识别、型式识别)是近30年来得到迅速发展的一门新兴边缘学科。关于什么是模式或者机器所能辨认的模式，迄今还没有一个确切而严格的定义。卡纳尔(Kanal，L.)曾经说过这样一段话：

> “关于什么是模式识别和机器所能辨认的模式，至今还没有人能象香农(Shannon)对‘信息’一词做出定义那样，给出一个确切的定义。如果一旦出现了这样一个定义并被证实能够推动理论的发展，那将标志着人类智力的一大进展。虽然如此，目前的局面并不影响模式识别在各领域中的广泛应用。”

在数学中，没有定义的概念可以大体上分为两类：一是初始或本原(initial)概念，相当于推理系统中的公理，例如点、集合、原子命题等；二是暂时还找不到确切定义的概念。模式一词大概便属于后一类。因此，目前我们倾向于采用沈永欢（1979）的提法：“模式是对种种物质的或精神的对象进行的分类、描述和理解。”在一些应用领域中，有些专家则干脆将模式识别称为数量（数值）分类学，例如阳含熙、卢泽愚和史尼斯(Sneath，P.)、索科尔(Sokal，R.)等。

严格地说，模式识别不是简单的分类学。它的目标包括对于系统的描述、理解与综合。模式识别的高级阶段是通过大量信息对复杂过程进行学习、判断和寻找规律。从这个意义上说，模式识别与“学习”或“概念形成”的意义是相近的。模式识别与机器智能的结合将为人类认识世界和做出新的发现开辟广阔的前景。

在模式识别的初创时期，人们企图通过仿生学，即对动物感受器官和机能的模拟来实现自动模式识别。但是，从1968年以来，由于实际上的困难，关于生物模型的热情有所下降。多数模式识别学者接受了这样的观点：模式识别是一个力图达到的目标，而实现这一过程和方法的技术可以各异。因此识别过程的机制不一定与生物系统相同。由此出发，人们开始大量引入数学理论和方法作为识别的工具。近几年来，这些理论和方法已经日臻成熟，以至有人认为模式识别已从一门"艺术"发展为系统的科学。当然，象心理学、生物学或者光学方法仍是不可忽视的一些方面。

模式识别的数学方法中最重要的是统计决策方法。它的一个基础是多元统计学，另一个支柱则由多种数学分支构成。此外在统计决策模式识别中也大量地使用着代数学、运筹学、图论以及模糊数学方法。本书所介绍的方法都属于这一范畴。

另一类模式识别方法称为语言或结构方法。这一方法是富有启发性的，并且比统计方法更为重视模式的结构性质。

马瑟荣(Matheron，G.)与塞拉(Serra，J.)创立的数学形态学(mathematical morphology)方法是另一种新颖的模式识别模型。在这一模型中，理论与实践得到了自觉的和谐的结合。这一方法主要用于对图形的识别。

现在我们用一个例子说明模式识别的任务与意义。假定我们需要研究某一特定地区的强震孕震规律并据此预测强震的发生时间。此时被研究的对象是一批"时间段"，我们可以对每段时间选取若干个孕震因素（地震前兆）并测量它们的数据。按照传统的数学观点，下一步工作似乎应当是构造一个精确的方程或者函数，用各个孕震因素作为变量，然后通过求解来描述孕震过程并进行预报。但是，由于我们目前对孕育地震过程的认识还不十分清楚，所以构造上述的方程或函数是有困难的。

模式识别采取另一种处理方式。它暂不去追求精确的数学方

程，而是在专家经验和已有认识的基础上，从所获得的大量数据和历史事实出发，利用数学方法来完成识别过程。本例中在进行模式识别前，首先将所研究的问题化为分类的问题。以 ΔT 为间隔将所研究的全部时间划分为 N 个时间段，每一时间段称为一个样品（或对象），记为 X_1，…，X_i，…，X_N，$i=1$，2，…，N。在用分类判别方法识别时，还需事先确定样品的类别。这里取曾发生过震级 $M\geqslant 7$ 地震的时间段为危险类样品（记为 D），否则为不危险类样品（记为 N）。如此，将预测强震发生时间的问题转化成了对时间段进行分类并识别 D 类时间段的模式识别问题。

强震发生前，可观测到一些前兆现象，例如地震频度增高，b 值下降等等。本例中将这些可能反映孕育强震的每一因素称为一项特征。共选了 n 个特征，并测量出这些特征的数据，这一步骤称为特征提取。另外我们还利用一些方法对所提取的特征进行选择，淘汰掉一些特征，保留一些起主要作用的特征用于识别（如科拉-3 方法），或者采用线性或非线性方法得到 n' 个新特征（$n'<n$）以降低特征空间的维数（例如主成分分析法），这一步骤称为特征选择。然后采用各种分类判别方法或聚类方法，根据 n 个（或 n' 个）特征对样品进行研究，找出 n 维特征空间中合适的超曲面，将两类样品分开，此超曲面称分类器或判决函数，这一步骤称为分类识别。整个模式识别过程如图 1 所示。

