

模式识别

杨光正 吴 岷 张晓莉 编著

中国科学技术大学出版社

模 式 识 别

杨光正 吴 岷 张晓莉 编著

中国科学技术大学出版社

2001·合肥

内 容 简 介

本书是在我们十余年来对大学本科生和研究生讲授《模式识别》课程的基础上,结合近年一些国外教材和我们的模式识别研究成果编写而成,可作为大学本科和研究生学习模式识别课程的参考书。

虽然近年有许多新的模式识别研究方法提出,但模式识别的基础仍是由决策论方法(统计方法)和结构方法(句法方法)两大部分构成。因此,本书主要介绍统计识别方法和句法方法的基本理论和方法。本书共有14章,第1章引论,介绍模式识别的研究内容和研究方法;第2章至第6章介绍统计模式识别方法;第7章至第11章介绍句法模式识别方法;第12章介绍把句法模式识别推广到句法知识系统,从而实现基于知识的模式识别;第13章介绍模糊集论在模式识别中的应用,指出模式识别的模糊本质;第14章介绍人工神经网络在模式识别中的应用,提出用模式识别的观点研究神经网络,使前馈神经网络透明化。

图书在版编目(CIP)数据

模式识别/杨光正,吴岷,张晓莉编著. —合肥:中国科学技术大学出版社,2001.7
ISBN7-312-01254-X

I. 模... II. ①杨...②吴...③张... III. 模式识别 IV. O235

中国版本图书馆 CIP 数据核字(2001)第 80521 号

中国科学技术大学出版社出版发行
(安徽省合肥市金寨路 96 号,230026)
中国科学技术大学印刷厂印刷
全国新华书店经销

开本:787mm×1092mm/16 印张:15 字数:400 千
2001 年 7 月第 1 版 2001 年 7 月第 1 次印刷
印数:1—3000 册
ISBN7-312-01254-X/O·244 定价:20.00 元

前 言

模式识别是一门很受人们重视的技术科学。虽然早在 30 年代就有人试图以当时的技术来解决一些识别问题,但只是随着近代计算机科学技术的进步,模式识别才真正发展起来。从 60 年代至今,在模式识别领域中已取得了不少成果。它的迅速发展和广泛应用的前景引起各方面的关注。

模式识别属于人工智能范畴,其目的就是用机器去完成人类智能中通过视觉、听觉、触觉等感官去识别外界环境的那些工作。这是一个新的研究领域,到现在为止,它的理论和技术都远未完善,很多课题有待人们去研究。

目前模式识别主要用两类方法进行,即决策论方法(统计方法)和结构(句法)方法。在模式识别中也应用了一些模糊数学的方法,取得了良好效果;对于某些复杂的模式识别问题,可以与人工智能领域中的专家系统结合,研究出“基于知识”的模式识别方法;神经网络作为一个有“学习能力”的工具,近来在模式识别中也得到许多应用。这些方法对模式识别学科起到很大的推进作用。要指出的是,这些新的模式识别方法,仍可归结于以上两种基本的模式识别方法或它们的结合。

本书是在我们十余年来对大学本科生和研究生讲授《模式识别》课程的基础上,结合近年一些国外教材和我们在模式识别方面的研究成果编写而成的,可作为大学本科和研究生学习模式识别课程的参考书。

本书主要介绍模式识别的基本方法,即决策论方法(统计方法)和结构方法(句法方法)的基本理论,同时也介绍模式识别方法的一些发展方向,其中包括作者近年来的部分研究成果。并介绍主要方法的实现途径。全书共 14 章,第 1 章引论,介绍模式识别的研究内容和研究方法;第 2 章到第 6 章介绍统计模式识别方法,包括统计识别的基本理论和方法以及模式特征的抽取和选择;第 7 章到第 11 章介绍句法结构模式识别方法,在第 7 章中简要叙述了与模式识别应用有关的形式语言的基本理论;第 12 章介绍把句法模式识别推广到句法知识系统,从而实现基于知识的模式识别;第 13 章介绍模糊集论在模式识别中的应用,指出模式识别的模糊本质;第 14 章介绍人工神经网络在模式识别中的应用,提出用模式识别的观点研究神经网络,使前馈神经网络透明化。

本书第 12 章 12.3 节和第 14 章 14.1、14.3 节由吴岷执笔;第 13 章 13.1、13.3 节和第 5 章 5.1.5 节由张晓莉执笔;其余由杨光正执笔。全书由杨光正统稿。

