国家精品课程教材

新编《信息、控制与系统》系列教材

模式识别（第三版）

Pattern Recognition (Third Edition)

张学工 编著
Zhang Xuegong

清华大学出版社

国家精品课程教材

新编《信息、控制与系统》系列教材

模式识别（第三版）

Pattern Recognition (Third Edition)

张学工 编著
Zhang Xuegong

清华大学出版社
北京

内 容 简 介

本书是清华大学自动化系国家精品课程“模式识别基础”的教材，是在《模式识别》第一版和第二版基础上重写而成的。本教材系统地讨论了模式识别的基本概念和代表性方法，包括监督模式识别中的贝叶斯决策理论、概率密度函数的估计、线性判别函数、非线性判别函数、近邻法、特征选择与提取的典型方法以及非监督模式识别中的基于模型的方法、混合密度估计、动态聚类方法、分级聚类方法等，并在相应章节包括了人工神经网络、支持向量机、决策树与随机森林、罗杰斯特回归、Boosting 方法、模糊模式识别等较新进入模式识别领域的内容。整体内容安排力求系统性和实用性，并覆盖部分当前研究前沿。

本书可以作为高等院校自动化、计算机等相关专业高年级本科生和研究生学习模式识别的教材，也可以供计算机信息处理、生物信息学、数据挖掘、统计等各领域中从事模式识别相关工作的广大科技人员和高校师生参考。

图书在版编目(CIP)数据

模式识别/张学工编著. —3 版. —北京：清华大学出版社，2010.8
(新编《信息、控制与系统》系列教材)
ISBN 978-7-302-22500-3

Ⅰ. ①模… Ⅱ. ①张… Ⅲ. ①模式识别 Ⅳ. ①TP391.4

中国版本图书馆 CIP 数据核字(2010)第 067622 号

责任编辑：王一玲
责任校对：梁　毅
责任印制：李红英
出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　**邮　　编**：100084
社　总　机：010-62770175　　**邮　购**：010-62786544
投稿与读者服务：010-62795954，jsjjc@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：北京市清华园胶印厂
经　　销：全国新华书店
开　　本：185×260　**印　张**：16　**字　数**：370 千字
版　　次：2010 年 8 月第 3 版　　**印　　次**：2010 年 8 月第1 次印刷
印　　数：1～3000
定　　价：25.00 元

产品编号：023906-01

新编《信息、控制与系统》系列教材
出 版 说 明

信息、控制与系统学科是在20世纪上半叶形成和发展起来的一门新兴技术科学。在人类探索自然和实现现代化的进程中，信息、控制与系统学科的理论、方法和技术始终起着重要的和基础的作用。基于信息、控制与系统科学的自动化的发展和应用水平在一定意义上是一个国家和社会的现代化程度的重要标志之一。本系列教材是关于信息、控制与系统学科所属各个领域的基本理论和前沿技术的一套高等学校系列教材。

本系列教材所涉及的范围包括信号和信息处理、模式识别、知识工程、控制理论、智能控制、过程和运动控制、传感技术、系统工程、机器人控制、工业自动化、计算机控制和仿真、网络化系统、电子技术等方面。主要读者对象为自动控制、工业自动化、计算机科学和技术、电气工程、机械工程、化工工程和热能工程等专业有关的高年级大学生和研究生，以及工作于相应领域和部门的科学工作者和工程技术人员。

十多年前，清华大学出版社会同清华大学自动化系，曾经组编出版过一套《信息、控制与系统》系列教材，产生了较大的社会影响，其中多数著作获得了包括国家级教学成果奖和部委优秀教材奖在内的各种奖励，至今仍为国内众多院校所采用，并被广大相关领域科技人员作为进修和自学读物。我们现在组编的这套新编《信息、控制与系统》系列教材，从一定意义上说，就是先前那套教材的延伸和发展，以反映近年来学科的发展和在科学研究与教学实践上的新成果和新进展，以适应当前科技发展和教学改革的新形势和新需要。列入这套新编系列教材中的著作，大多是清华大学自动化系开设的课程中经过较长教学实践而形成的，既有多年教学经验和教学改革基础上的新编著的教材，也有部分原系列教材的更新和修订版本。这套新编系列教材总体上仍将保持原系列教材求新与求实的风格，力求反映所属领域的基本理论和新近进展，力求做到学科先进性和教学适用性的统一。需要说明的是，此前我们曾以《信息技术丛书》为名组编这套教材，并已出版了若干种著作。现为使“书”和“名”更为相符，这些已出版的著作将在重印或再版时列入这套新编系列教材。