首先对已知类别样品（学习样品）进行上述过程，以找出分类器，这一过程称为学习或分析过程。对未知类别样品的识别即预测，称为识别过程。在后一过程中，数据输入至特征选择等步骤均同前一过程，而分类识别时只要用前一过程中找出的分类器对该样品分类即可。当我们需要知道未来某一时间段内是否会发生强震时，就将这一时间段当作未知类别的样品，视其是否分到 D 类来预测这一时间段内是否会发生强震。

这个例子是具有普遍意义的。在大多数复杂系统中，由于参

数的众多，机理的复杂以及流程的繁复，建立精确的数学模型都是相当困难的。早在60年代，就有人提出用模式识别方法“学习”最优控制方式的设想，这一设想正在不断趋于完善和在各个领域中得到实施。

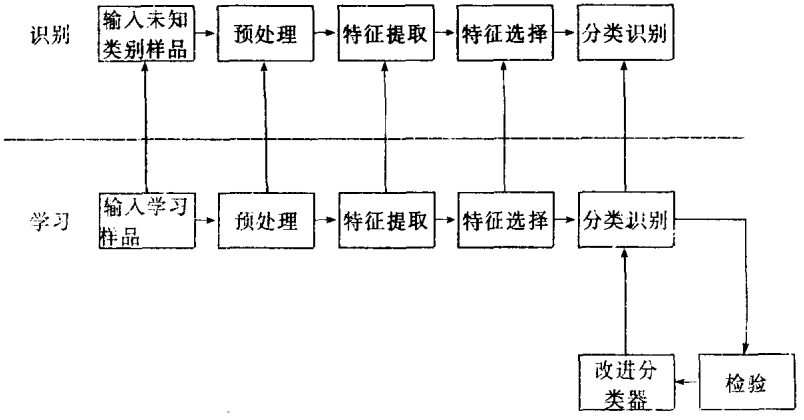

图1　模式识别流程图

另一方面，即使机理并不十分复杂的问题，也常常需要使用模式识别方法加以解决。例如，我们希望用机器去识别英文字母。要实现这个任务似乎采用简单的“样板匹配”法就足以解决了。换句话说，只要对每个字母构造一个标准“样板”，并将输入机器的每个字母与它们一一匹配即比较，然后按照相似程度做出决策。

但是，这种方案实际上是不好的。这是因为，每个人书写同一字母时都会出现或多或少的差异，而每个字母的位置、大小和角度也会有所区别。所有这些都属于随机性的干扰因素。因此，明智的做法仍然是对每个字母提取一批特征，并且根据这些特征的统计规律去构造分类器，即采用典型的模式识别方法。

由此可见，模式识别是一种借助于大量信息和经验进行推理的方法。它属于机器智能的范畴。

模式识别的发展速度是相当迅速的。1965 年以前，关于这一领域还只有一本专著，即尼尔森(Nilsson，N.T.)的一本著作。但是近十年来的专著已经不下数十本，还有许多文献汇编和会议报告汇集，仅英文文献，目前每年已超过 500 篇。

1973 年，由国际电气与电子工程师协会(IEEE)发起召开了第一次关于模式识别的国际学术会议。此后成立了国际模式识别协会(IAPR)，每两年召开一次国际模式识别学术会议(ICPR)。此外，IEEE 的计算机学会于 1977 年成立了模式分析与机器智能(PAMI)委员会，在不召开 ICPR 的年份则举办 IEEE 的模式识别与图象处理学术会议。最近一次 ICPR（1984，加拿大，蒙特利尔）上，我国学者（不包括台湾省的学者）有 10 篇论文入选，其中包括本书作者的两篇。

在国内，自动化学会和计算机学会都设有模式识别与机器智能专业委员会，迄今已举行过 5 次学术会议。有关模式识别的专著和译著已出版了近 10 本之多。

模式识别的应用范围是极为广泛的，几乎不用费力便可以举出以下一些方面:

(1) 文字和字符识别：信函分拣，文件处理，卡片输入，稿件输入，支票查对，期刊阅读，自动排版等。

(2) 图形识别：遥感和航空照片分析，金相图分析和鉴定，目标搜索，指纹、唇纹和面貌辨认，X 光、显微图象、热象及超声图象检查等。

(3) 声音识别：语音识别和鉴定，侦听和机器故障判断等。

(4) 生物医学应用：疾病诊断，癌细胞、白血球、染色体和疟原虫检查，修复手术控制设计等。

(5) 工业应用：产品质量检验，集成电路设计，自动键合，图形设计等。

(6) 预报问题：天气预报，工业烟雾预报，地震预报，经济预报等。

(7) 社会科学与管理科学：社会人类学，体质人类学，心理学，考古学，语言学，决策问题，系统可靠性分析与对系统未来行动的判断，市场研究，文艺风格鉴定等。

(8) 其他：生态学，生物地理学，地球科学，孢粉学等。

以上举出的还只是模式识别应用范围中的一部分。例如，军事方面的应用我们没有提到。"有人甚至认为，模式识别是本世纪雄心最大的学科，需要电子学家、数学家、生物学家、心理学家、哲学家和社会学家等的通力合作"（沈永欢，1979）。例如，模式识别在地震学中的应用，就是由苏联科学院应用数学家、地球物理学家、生物医学家以及美国加州大学、麻省理工学院的教授们合作并首先开始的。他们所研制的"PR 程序包"由联合国科教文组织支持已向世界各国进行了推广。在我国，国家地震局地球物理研究所、北京工业大学应用数学系和中国科学院生物物理研究所在这一领域内已进行了多年的合作。

> "人类非常善于把感觉信号变为认识符号，也善于用常识解决问题。但是，我们在大量数据面前就感到畏怯了：我们处理数据往往没有系统，健忘，会觉得厌烦，而且容易思想旁骛。……具有对话能力的计算机将帮助我们解决一些问题。我们应该为认识到自己的能力有限，并发明了弥补这些不足的技术而感到庆幸。"
>
> —— 费吉鲍姆(Feigenbaum，E.A.)等

本书第一章介绍模式识别的基本概念和主要内容，第二至五章分别介绍一些最常用的方法和算法，第六章介绍模糊(fuzzy)方法在模式识别中的应用；最后，第七至九章分别介绍模式识别方法在三个实际领域中的应用。通过这些介绍，我们希望能使读者对于模式识别的主要内容、典型方法和使用方式形成一个初步的印象。

第一章　基本概念和主要问题

在这一章里，我们首先介绍模式识别中的各种基本概念，然后分别叙述模式识别研究领域中的几个主要方面，即聚类、分类判别和特征选择。

第一节　样品与特征

一、样品与特征

在模式识别中，被观测的每个对象称为一个样品。

例如，在地震危险区域划分中，每个地面区域是一个样品；在强震发震时间预报中，每个时间段是一个样品；在疾病诊断中，每个病人是一个样品；在质量检验中，每个产品是一个样品，等等。

我们用大写英文字母 X，Y 或 Z 表示样品。如果一批样品共有 N 个，我们把它们分别记作 X_1，X_2，…，X_N。如果一批样品分别来自 m 个不同的类，来自第 1 类的样品有 N_1 个，来自第 2 类的有 N_2 个，等等，则可把类号加上圆括号后附在样品名的右上角，即写作

$$X_1^{(1)}, X_2^{(1)}, \cdots, X_{N_1}^{(1)}, X_1^{(2)}, X_2^{(2)}, \cdots, X_{N_2}^{(2)}, \cdots, X_{N_m}^{(m)}.$$

其中记号 $X_i^{(j)}$ 表示第 j 类的第 i 个样品。

对每个样品必须确定一些与识别有关的因素，作为研究的根据；每个因素称为一个特征。

例如，在研究地震危险区域划分时，每个样品是一块区域，与之对应的特征可以取该区域内的各项地质地貌特征，如主活动断裂数，主活动断裂的端点及交汇点个数，区域内的最大高程

等。又如，在医学诊断中，每个样品是一个患者，特征便可取与诊断有关的各项病理指标，如体温、血压、白血球数目等。

我们对特征用小写的英文字母 x，y 或 z 表示。如果对每个样品 X 共取 n 项特征，便可把 X 看成 n 维空间中的一个点或者一个列向量，记作

$$X=\begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix}=(x_1,x_2,\cdots,x_n)^T.$$

类似地可以采用以下的记号:

$$X_i=(x_{1i},x_{2i},\cdots,x_{ni})^T,$$

$$X_i^{(j)}=(x_{1i}^{(j)},x_{2i}^{(j)},\cdots,x_{ni}^{(j)})^T.$$

把样品看成点有利于从几何上考虑问题，看成向量则在计算时比较方便。由特征 $x_1,x_2,\cdots,x_n$ 构成的空间称为 n 维特征空间。

若有一批样品共 N 个，写出每个样品所对应的 n 个特征之值，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	X_1	X_2	…	X_j	…	X_N
x_1	x_{11}	x_{12}	…	x_{1j}	…	x_{1N}
x_2	x_{21}	x_{22}	…	x_{2j}	…	x_{2N}
…	…	…	…	…	…	…
x_i	x_{i1}	x_{i2}	…	x_{ij}	…	x_{iN}
…	…	…	…	…	…	…
x_n	x_{n1}	x_{n2}	…	x_{nj}	…	x_{nN}