我们力求将本书编写好,但错误和疏漏仍在所难免,欢迎读者提出宝贵意见。

编著者

2000.12.12

目 次

第 1 章 引论	(1)
1.1 模式和模式识别	(1)
1.2 模式识别的发展和应用	(2)
1.3 模式识别的研究方法	(4)
第 2 章 线性判别函数	(7)
2.1 线性判别函数和决策面	(7)
2.2 感知准则函数和梯度下降法	(11)
2.3 固定增量算法及其收敛性	(13)
2.4 最小平方误差准则函数	(16)
2.5 多类情况下的线性判别函数和固定增量算法	(18)
2.6 分段线性判别函数	(21)
2.7 Fisher 线性判别函数	(23)
2.8 支持向量机	(26)
第 3 章 Bayes 决策理论	(30)
3.1 最小错误概率的 Bayes 决策	(30)
3.2 最小风险的 Bayes 决策	(34)
3.3 Neyman-Pearson 决策	(37)
3.4 最小最大决策	(38)
3.5 Bayes 分类器和判别函数	(41)
3.6 正态分布时的 Bayes 决策法则	(44)
3.7 离散情况的 Bayes 决策	(51)
第 4 章 概率总体的估计	(53)
4.1 参数估计	(53)
4.1.1 极大似然估计	(53)
4.1.2 Bayes 学习	(56)
4.2 非参数技术	(61)
4.2.1 概率密度函数估计的基本方法	(61)
4.2.2 Parzen 窗法	(62)
4.2.3 密度函数估计值的收敛性	(65)
4.2.4 应用 Parzen 窗法的一个例子	(66)

4.2.5 后验概率的估计	(67)
第 5 章 近邻法则和集群	(69)
5.1 近邻法则	(69)
5.1.1 近邻法则的一般概念	(69)
5.1.2 近邻法则的错误率	(70)
5.1.3 K -近邻法则	(73)
5.1.4 关于近邻法则的讨论	(73)
5.1.5 快速近邻算法	(75)
5.2 集群	(77)
5.2.1 样本间类似性的度量	(77)
5.2.2 集群的准则函数	(80)
5.2.3 迭代最优化方法	(83)
5.2.4 等级集群方法	(85)
5.2.5 基于近邻法则的集群算法	(90)
第 6 章 模式特征的抽取和选择	(92)
6.1 概述	(92)
6.2 离散的 Karhunen-Loève 变换	(94)
6.3 多个分布的特征选择	(98)
6.4 特征抽取方法举例	(100)
6.4.1 手写体字符识别的特征抽取	(100)
6.4.2 心电图特征的抽取	(102)
第 7 章 结构模式识别方法和形式语言	(105)
7.1 句法结构模式识别方法	(105)
7.2 形式语言和短语结构文法	(107)
7.3 正规语言和有限自动机	(112)
7.4 前后文无关语言和下推自动机	(119)
7.5 图灵机和线性界限自动机	(122)
第 8 章 形式语言在模式识别中的发展	(125)
8.1 程序文法	(125)
8.1.1 应用于模式识别时各型文法功能的比较	(125)
8.1.2 前后文无关程序文法	(128)
8.2 高维模式文法	(129)
8.2.1 阵列文法	(129)
8.2.2 矩阵文法	(130)
8.3 PDL 文法	(132)

8.4 树文法	(135)
第9章 句法分析	(141)
9.1 句法分析的一般概念	(141)
9.2 正规语言的句法分析方法	(146)
9.3 算子优先算法	(148)
9.4 Cocke-Younger-Kasami 算法(CYK 算法)	(152)
9.5 Earley 算法	(155)
9.6 Earley 算法分析	(159)
9.6.1 Earley 算法的实现	(160)
9.6.2 Earley 算法的改进	(161)
9.7 几种前后文无关语言句法分析算法的比较	(163)
第10章 句法结构模式识别方法的发展	(165)
10.1 随机文法	(165)
10.1.1 随机文法的概念	(165)
10.1.2 随机有限自动机	(168)
10.1.3 随机文法在模式识别中的应用	(169)
10.1.4 随机文法中产生式概率的估计	(171)
10.2 属性文法	(173)
10.2.1 词义信息的利用	(173)
10.2.2 前后文无关属性文法	(173)
10.2.3 利用属性文法的统计考虑	(178)
10.2.4 关于属性文法的讨论	(180)
第11章 文法推断和基元选择	(181)
11.1 文法推断的基本概念	(181)
11.2 正规文法的推断	(182)
11.2.1 规范确定文法	(182)
11.2.2 规范微商文法	(183)
11.2.3 K -尾文法	(185)
11.3 前后文无关文法的推断	(186)
11.3.1 利用前后文无关文法的自嵌套特性	(186)
11.3.2 利用具有结构信息的样本	(187)
11.4 模式基元的抽取	(188)
11.5 依据边界或骨架的模式基元抽取方法	(189)
11.6 依据区域的模式基元抽取方法	(191)
第12章 句法模式识别与专家系统的关系	(194)

