“十二五”国家重点图书
Springer 精选翻译图书

模式识别：算法及实现方法

Pattern Recognition: An Algorithmic Approach

[印] M. Narasimha Murty
V. Susheela Devi 著
王振永 译

哈尔滨工业大学出版社
HARBIN INSTITUTE OF TECHNOLOGY PRESS

“十二五”国家重点图书
Springer 精选翻译图书

模式识别：算法及实现方法

Pattern Recognition: An Algorithmic Approach

[印] M.Narasimha Murty
V.Susheela Devi 著
王振永 译

哈爾濱工業大學出版社
HARBIN INSTITUTE OF TECHNOLOGY PRESS

内容简介

本书主要介绍模式识别的基本概念与算法，全书分为 11 章，内容包括：模式识别概述、模式的表示、最近邻分类器、贝叶斯分类器、隐式马尔可夫模型、决策树、支持向量机、组合分类器、聚类方法等。希望本书有助于读者更好地理解模式识别技术以及该技术对各个领域的重要作用。本书包含了大量的工作实例，安排了适量的练习，提供了丰富的延伸阅读材料。希望每一位读者都能从中受益。

本书适用于电子信息、计算机、自动控制等专业的本科生和研究生及本领域的研究者。

黑版贸审字 08－2017－059

Translation from English language edition:
Pattern Recognition: An Algorithmic Approach
by M. Narasimha Murty and V. Susheela Devi

图书在版编目(CIP)数据

模式识别：算法及实现方法/(印)纳拉辛哈·穆尔蒂(M. Narasimha Murty)，(印)苏席拉·提毗(V. Susheela Devi)著；王振永译.
—哈尔滨：哈尔滨工业大学出版社，2017.10
ISBN 978－7－5603－6327－1

Ⅰ.①模… Ⅱ.①纳… ②苏… ③王… Ⅲ.①模式识别
Ⅳ.①TP391.4

中国版本图书馆 CIP 数据核字(2016)第 285023 号

电子与通信工程
图书工作室

责任编辑　李长波
封面设计　高永利
出版发行　哈尔滨工业大学出版社
社　　址　哈尔滨市南岗区复华四道街 10 号　邮编 150006
传　　真　0451－86414749
网　　址　http://hitpress.hit.edu.cn
印　　刷　哈尔滨市经典印业有限公司
开　　本　660mm×980mm　1/16　印张 16　字数 290 千字
版　　次　2017 年 10 月第 1 版　2017 年 10 月第 1 次印刷
书　　号　ISBN 978－7－5603－6327－1
定　　价　40.00 元

译者序

模式识别作为一种依据事物所抽象出的统计信息的数据分类方法，是信息科学和人工智能的重要组成部分。模式识别在生物信息学、心理分析、生物识别技术和许多其他应用中有重要作用。但是目前国内相关的经典书籍并不多。为了方便广大学者在这一领域进行更广泛和深入的研究，译者在查阅了相关英文原版书籍后，决定将印度学者 M. Narasimha Murty 和 V. Susheela Devi 的著作《模式识别》(Pattern Recognition：An Algorithmic Approach)引入我国。

本书不仅包含模式识别的基本概念和基础知识，同时包含大量的实例。这使得本书既适合于相关学科的本科生、研究生学习，适合作为相关专业的教材进行课堂讲授，各章末的延伸阅读材料和参考文献又适合于本领域的研究者阅读。

本书的翻译工作由哈尔滨工业大学电子与信息工程学院王振永老师及其研究团队共同完成。其中王振永翻译了全书，并负责全书的统稿、修改与校对工作，对内容进行了反复修改和推敲，以提高本书的可读性，并对原书中存在的某些疏漏进行了修订。本书的出版要感谢崔晨、田园、王洪云和袁泉这四位学生，他们在专业术语翻译、公式符号的计算机录入以及校对等方面付出了大量的时间和精力。感谢顾学迈教授和郭庆教授对本译著部分内容提出的建设性意见，感谢李德志博士对本书第 5 章提出的修改意见和唐弢博士对本书第 8 章给予的帮助。

本书的翻译是在国家自然科学基金(No. 61601147，No. 61571316)支持下完成的，特此感谢；还要感谢哈尔滨工业大学提供的各种设施，保证了本书翻译所需的各种资源。