我们希望，这套新编系列教材，既能为在校大学生和研究生的学习提供内容先进、论述系统和教学适用的教材或参考书，也能为广大科学工作者与工程技术人员的知识更新与继续学习提供适合的和有价值的进修或自学读物。我们同时要感谢使用本系列教材的广大教师、学生和科技工作者的热情支持，并热忱欢迎提出批评和意见。

新编《信息、控制与系统》系列教材编委会

2002年6月

新编《信息、控制与系统》系列教材编委会

序

张学工教授长期从事模式识别课程的教学和科研工作，取得了优异的成绩，积累了丰富的经验。他还是国内首先著文介绍和翻译 Vapnik 统计学习理论著作的作者，这为他编写新一版的《模式识别》教材打下了坚实的基础。新版的《模式识别》取材更加精炼，安排更加符合学生学习的规律，特别是把传统的统计模式识别方法与人工神经网络和支持向量机有机地结合起来，使学生能够更好地掌握三者的内在联系，进一步理解学习样本集对于设计模式识别系统的重要性，对于以后在实际应用中确定合适的方法有很好的指导意义。

模式识别学科的发展和模式识别在各个学科的应用研究紧密联系在一起。张学工教授多年从事地震勘探信号识别和生物信息学领域的科研工作，对模式识别方法的实际应用有深入的研究，部分的研究成果在本书中有所反映，这使得本书在理论联系实际方面比原有版本有很大的改进。我相信本书的出版不仅能提高国内模式识别课程教学水平，而且对广大科技工作者更好地把模式识别方法应用于研究工作中也会起到很好的推动作用。

边肇祺　教授

2010 年 6 月 19 日

前　言

从本书第二版出版到现在已经又是十年了。在这十年里，我们真切地感受到了信息时代的到来。对信息的处理和分析，已经不仅仅是信息科学家所关心的问题，也不仅仅是信息技术产业所关心的问题，而是为很多学科和很多领域共同关心的问题。作为信息处理与分析的重要方面，模式识别也开始从一个少数人关心的专业，变成一个在工程、经济、金融、医学、生物学、社会学等各个领域都受到关注的学科。

模式识别学科的发展，可以从笔者所在的清华大学自动化系在模式识别专业教学和教材上的沿革窥见一斑。早在 1978 年，在已故中科院学部委员常迥教授的领导下，自动化系成立了信号处理与模式识别教研组，后更名为信息处理研究所，1981 年获准成立“模式识别与智能系统”学科（当时称“模式识别与智能控制”）的第一个硕士点、博士点。从那时起，边肇祺等教授就开始为研究生开设模式识别课程，后逐渐包括进少部分五年级本科生（当时清华大学本科学制为五年）。80 年代中期，边肇祺、阎平凡、杨存荣、高林、刘松盛和汤之永等老师组成了教材编写小组，开始编写模式识别教材，这就是 1988 年出版的《模式识别》（第一版）。该教材的出版，为我国模式识别学科的发展做出了历史性的贡献，被很多高校和科研院所作为教材或参考书。十年以后，模式识别学科的内容有了很多更新和发展，我们成立了由边肇祺、阎平凡、赵南元、张学工和张长水组成的改写小组，由笔者与边肇祺老师共同组织编写了本书的第二版，2000 年正式出版。此时的模式识别课程，已经由最初只有十几位研究生参加的小课，发展为由上百名研究生和高年级本科生参加的大课。第二版教材也得到了国内同行的欢迎，9 年内已经重印 15 次。

随着模式识别学科的日益发展，我们很快认识到，对模式识别课程的需求已经超出了本专业研究生的范围。于是我们将模式识别课程分为两门：面向研究生的“模式识别”和面向本科生的“模式识别基础”。到今天，本科生“模式识别基础”每年的选课人数也已达到 100～150 人，除了来自本系的学生，每年还有多位来自其他院系的学生选课。2007 年，该课程荣幸地被评为国家级精品课程。

在近几年的教学实践中，我们体会到，原来的教材有些地方不太适应大范围教学的需要，而且近十年来模式识别自身以及它在很多领域中的应用又有了很多新发展。因此，笔者

从两年前开始着手编写新版教材。新版教材的出发点是：一方面，结合当前的最新发展，精炼传统内容，充实新内容，进一步增强实用性，接触学科前沿；另一方面，在教材的深度和广度上兼顾广大本科生学习的特点和本专业研究生的需求，力求达到使非本专业学生通过本教材能学到足够系统的基本知识，而本专业学生又能以本教材作为其专业研究的重要起点。