12.1	句法模式识别的进一步考虑	(194)
12.2	用句法方法的基于知识的模式识别系统	(195)
12.3	句法专家系统的算法	(198)
12.3.1	ED 算法	(198)
12.3.2	EDS 算法	(200)
12.3.3	ES 算法	(201)
第 13 章	模糊集论在模式识别中的应用	(203)
13.1	模糊集论的基本概念	(203)
13.1.1	模糊性和模糊子集	(203)
13.1.2	模糊子集的基本运算	(203)
13.1.3	水平截集	(204)
13.1.4	模糊相关性的度量	(205)
13.1.5	模糊关系及其运算	(207)
13.2	模式识别的模糊本质	(210)
13.3	模式识别的模糊方法	(212)
13.3.1	隶属度原则和择近原则	(212)
13.3.2	基于模糊等价关系的模式分类	(213)
13.3.3	模糊文法	(214)
第 14 章	人工神经网络在模式识别中的应用	(216)
14.1	人工神经网络概要	(216)
14.1.1	最简单的感知器	(216)
14.1.2	误差反向传播训练算法(BP 算法)	(216)
14.2	用模式识别方法研究神经网络	(219)
14.2.1	前馈神经网络的分解	(219)
14.2.2	用分段线性划分研究神经网络的几个问题	(222)
14.3	神经网络的 PLD 算法	(224)
主要参考文献		(229)

第 1 章 引 论

1.1 模式和模式识别

模式识别是一门很受人们重视的学科。虽然早在 30 年代就有人试图以当时的技术来解决一些识别问题,但只是在近代随着计算机科学技术的发展和应用,模式识别才真正发展起来。从 60 年代至今,在模式识别领域中已取得了不少成果。它的迅速发展和广泛应用的前景引起各方面的关注。

模式识别属于人工智能范畴,人工智能就是要用机器去完成过去只有人类才能做的智能活动。在这里,“智能”指的是人类在认识和改造自然的过程中表现出来的智力活动的的能力。例如:

- 通过视觉、听觉、触觉等感官接受图像、文字、声音等各种自然信息去认识外界环境的能力;
- 将感性知识加工成理性知识的能力,即经过分析、推理、判断等思维过程而形成概念、建立方法和作出决策的能力;
- 经过教育、训练、学习不断提高认识与改造客观环境的能力;
- 对外界环境的变化和干扰作出适应性反应的能力等。

模式识别就是要用机器去完成人类智能中通过视觉、听觉、触觉等感官去识别外界环境的自然信息的那些工作。

虽然模式识别与人工智能关系很密切,但由于历史的原因,它已经形成了独立的学科,有其自身的理论和方法。在许多领域中,模式识别已有不少比较成功的实际应用,这也丰富了人工智能的内容。

“模式”这个概念的内涵是很丰富的。我们把凡是人类能用其感官直接或间接接受的外界信息都称为模式。比如,文字、图片、景物是模式;声音、语言是模式;心电图、脑电图、地震波等也都是模式,进一步,诸如社会经济现象、某个系统的状态等也都是模式。从这个意义上说,人们在做每个行动时都要先进行模式识别。例如,我们要去听课,就必须对课程表作文字识别;为了找到教室和座位,还要进行景物分析;而听课本身就是在做声音识别。再如一个医生给病人看病,他首先要了解病情,再做一些必要的检验,根据找到的能够以此诊断病情的主要特征(如体温、血压、血相等)作出病情的分类决策(也就是诊断),这也是在作模式识别。

对于比较简单的问题,可以认为识别就是分类。例如,对于识别从“0”到“9”这十个阿拉伯数字的课题,可以将它转化为把待识别的字符分到从“0”到“9”这十类中某一类的问题。待识别的字符被模式识别系统分到哪一类(例如“2”那一类),就可以认为它就是那一类所代表的数字。在这个例子中,系统就“识别”出这个字符是阿拉伯数字“2”。但是,对于比较复杂的识别问题,就往往不能用简单的分类来解决,还需要对待识别模式的描述。例如,汉字识别和景物识别

就是如此。

模式识别作为一门技术科学,目的就是要研究出能自动进行模式分类和描述的机器系统,以完成人类的模式识别的功能。可见它的目标同人工智能范畴的其它分支的目标是一致的,都是要用机器来代替人类的部分智力活动。