最后，由于许多专业术语还没有统一的中文译法，因此本书的术语除了借鉴于李晶皎等翻译的《模式识别》(第四版)和李宏东等翻译的《模式分类》(原书第二版)外，其余未有标准中文译法的专业术语我们依据其物理

含义及中文习惯给出了可以接受的中文术语。在本书的最后还给出了中英文术语的对照表，便于读者进行专业术语查找和比对。

由于译者水平有限，翻译中难免存在疏漏和不当之处，敬请读者批评指正。

本书为尊重原著，所有向量、矢量、矩阵等量均未用黑斜体表示。

译　者

2017 年 1 月于哈尔滨

前　言

作者写这本书的主要目的是为了使本学科的本科生和研究生对模式识别有更清晰的概念。本书不考虑数据的预处理，而是假设模式是用经过恰当预处理技术的数字向量表示的，描述利用数字数据进行重要决策（如分类）的算法。本书有大量的工作实例，并在每章结束时安排了练习。

本书也适用于本领域的研究人员。想对本学科知识有更深入了解的读者可以参阅延伸阅读材料和各章末尾的参考文献。

模式识别在很多领域都有应用，包括地质学、地理学、天文学和心理学。更具体地说，它对生物信息学、心理分析、生物识别技术和许多其他应用有重要作用。作者相信，本书对所有需要应用模式识别技术来解决问题的研究人员都有帮助。

感谢 Vikas Garg，Abhay Yadav，Sathish Reddy，Deepak Kumar，Naga Malleswara Rao，Saikrishna，Sharath Chandra 和 Kalyan 同学对本书的部分内容提出了阅读意见。

M. Narasimha Murty

V. Susheela Devi

目　录

第 1 章　导论 …… 1
1.1　什么是模式识别? …… 2
1.2　模式识别的数据集合 …… 3
1.3　模式识别的理论框架 …… 3
问题讨论 …… 4
延伸阅读材料 …… 5
习题 …… 5
本章参考文献 …… 5
第 2 章　模式集合的表征 …… 7
2.1　模式集合表征的数据结构 …… 8
2.2　模式聚类的表征 …… 14
2.3　相似度度量方法 …… 14
2.4　模式的尺寸 …… 20
2.5　数据集合的抽象 …… 21
2.6　特征提取 …… 23
2.7　特征选择 …… 27
2.8　分类分析方法 …… 36
2.9　聚类分析方法 …… 37
问题讨论 …… 38
延伸阅读材料 …… 38
习题 …… 38
上机练习 …… 40
本章参考文献 …… 40
第 3 章　最近邻分类器 …… 42
3.1　最近邻算法 …… 42
3.2　典型的最近邻算法 …… 44
3.3　最近邻算法在交易数据库中的应用 …… 47
3.4　高效最近邻算法 …… 48
3.5　数据约简方法 …… 57

3.6　原型选择方法 …… 57
问题讨论 …… 68
延伸阅读材料 …… 68
习题 …… 69
上机练习 …… 71
本章参考文献 …… 71
第 4 章　贝叶斯分类器 …… 74
4.1　贝叶斯理论 …… 74
4.2　最小差错率分类器 …… 75
4.3　概率估计方法 …… 77
4.4　与 NNC 方法的比较 …… 79
4.5　朴素贝叶斯分类器 …… 80
4.6　贝叶斯置信网络 …… 84
问题讨论 …… 86
延伸阅读材料 …… 86
习题 …… 86
上机练习 …… 87
本章参考文献 …… 88
第 5 章　隐式马尔可夫模型 …… 89
5.1　面向分类的马尔可夫模型 …… 90
5.2　隐式马尔可夫模型 …… 95
5.3　基于马尔可夫模型的分类方法 …… 100
问题讨论 …… 102
延伸阅读材料 …… 102
习题 …… 102
上机练习 …… 104
本章参考文献 …… 104
第 6 章　决策树 …… 105
6.1　简介 …… 105
6.2　面向模式分类的决策树方法 …… 107
6.3　决策树的构建 …… 111
6.4　节点拆分方法 …… 115
6.5　过度拟合和修剪 …… 118
6.6　决策树归纳实例 …… 119