编写新版教材所需要的时间超出了我的预想，很高兴她今天终于能和读者见面了。在此要感谢在本书编写过程中给了我很多帮助的同事和同学们，尤其是：美国南加州大学的 Jasmine X. Zhou 教授在 2007 年给我提供了短期访问机会，使我能够有一段相对完整的时间集中开始本书的写作；蒋博同学通读了本书三分之二的初稿并作了多处补充；现在已经分别是电子科技大学和北京大学教师的凡时财、李婷婷同学帮助准备了本书部分素材。我还要感谢清华大学出版社王一玲编辑在本书编写过程中的一贯支持。当然，最重要的，我要感谢参加本书第一版和第二版编写的所有老师，这不但是因为在这一版中仍使用了前两版的一些内容，更是因为，是这些老师们把我带进了模式识别的大门，使我受益至今。

由于时间仓促和个人水平所限，教材中难免有错误或不足之处，敬请广大同行和读者批评指正，相关内容请发电子邮件到 zhangxg@tsinghua.edu.cn，以便在再版时补充和修改。

在本书最终完稿的时候，我十岁的女儿以极大的兴致看完了我讲“模式识别基础”第一课的录像，并说将来长大了要听我讲课。谨以此书献给我的妻子和女儿。

张学工

2009 年 11 月 29 日

目　录

第 1 章
概　论

1.1　模式与模式识别

人类智慧的一个重要方面是其认识外界事物的能力。这些能力可能是从一个人的孩童时期就具备并且不断增强的，并且这种能力在很多动物身上也不同程度地存在。人们往往对这种能力习以为常，并意识不到它是复杂的智能活动的结果。但是，如果仔细分析我们日常所进行的很多活动，就会发现，几乎每一项活动都离不开对外界事物的分类和识别。

例如，当看到图 1-1 的照片时，很可能会得出这样的印象或结论：这是一幅风景照片，表现的是中国某一江南水乡的景色。这一看似简单的认知过程实际上是由一系列对事物类别的识别构成的，比如，我们会识别出这是一幅照片（而不是绘画），是一幅风景照片（而不是人物或其他照片），照片中有小河、房屋、游船等。进一步，这种傍水而建的民居建筑风格让我们在这些具体的观察之上形成了“这是江南水乡”的判断。在整个过程中，照片、风景照片、小河、房屋、游船、江南水乡等都是代表着客观世界中的某一类事物的概念，人们对这些概念的识别并不是依靠对每一个具体对象的记忆，而是依靠在以往对多个此类事物的具体实例进行观察的基础上得出的对此类事物整体性质和特点的认识。比如，这幅照片中的游船或许和我们以前见过的任何游船都不完全一样，但是由于我们见过很多游船，在头脑中已经形成了对“游船”这一类事物所具有的特征的认识，因此，尽管这些游船我们并没有见过，我们仍然能毫不困难地识别出它们是游船。换句

图 1-1　一幅风景照片

话说，我们已经通过以往看到很多的游船在头脑中形成了“游船”这一类事物的一种模式，当看到新的游船时，我们能把这种模式识别出来。这就是每个人每天都在大量进行的模式识别活动的一个简单例子。同样地，我们从这幅照片中看到小河、房屋，是对“小河”、“房屋”模式的识别；而对于这是一张风景照片、照片中的风景是江南水乡这种更抽象的判断，则是在对具体物体的识别的基础上形成的更上一层的模式识别。

图 1-2 是一段心电图信号的片段。多数读者可能只能认出这是心电图，而有一定医学知识的读者则能从这个信号片段中找到所谓的 T 波、P 波、U 波等不同的部分，这些都是心电图信号中的特殊模式。根据这些模式，有经验的医生还可以通过心电图判断病人的心脏健康状况，进行更高层次的模式识别。现代生物医学的发展为临床医生提供了大量的检验手段，从宏观到微观，从医学影像到基因和蛋白质标记物的表达，应有尽有，而医生根据这些结果对疾病进行诊断就是在进行对疾病的模式识别。