模式识别是一个新的研究领域,到现在为止,它的理论和技术都远未完善,很多课题有待人们去研究和探索。

模式识别也是一门边缘技术科学。它与人工智能、信号处理、计算机科学技术、概率统计、模糊集论、信息论、数字图像处理、形式语言学、心理学、语言学等都有密切的关系,而且随着模式识别这门学科的发展,还会与其它更多的学科发生关系。另一方面,许多不同学科的人们也从各自的角度对它产生了越来越多的兴趣。

下面以二维图像为例说明模式识别与相邻学科间的关系和区别:

(1) 图像处理,它包括图像增强、图像复原、图像编码、图像分割、边缘检测等,它的输入是图像,输出是经过处理的图像。

(2) 图像识别,它是模式识别的一个分支,输入是图像,输出是图像的分类和结构描述。

(3) 图像理解,它是人工智能的一个分支。它的输入是图像,输出是对图像的描述和解释。

这三门学科是互相区别而又互相联系的。在进行图像识别前先要作图像处理,使原始的图像能满足识别的要求;而对于某些复杂的图像,如果有图像理解的帮助,就会有助于它们的正确识别。从图像处理的方面看来,在图像的完善处理后,人们会很自然地想到对处理过的图像进行识别。而从人工智能的方面看,智能输入是一个人工智能系统的必要组成部分。

1.2 模式识别的发展和应用

随着模式识别的迅速发展,识别技术在越来越多的领域得到应用,下面介绍它的几个主要应用范围和发展概况。

1. 文字识别

迄今为止在模式识别领域中发展得最成熟并得到最广泛应用的一个方面是文字识别。早在1929年Tauschek就试图用模板匹配的方法去识别十个印刷体阿拉伯数字。机器中有十个相应于阿拉伯字符0~9的模板,这些模板类似于照相的负片。当文件上的阿拉伯数字(例如“2”)与模板的“2”字相吻合时,透过模板的光线最少,在模板后面的光接收器输出很小。而当文件上的阿拉伯数字与模板的字符不相吻合时,字符以外纸面反射的白光能部分透过模板的透明部分,使光接收器有较大的输出。对每个待识别的字符,依次试过十块模板,输出最小的模板的数字就是待识别的字符。装置示意图1.1。

字符识别在这个原始方法的基础上不断地加以改进。但这类方法对于字体、位置和印刷油墨、纸张质量等都有特别的要求,终因光—电方法的条件限制而没有得到推广应用。文字(字符)识别的真正发展还是在计算机科学技术有了长足的进步之后。

按识别对象,文字识别分为西文字符,阿拉伯数字和汉字识别等。它们又分为印刷体和手写体的识别。显然,手写体的识别要比印刷体的识别困难得多。由于汉字结构复杂,种类多,它是我国模式识别研究者的重点课题之一。此外,还有利用书写板输入的在线文字识别,它利用汉字的笔顺信息,从而降低了识别的难度。