问题讨论 …… 122
延伸阅读材料 …… 122
习题 …… 123
上机练习 …… 124
本章参考文献 …… 124
第 7 章　支持向量机 …… 126
7.1　简介 …… 126
7.2　学习线性判别函数 …… 132
7.3　神经网络 …… 146
7.4　面向分类的支持向量机 …… 152
问题讨论 …… 158
延伸阅读材料 …… 159
习题 …… 159
上机练习 …… 162
本章参考文献 …… 162
第 8 章　多分类器组合 …… 163
8.1　简介 …… 163
8.2　构建集成分类器的方法 …… 164
8.3　多分类器组合方法 …… 172
问题讨论 …… 175
延伸阅读材料 …… 176
习题 …… 176
上机练习 …… 177
本章参考文献 …… 177
第 9 章　聚类方法 …… 180
9.1　简介 …… 180
9.2　聚类方法的重要性 …… 188
9.3　分级聚类方法 …… 193
9.4　划分聚类方法 …… 199
9.5　大规模数据集合的聚类方法 …… 202
问题讨论 …… 207
延伸阅读材料 …… 208
习题 …… 208
上机练习 …… 210

本章参考文献 …… 210
第 10 章　本书总结 …… 213
第 11 章　应用实例:手写数字识别 …… 215
11.1　数字数据的描述 …… 215
11.2　数据预处理 …… 217
11.3　分类算法 …… 217
11.4　典型模式的选择 …… 217
11.5　识别结果 …… 218
问题讨论 …… 221
延伸阅读材料 …… 222
本章参考文献 …… 222
名词索引 …… 223

第1章 导 论

学习目标：

通过本章的学习，需要掌握：

①能够定义模式识别。

②理解模式识别在不同应用中的重要性。

③能够解释模式识别问题的两种主要模式。

i. 统计模式识别。

ii. 结构模式识别。

模式识别可以定义为一种基于已知知识或者依据模式表述所抽象出的统计信息进行数据分类的方法。

模式识别有很多重要的应用，例如多媒体文档识别（MDR）和自动医学诊断。在进行 MDR 时，必须处理文本、音频和视频数据的集合。文本数据可以由对应一个或多个自然语言的字母和数字组成。音频数据可能是语音或音乐。视频数据可能是一个单一的图像或者图像序列，例如，一个罪犯的照片、指纹以及签名可以作为一个图像出现。同样，也可以使用一系列的图片来记录一个人在机场移动的画面，这样就形成了一个视频。

在一个典型的模式识别应用程序中，需要对原始数据进行处理，将其转换成一种可被机器使用的形式。例如，可以将各种形式的多媒体数据转换成一个由一些特征值组成的向量。在文本中，模式的表示可以是关键词出现的概率。音频数据可以表示为线性预测编码（LPC）的系数。而视频数据则可以转换到变换域来表示，比如小波变换和傅里叶变换。信号处理可以将原始数据转换为矢量数据（这是预处理的部分内容）。本书不会讨论数据的预处理过程。

相反，假设模式是用经过恰当预处理技术的数字向量表示的，本书描述利用数字数据进行重要决策（如分类）的算法，也就是说，将讨论模式识别的算法。模式识别涉及模式的分类和聚类。在模式分类中，使用一组训练模式或领域知识为模式分配类标签。聚类可以将数据分区，这有助于我们制定决策，我们感兴趣的决策制定是数据分类。例如，基于个人的数据，

脸、指纹以及声音，可以判断他是否是一个通缉犯。在这个过程中，不需要处理数据中所有的细节。

对数据的总结或恰当抽象是很有意义的。适当的抽象化数据对人类和机器是有利的。对人类来说，这有助于理解问题，而对于机器来说，它减少了时间和空间上的计算负担。将例子抽象化是机器学习的一个著名范例。具体来说，机器学习有两种重要方式——通过示例和监督学习以及通过观察和聚类学习。在人工智能中，领域知识可以帮助丰富机器的学习活动。这种情况下，基于规则系统的抽象形式被广泛使用。此外，当数据量很大时，数据挖掘工具很有用。所以，模式识别可以很自然地与机器学习、人工智能和数据挖掘相关联。

1.1 什么是模式识别?