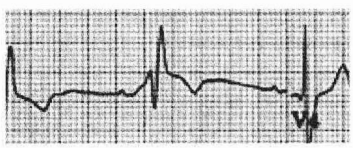

图 1-2 一段心电图片段

人们对外界事物的识别，很大部分是把事物按分类来进行的。比如，看到一幅照片，我们很自然地就会知道这是一幅风景照片还是人像照片，这实际上是把各种照片分成了若干类，包括风景、人像、体育、新闻等，在看到一幅具体的照片时我们就把它归到其中的某一类（或同时归属某几类）。事实上，我们对外界对象的几乎所有认识都是对类别的认识，在我们的心目中，“房子”的概念代表的是一类对象而不是一座座具体的房子，“人”、“大人”、“小孩”、“男生”、“女生”、“树”、“花”、“草”、“桌子”、“椅子”、“床”等，都是类别的概念。极端来说，我们认识每一个人其实也都是作为一个类别来认识的，因为我们此时看到的张三（即他在我们视网膜上的成像）与彼时他的模样是不完全一样的，此时听到的他的声音和彼时听到的他的声音也是不完全一样的，之所以我们能够把这些不同的图像、不同的声音都识别为张三，就是因为我们在大脑中已经形成了关于他的一种模式，只要符合这种模式的图像和声音就会被划分为“张三”这一类。

我们对外界对象的类别判断并不限于直接从五官获得的信号，也存在于很多更高级的智能活动中。

我们在与人交往的过程中，会通过对每个人多方面特点的观察逐步形成一些对他们的看法，比如觉得某个人很聪明，某个人很可亲，某个人很难相处等，这也是一种对模式的识别，只不过这些模式的定义更模糊、更抽象。

在金融行业，一个成功的信用卡公司往往需要通过认真分析客户的信用资料和消费习惯等将用户分类，以便更好地判断用户的信用程度，并且可以通过消费模式的突然变化检测可能的信用卡盗用等行为；一个好的保险公司也需要根据客户的收入、职业、年龄、受教育程度、健康状况、家庭情况、行为记录等将客户细分，以便更有针对性地对客户提供最恰当的保险产品。

在人来人往的公共场所，训练有素的反扒警察可以很准确地发现正在伺机作案的扒手，靠的正是对这些人与常人不同的行为模式来识别的。

模式识别一词的英文是 pattern recognition。在中文里，“模”和“式”的意思相近。根据《说文》，模，法也；式，法也。因此，模式就是一种规律。英文的 pattern 主要有两重含义，一是代表事物（个体或一组事物）的模板或原型，二是表征事物特点的特征或性状的组合。在

模式识别学科中，模式可以看作是对象的组成成分或影响因素间存在的规律性关系，或者是因素间存在确定性或随机性规律的对象、过程或事件的集合。也有人把模式称为模式类，模式识别也被称作模式分类(pattern classification)。

在《说文》中，识，知也；别，分解也。识别就是把对象分门别类地认出来。在英文中，识别(recognition)一词的主要解释是对以前见过的对象的再认识(re-cognition)。因此，模式识别就是对模式的区分和认识，把对象根据其特征归到若干类别中适当的一类。

从前面的举例我们可以看到，人类智能活动中包含大量的模式识别活动。作为一门学科，模式识别所研究的重点并不是人类进行模式识别的神经生理学或生物学原理，而是研究如何通过一系列数学方法让机器(计算机)来实现类似人的模式识别能力。

为了在本书后面章节中讨论方便，我们在这里把一些基本术语的含义约定一下。这些术语在其他文章或书籍中的含义和用法可能会略有不同，但只要参考上下文就不难明确其确切含义。

样本(sample)：所研究对象的一个个体。注意，这与统计学中通常的用法不同，相当于统计学中的实例(example 或 instance)。

样本集(sample set)：若干样本的集合。统计学中的“样本”通常就是指样本集。

类或类别(class)：在所有样本上定义的一个子集，处于同一类的样本在我们所关心的某种性质上是不可区分的，即具有相同的模式。习惯上，我们经常用 ω_1、ω_2 等来表示类别，在两类分类问题中也有时用{−1，1}或者{0，1}等来表示。

特征(feature)：指用于表征样本的观测，通常是数值表示的某些量化特征，有时也被称作属性(attribute)。如果存在多个特征，则它们就组成了特征向量。样本的特征构成了样本的特征空间，空间的维数就是特征的个数，而每一个样本就是特征空间中的一个点。某些情况下，对样本的原始描述可能是非数值形式的，此时通常需要采用一定的方法把这些特征转换成数值特征。在本书中，除特别说明外，特征都是指取值为实数的数量特征。