一般说来,如果对于字符的书写予以一些限制,则识别过程会简单些,所以有的识别系统规定字符要写在指定的方框内,有的系统还在框内加上限制字符形状的格点等。

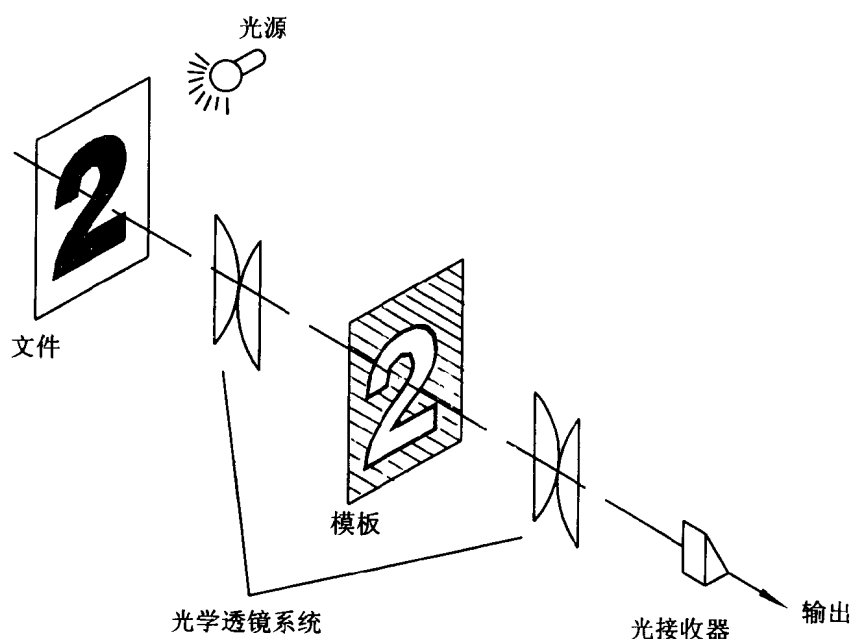


图 1.1

手写体阿拉伯数字的识别在邮政信函自动分拣上起重要的作用。现在信封上的邮政编码就是为自动分拣而设的。

汉字识别的难度较大,这主要因为类别太多、图形复杂的缘故。文字识别可用于各种大、中型计算机的输入,诸如出版印刷、新闻通讯、银行、邮政、资料文献、管理等部门使用的计算机,都有大量文字输入的问题,所以文字识别研究的突破必将显著提高这些部门使用计算机的水平,扩大它的应用范围。

2. 语音识别

语音识别的难度和复杂性都很高,因为要抽取语音的特征,不仅需要分析语言的结构和语音的物理过程,而且还涉及听觉的物理和生理过程。人们可以听懂不同噪音,不同速度的连续语句,但是要用机器来实现,困难就很大。

语音识别有两类课题,第一类是识别人们的语言,它可能是不同人们在不同环境背景下的声音。孤立语音的识别已经取得不少好成绩。语音识别的最终目的是识别连续语音,困难在于连续语音的分割、节拍信息的提取、某些辅音的准确和稳定的检出等,这些问题还没有得到很好的解决。

语音识别的另一类课题是发声者的识别。这个课题的任务是识别出发声者是谁,而对于说话内容则并不感兴趣。这在身份鉴别中能起很大作用。

3. 医学上的应用

模式识别应用于医学问题已取得不少成绩。主要有以下几方面:

(1) 心电图和心电向量图的分析;

- (2) 脑电图的分析；
- (3) 染色体的自动分类；
- (4) 癌细胞分类；
- (5) 血相分析；
- (6) 医学图片的分析,包括 X 光片,CT 片,磁共振片的分析等。

模式识别在医学上的应用很多,前景是很好的。随着卫生保健事业的发展,各种健康普查的工作量也会大大增加,单纯依靠人工来分析显然不能满足要求。所以,模式识别在医学上的应用会越来越广泛和深入。

4. 其它方面的应用

现在遥感技术用多光谱扫描仪及多波段航空摄影可以得到大量的地球图像数据,为人们提供大量的地下矿藏、农作物分布、气象情况、野生动物资源等资料。这些数量庞大的信息,促进了遥感技术的发展,也带来了处理这些信息需要极大的工作量的问题。应用模式识别技术进行遥感图片的分类,可以完成大量的信息处理工作。

模式识别的另一个重要应用方面是军事。可见光、雷达、红外图像的分析与识别的主要目的是检出和鉴别目标的出现,判断目标的类别并对运动中的目标进行监视和跟踪。这方面的识别问题已经受到注意,并且做了一些重要的研究工作。另外,采用地形匹配方法校正飞行轨道以提高导弹命中精度,也是模式识别的一个重要的应用课题。

此外,模式识别在鉴别人脸和指纹、地质勘探、高能物理等方面都有很多用途。作为机器人的视觉系统和听觉系统,模式识别技术在人类不宜工作的环境中,太空、放射性和其它有害污染环境,都可有广泛的应用。

目前对模式识别应用的限制主要是系统价格高昂和复杂的问题,从而向更多的应用领域发展、降低价格和合理简化模式识别系统正是这门新兴学科发展的方向。

1.3 模式识别的研究方法

图 1.2 所示是一个模式识别系统的简单框图,它由预处理、特征或模式基元的抽取和选择以及识别三大部分组成。各部分的作用简单介绍如下:

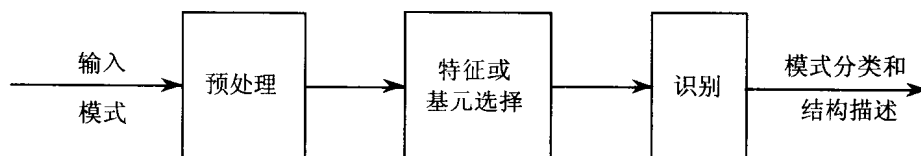


图 1.2

1. 预处理

前面提到,“模式”的概念的内涵是很丰富的,凡是人类能用其感官直接或间接接受的外界信息都可称为模式。所以,一个模式识别系统的输入就可能是各种不同介质上的信息。因为现代模式识别技术是建立在使用数字电子计算机的基础上的,所以对于非电量输入模式,必须首先把它们转换成电信号,然后通过模/数转换,使之成为数字计算机能接受的数字量。为了使输