在模式识别中，为模式指定标签。在图 1.1 中，分别有属于类 X 和类 O 的模式。模式 P 是一个新的样本，它需要被分到类 X 或者类 O 中。在一个将人类划分为“高”“中等”“矮”三类的系统中，通过示例学习，系统学习到了把一个指定的人分到这些类中的方法。这里的类标签是具有语义的，它们传达一些意义。在聚类的情况下，将一组未标记模式放在一起。这时，分配给每组的标签是结构式的，或者仅仅表示这个集群的身份。

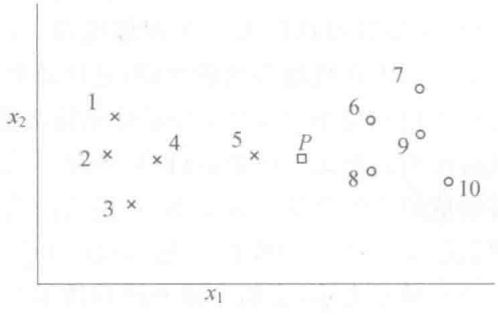

图 1.1 模式的例子

有时，可以使用分类规则而不进行任何抽象，这时相邻/相似性(或距离)的概念可以用来进行模式分类。这种相似性函数的计算基于模式的表示方式。一个模式用一个由特征值组成的向量来表示，用来表示模式的特征非常重要。例如，在表 1.1 中的数据中，人被分为“高”和“矮”两类，此时分类所使用的特征为“体重”。

对于一个体重为 46 kg 的人，那么他/她可能会被分类为“矮”，因为 46

更接近50。然而，这样的分类结果并不合理，因为，一个人的体重和类标签“高”和“矮”并不相关，而使用特征“身高”会更合适。在第2章中将讨论模式和类的表示方式。

表1.1 使用特征“体重”，将人分为高和矮两类

体重/kg	类标签
40	高
50	矮
60	高
70	矮

模式识别的重要方面之一就是其应用前景，在农业、教育、安全、交通、金融、医学和娱乐等领域中有广泛应用，具体包括生物识别技术、生物信息学、多媒体数据分析、文档识别、故障诊断以及专家系统等。分类是人类的一种基本的思维模式，所以模式识别可以应用在任何领域。一些使用频率最高的应用包括字符识别、语音/说话人识别和图像中的对象识别。如果读者对这些应用感兴趣，可以在 *Pattern Recognition* (*www.elsevier.com/locate/pr*) 或 *IEEE Transactions on Pattern Analysis and Machine Intelligence* (*www.computer.org/tpami*) 中的文献中进一步了解。

1.2 模式识别的数据集合

在互联网上有大量的数据集可供使用。一个受欢迎的网站是UC Irvine的机器学习库(*www.ics.uci.edu/MLRepository.html*)，它包含许多不同大小的数据集，可以用于各种分类算法。其中很多甚至给出了一些分类方法的分类精度，可以作为研究的基准。用于数据挖掘的大型数据集可在网站 *kdd.ics.uci.edu* 和 *www.kdnuggets.com/datasets/* 中找到。

1.3 模式识别的理论框架

有多种理论框架能解决模式识别问题，其中最主要的两种为：

(1)统计模式识别。

(2)结构模式识别。

在这两种方式中，统计模式识别使用更为广泛，在文献中大量出现。

其主要原因是在这一领域大多数实际问题需要处理有噪声数据和不确定性，统计和概率是处理此类问题的有效工具。另一方面，形式语言理论为结构模式识别提供了有利条件，然而这样的语言工具不适合处理噪声环境中的数据。这使得本书的重点在于统计分类和聚类。

在统计模式识别中，使用向量空间表示模式和类。数据抽象通常用来处理多维空间中点的概率分布。使用向量空间来表示模式，就需要讨论子空间以及点的相似度。与这一概念相关的有几个计算工具，例如，神经网络、模糊集和粗糙集的模式识别方案都使用矢量表示点和类。

最常用并且最简单的分类器是基于最近邻规则的。这里一个新的模式是根据它最近邻的类标签进行分类的，这类分类中没有训练过程。在第3章中将详细论述最近邻分类。存在不确定性时要注意分类器的理论边界。贝叶斯分类器是针对最小错误率的最优分类器。在第4章中会讨论贝叶斯分类器。隐式马尔可夫模型(HMM)在语音识别等领域被广泛应用，在第5章中将会讨论 HMM。决策树是一个透明的数据结构，可以处理数值与类别特征。在第6章中将讨论决策树分类器。