已知样本(known sample)：指事先知道类别标号的样本。

未知样本(unknown sample)：指类别标号未知但特征已知的样本。

所谓模式识别的问题就是用计算的方法根据样本的特征将样本划分到一定的类别中去。

1.2 模式识别的主要方法

解决模式识别问题的方法可以归纳为基于知识的方法和基于数据的方法两大类。

所谓基于知识的方法，主要是指以专家系统为代表的方法，一般归在人工智能的范畴中，其基本思想是，根据人们已知的(从专家那里收集整理的)关于研究对象的知识，整理出若干描述特征与类别间关系的准则，建立一定的计算机推理系统，对未知样本通过这些知识推理决策其类别。

句法模式识别(syntax pattern recognition)也可以看作是一种特殊的基于知识的模式识别方法。它的基本思想是，把对象分解描述成一系列基本单元，每一个基本单元表达成一定的符号，而构成对象的单元之间的关系描述成单元符号之间的句法关系，利用形式语言、

句法分析的原理来实现对样本的分类。

另一大类模式识别方法是基于数据的模式识别方法。

在确定了描述样本所采用的特征之后，这些方法并不是依靠人们对所研究对象的认识来建立分类系统(在很多情况下人们是不具备这样的认识的)，而是收集一定数量的已知样本，用这些样本作为训练集(training set)来训练一定的模式识别机器，使之在训练后能够对未知样本进行分类。这种模式识别方法可以看作是基于数据的机器学习(machine learning)的一种特殊情况，学习的目标是离散的分类，这也是机器学习中研究最多的一个方向。

图1-3给出了这种机器学习系统的基本思想。

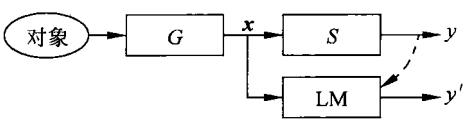

图1-3 基于数据的机器学习

在图1-3中，G表示从对象观测特征的过程，特征用向量$\boldsymbol{x}$表示，y表示我们所关心的对象的性质，在模式识别中就是分类。S表示决定$\boldsymbol{x}$和y之间关系的系统，它存在但我们不知道其内部机理(如果知道就可采用基于知识的方法)。我们可以得到一定数量的已知样本，即一定数量的$\boldsymbol{x}$和对应的y的数据对$\{(\boldsymbol{x},y)\}$。基于数据的模式识别就是利用这样的训练样本来训练学习机器LM，也就是建立实现从特征向量$\boldsymbol{x}$判断类别y'的一个数学模型，用来对未知样本计算(预测)其类别。

基于数据的方法是模式识别最主要的方法，在无特别说明的情况下，人们说模式识别通常就是指这一类方法，其任务可以描述为：在类别标号y与特征向量$\boldsymbol{x}$存在一定的未知依赖关系、但已知的信息只有一组训练数据对$\{(\boldsymbol{x},y)\}$的情况下，求解定义在$\boldsymbol{x}$上的某一函数$y'=f(\boldsymbol{x})$，对未知样本的类别进行预测。这一函数叫做分类器(classifier)。这种根据样本建立分类器的过程也称作一种学习过程。

基于数据的模式识别方法，基础是统计模式识别，即依据统计的原理来建立分类器，这也是本教材的主要内容。通常，人们说模式识别方法主要是指统计模式识别方法。

统计模式识别方法中的线性判别函数等内容诞生于20世纪30年代，而整个模式识别学科从20世纪60年代起得到了很大的发展，逐渐形成比较完整的体系。这是一个很年轻的学科，其发展与现代计算机技术的发展相辅相成，近二十年来仍然不断有新的方法和理论涌现出来，并且应用领域日益扩大。同时，这也是一个学科高度交叉的领域，一大批来自统计学、数学、自动化与计算机科学、工程技术以至心理学和神经生理学领域的科学家活跃在其理论、方法和应用研究的各个方面。

近二十年来发展起来的模式识别方法新成员中，最有代表性的是20世纪80年代中期发展起来的人工神经网络，和以20世纪90年代中期出现的支持向量机为代表的统计学习理论与核函数方法。它们虽然有相对独立的理论体系，但是与传统的统计模式识别有密切的联系，我们也把它们纳入到本教材的体系中。

基于数据的模式识别方法(以后就简称模式识别方法)，适用于我们已知对象的某些特征与我们所感兴趣的类别性质有关系，但无法确切描述这种关系的情况。图1-4示意了模式识别研究的范畴。之所以无法确切地描述这种关系，一方面可能是因为目前对相关机理的研究还比较初步，不足以揭示所研究类别的内在规律；另一方面则可能是由于问题本身的不确定性、样本间的异质性和观测数据的不准确等造成的。如果分类和特征之间的关系可以完全确切地描述出来，那么采用基于知识的方法可能会更有效；而如果二者的关系完