入模式满足识别的要求,还要根据具体情况对模式进行处理,如滤波、坐标变换、图像增强、图像复原、区域分割、边界检测、骨架提取等,以减少外界干扰和噪声的影响,使模糊的模式变清晰,以便抽取模式识别所需要的特征。这是模式识别过程的一个重要的环节。

2. 特征和模式基元选择

经过预处理,满足识别要求的模式要根据识别方法的要求抽取选择特征和基元,作为识别的依据。一般说来,要求选择出来的特征和基元能够足够代表这个模式,另一方面要求它们的数量尽量少,从而能有效地进行分类和描述。模式特征和基元的选择对识别的效果有直接的影响,所以它们的选择是模式识别的关键,但是,目前还没有一个有效的,一般的抽取、选择特征和基元的方法。模式识别使用的预处理方法仍是非常“面对问题”的,对于不同的识别对象要用不同的方法抽取特征和基元,而且往往不止一种方法。

3. 识别

模式识别主要用两类方法进行,即决策论方法(统计方法)和结构(句法)方法。许多具体的模式识别方法都可以归结到这两种方法中来。

(1) 决策论模式识别方法一般称为统计模式识别方法,它是从被研究的模式中选择能足够代表它的若干特征(设有 d 个),每一个模式都由这 d 个特征组成的在 d 维空间的一个 d 维特征向量来代表,于是每一个模式就在 d 维特征空间占有一个位置。一个合理的假设是同类的模式在特征空间相距很近,而不同类的模式在特征空间则相距较远,这是因为相距近的模式意味着它们的各个特征相差不多,从而在同一类中的可能性也较大。如果用某种方法来分割特征空间,使得同一类模式大体上都在特征空间的同一个区域中,对于待分类的模式,就可根据它的特征向量位于特征空间中哪一个区域而判定它属于哪一类模式。决策论模式识别的任务就是用不同的方法划分特征空间,使得识别的目的能够达到。

(2) 结构方法立足于分析模式的结构信息。至今比较成功的是句法结构模式识别方法。在这个方法中,把模式的分层结构类比于语言中句子的构造,这样,就可利用形式语言学的理论来分析模式。大家知道,句子由单词按文法规则构成。同样,模式由一些模式基元按一定的结构规则组合而成,分析模式如何由基元构成的规则就是结构分析的内容,这相当于在形式语言学中对一个句子作句法分析。句法结构模式识别就是检查代表这个模式的句子,是否符合事先规定的某一类文法规则,如果符合,那么这个模式就属于这个文法所代表的那个模式类。除了分类的信息外,句法结构方法还能给出模式的结构信息。

这两种识别方法是并行不悖的,是可以取长补短,互相补充的。统计方法发展较早,取得了不少应用成果,但是它对模式本身的结构关系很少利用,而很多识别问题,并不是用简单的分类就能解决的,往往更重要的是要弄清楚这些模式的结构关系。另一方面,单纯的句法结构方法没有考虑到模式所受到的环境噪声的干扰等不稳定因素的影响。因此,把这两种方法结合起来,是研究模式识别问题的一个方向。在这方面,提出了随机文法、属性文法等一些新的研究方法,取得了一定的成果。

要指出的是,句法结构模式识别方法中关键的基元选择是要用统计识别方法得到的。所以可以说,统计方法是模式识别的基础。

近来,在模式识别中也应用了模糊数学的方法,取得良好效果。对于某些复杂的模式识别问题,可以采用人工智能中的一些方法,尤其是与专家系统相结合,研究出新的“基于知识”的模式识别方法。本书将要提到,句法方法在构造基于知识的模式识别系统中能起很大的作用。

神经网络作为一个有“学习能力”的工具，在模式识别中也得到越来越多的应用。本书将介绍用模式识别的观点对神经网络作出解释，使得神经网络具有透明性和便于应用。这些方法对模式识别学科都有很大的推进。要指出的是，无论是用模糊集论还是神经网络或是采用基于知识的模式识别方法，它们仍可归结为以上两种基本的模式识别方法或它们的结合。

第 2 章 线性判别函数

2.1 线性判别函数和决策面

线性判别函数是决策论模式识别方法中的一个重要的基本方法。许多其它决策论识别方法也可用判别函数来研究,它也是研究神经网络的基础。本书就从线性判别函数开始介绍统计模式识别的各种方法。