神经网络模型是用来模拟人类大脑的学习过程的。有一种神经网络感知器，是用于查找高维空间中的线性决策边界的。支持向量机(SVMs)就是基于这一观点建立的。在第7章中，将探讨神经网络以及支持向量机的作用。使用多个分类器来得到一个新模式的类标签也是可行的，这样的组合分类器将在第8章中讨论。

通常，可能有可以直接用于分类的大量的训练数据集合。这时，可以通过聚类来生成抽象数据，并将这些抽象数据用于分类。例如，对应不同类的模式可以被聚类并形成一个子类。每一个这样的子类(集群)可以用一个典型的模式来表示。这些典型模式可以代替整个数据集用来构建分类器。在第9章中，将讨论一些常用的聚类算法。

问题讨论

模式识别是用来处理模式分类和聚类的，其可以应用在许多领域中。模式识别可以是统计型或结构型的，统计模式识别应用更广泛，因为它可以更好地在噪声环境下工作。

延伸阅读材料

Duda 等(2000)撰写了一本非常棒的关于模式识别的书。Tan 等(2007)的 *Introduction to Data Mining* 是一本很好的资料。Russell 和 Norvig (2003)撰写了一本关于人工智能的书,其中将学习和模式识别技术作为人工智能的一个组成部分进行讨论。Bishop (2003) 讨论了在模式识别中神经网络的使用。

习　题

1. 考虑一个识别数字 0 到 9 的任务。使用一组由计算机生成的数据。对于这个问题,可以使用哪些特性?这些特征是有语义的还是结构性的?如果它们全都是结构性的,能否想出另一些特征使得它们中的一部分是结构性的,而另一部分是具有语义的?

2. 举出一个不需要对训练数据进行抽象的分类方法。

3. 指出以下数据中哪些可以直接使用分类规则,哪些需要进行数据抽象?

①基于最近邻的分类器。

②决策树分类器。

③贝叶斯分类器。

④支持向量机。

4. 指出以下各项是否为统计模式识别或结构模式识别?

①模式为一组由特征值组成的向量,使用基于最近邻的分类器进行分类。

②模式本身很复杂,这些模式由简单的子模式组成,而这些子模式本身由更加简单的子模式组成。

③使用支持向量机进行分类。

④模式可以被看作某种语言的一个句子。

本章参考文献

[1] C. M. Bishop. *Neural Networks for Pattern Recognition*. New Delhi: Oxford University Press, 2003.

[2] R. O. Duda, P. E. Hart, D. G. Stork. *Pattern Classification*. JohnWiley and Sons. 2000.

[3] S. Russell, P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson India. 2003.

[4] P. N. Tan, M. Steinbach, V. Kumar. *Introduction to Data Mining*. Pearson India. 2007.

第2章　模式集合的表征

学习目标：

阅读本章之后，你将会：

(1)了解到模式可以表示为

—字符串。

—逻辑类型。

—模糊集和粗糙集。

—树和图。

(2)学会利用近似方法对模式进行分类，诸如：

—距离测量。

—非度量方法，包括：

①中位数距离。

②豪斯多夫(Hausdorff)距离。

③编辑距离。

④互近邻距离。

⑤概念内聚性。

⑥核函数。

(3)了解如何对数据进行抽象。

(4)发现特征提取的意义。

(5)了解特征选择的优点以及特征选择的不同方法。

(6)了解分类器评估所涉及的参数。

(7)理解完成聚类的评估需求。

模式是一个物理对象或抽象概念。如果讨论动物种类，那么对一种动物的描述就是一个模式。如果讨论不同类型的球，那么对一种球的描述(可能包括球的尺寸和材质)就是一个模式。这些模式由一系列描述所表征。根据分类问题的不同，使用不同的模式特征。这些特征被称作属性。模式是根据从属性中获取的数值对对象进行表征。在分类问题中，有一系列属性值已知的对象。有一系列的类型，每个对象属于其中一类。以动物这一模式为例，分类可以是哺乳动物、爬行动物等。在球类这一模式中，分类为足球、板球、乒乓球等。给定一个新模式，就需要去确定模式的分类。