在统计模式识别方法中,首先应把能代表模式的那些特征抽取出来,构成一个代表这个模式的特征向量。关于模式特征的抽取和选择问题,我们将在第六章中介绍。现在假定已经抽取到模式的若干特征,并且从中选择出 n 个最重要的:

$$x_1, x_2, \dots, x_n$$

如果这 n 个特征能够较好地描述原始的待识别的模式,则这个模式就可以用 n 维空间的一个列向量来代表,即:

$$\mathbf{X} = (x_1, x_2, \dots, x_n)^T$$

当我们观察待分类模式时,每次观察到的样本可能都是不同的,它们可以看成是随机产生的。所以每次抽取到的模式特征都应看成是随机变量,从而代表这些模式的 n 维向量也应是随机向量。由于假设抽取到的特征能够很好地代表这个模式,并且这些模式服从一定的总体分布规律,所以在模式空间这些随机的特征向量也应该服从模式的总体分布。如果样本数很大,则样本的分布就能相当可靠地反映这个总体分布规律。

图 2.1 表示一个简单的二维两类模式的例子。两个分布 I 和 II 代表两类模式的分布。如果从过去的经验我们可以知道这两个分布规律,就可以在这两个分布之间建立起一个边界 $g(\mathbf{X})=0$,它把这个二维空间分为两个部分,分别对应于两类模式。现在,如果有一个待识别的模式落在这个二维空间的某一点上(例如 $\mathbf{X}_1$),则只要能判断出这个点是在哪一类模式的范围

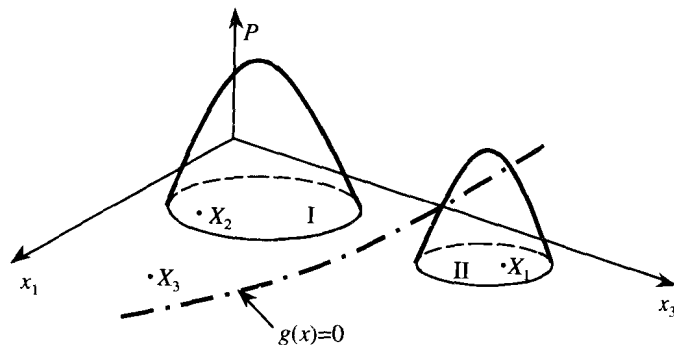


图 2.1

内,就有理由把这个模式分到那一类去,这是因为在同一范围内各个模式的特征比较相近的缘故。例如,在图 2.1 中,我们只要找出 I 类和 II 类模式的分类边界 $g(X)=0$,就可以将待分类的模式 X_1 分到 II 类去,而将模式 X_2 和 X_3 分到 I 类去。

多类别的情况与二类情况相似。图 2.2 是一个二维三类模式的例子。已经知道有三类模式 ω_1, ω_2 和 ω_3 ,它们在二维空间中的位置分别用符号“ $\times$ ”、“ $\circ$ ”和“ $\triangle$ ”表示。三类模式可用边界线 AD, BD 和 CD 分开。可以看出,待分类模式 X_1, X_2 和 X_3 可以分别分到模式类 ω_1, ω_3 和 ω_2 中去。

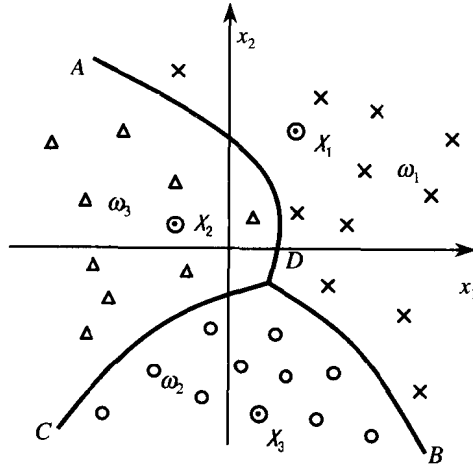


图 2.2

所以,如果根据以往大量的观察,知道模式类别的分布,从而能找出 n 维空间中模式类之间的分界,就能解决模式的分类问题。这实际上是一个通过给定样本的学习过程。本章中我们首先研究最简单的模式分类问题,即假设抽取到的模式样本的边界是“整齐”而不混杂的,而且以后遇到的待分类模式基本上不超过学习样本的分布范围,从而利用这些样本得出的分类边界是无误差的。为了找出这些模式类之间的分界面,可以利用判别函数来进行。

对于 n 维空间中的 c 个模式类别各给出一个由 n 个特征组成的单值函数,这叫做判别函数。在 c 类的情况下,我们共有 c 个判别函数,记为 $g_1(X), g_2(X), \dots, g_c(X)$,它们分别对应于模式类 $\omega_1, \omega_2, \dots, \omega_c$ 。

作为判别函数,它应具有如下的性质:假如一个模式 X 属于第 i 类,则有:

$$g_i(X) > g_j(X), \quad i, j = 1, 2, \dots, c, \quad j \neq i \quad (2.1)$$

而如果这个模式在第 i 类和第 j 类的分界面上,则有:

$$g_i(X) = g_j(X) \quad (2.2)$$

事实上,这是由 n 维模式降为一维或一个数的一种变换。

最简单的判别函数是线性判别函数,它是所有模式特征的线性组合。对于第 i 类模式,它有如下的形式:

$$g_i(X) = w_{i1}x_1 + w_{i2}x_2 + \dots + w_{in}x_n + w_{i0}, \quad i = 1, 2, \dots, c$$

或

$$g_i(X) = \sum_{k=1}^n w_{ik}x_k + w_{i0} \quad (2.3)$$

式中 w_{ik} 是特征的系数, 称为权, w_{i0} 为阈值权。这个式子说明, 每个特征对于判别函数作出不同的贡献, 贡献的大小就是它的权。

如果对第 i 类模式定义 n 维权向量为:

$$\mathbf{W}_i = (w_{i1} w_{i2} \cdots w_{in})^T \quad (2.4)$$

则判别函数可写成更简洁的形式:

$$g_i(\mathbf{X}) = \mathbf{W}_i^T \mathbf{X} + w_{i0}, \quad i = 1, 2, \cdots, c \quad (2.5)$$

判别函数不一定是线性的, 例如, 在判别函数中加上 $x_i x_j$ 项, 就得到二次判别函数:

$$g_i(\mathbf{X}) = \sum_{k=1}^d w_{ik} x_k + \sum_{l=1}^d \sum_{m=1}^d w_{ilm} x_l x_m + w_{i0}$$

式中 $x_l x_m = x_m x_l$ 。不失一般性, 可设 $w_{lm} = w_{ml}$, 这样, 对于同样数量的特征, 二次判别函数的系数就比线性判别函数多出 $\frac{n(n+1)}{2}$ 个。若用三次判别函数, 在 $g_i(\mathbf{X})$ 中再加入 $x_i x_j x_k$ 项, 则问题会变得更为复杂。

一般说来, 高阶判别函数可以提供超曲面的分界面, 但系数会增加许多, 甚至问题变成实际不可行的。所以在许多情况下, 还是尽量采用线性判别函数, 以及由它派生的分段线性判别函数。

在一些特殊例子中, 可以通过某种变换将非线性判别函数映射成线性的。例如:

设 $g(x) = a_1 + a_2 x + a_3 x^2$ 是一个二次判别函数, 定义三维向量

$$\mathbf{Y} = \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}, \quad (2.6)$$

则 $g(x)$ 可写成

$$g(x) = \mathbf{A}^T \mathbf{Y}$$

它就是一个关于 $\mathbf{Y}$ 是线性的判别函数了。要注意的是, 本例中 $g(x)$ 原来只是一个特征的判别函数, 这个映射导致维数的大量增加。

此外, 在一般情况下, 问题所需要的非线性映射是不容易得到的。本章将着重讨论线性判别函数的性质。

下面讨论二类情况下的线性判别函数。

在只有二类模式的情况下, 用 $c=2$ 展开式(2.5), 我们可以得到两个线性判别函数, 即:

$$g_1(\mathbf{X}) = \mathbf{W}_1^T \mathbf{X} + w_{10}$$

$$g_2(\mathbf{X}) = \mathbf{W}_2^T \mathbf{X} + w_{20}$$

如果 $\mathbf{X}$ 属于 ω_1 , 根据式(2.1), 有

$$g_1(\mathbf{X}) > g_2(\mathbf{X})$$

即:

$$\mathbf{W}_1^T \mathbf{X} + w_{10} > \mathbf{W}_2^T \mathbf{X} + w_{20}$$

或写成:

$$(\mathbf{W}_1^T - \mathbf{W}_2^T) \mathbf{X} + (w_{10} - w_{20}) > 0 \quad (2.7)$$

令 $\mathbf{W}^T = \mathbf{W}_1^T - \mathbf{W}_2^T$, $w_0 = w_{10} - w_{20}$, 得

$$g(\mathbf{X}) = \mathbf{W}^T \mathbf{X} + w_0$$