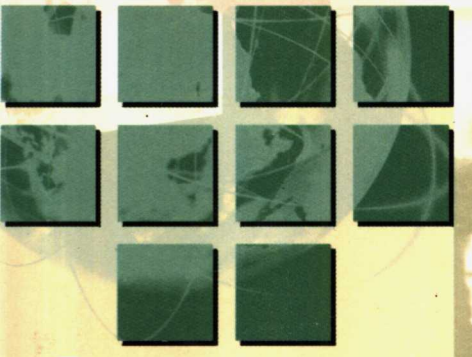
信息与计算科学系列教材

石峰 莫忠息 编著

信息论基础

（第二版）



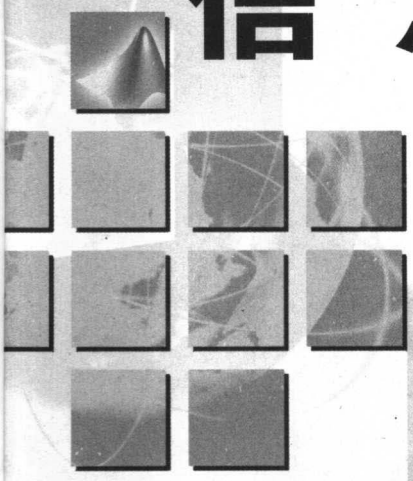
WUHAN UNIVERSITY PRESS
武汉大学出版社

·信息与计算科学系列教材·

石峰 莫忠息 编著

信息论基础

（第二版）



WUHAN UNIVERSITY PRESS
武汉大学出版社

图书在版编目(CIP)数据

信息论基础/石峰,莫忠息编著.—2版.—武汉:武汉大学出版社,2006.4

信息与计算科学系列教材

ISBN 7-307-04951-1

Ⅰ.信… Ⅱ.①石… ②莫… Ⅲ.信息论—高等学校—教材 Ⅳ.G201

中国版本图书馆CIP数据核字(2006)第012101号

责任编辑:顾素萍 解云琳 责任校对:刘 欣 版式设计:支 笛

出版发行:**武汉大学出版社** (430072 武昌 珞珈山)

(电子邮件:wdp4@whu.edu.cn 网址:www.wdp.com.cn)

印刷:湖北省通山县九宫印务有限公司

开本:787×980 1/16 印张:16.75 字数:284千字

版次:2002年7月第1版 2006年4月第2版

2006年4月第2版第1次印刷

ISBN 7-307-04951-1/G·806 定价:21.00元

内 容 提 要

本书从基本概念和基本方法入手，尽量使用初等的数学工具，系统而又全面地介绍信息论的基础理论、基本方法以及近年来发展的新成果，包括一些最新的、较为成熟的原理。全书共分 10 章。第 1 章介绍香农信息理论的基本概念、基本内容和发展简史。第 2 章讨论信源、信息的度量等。第 3 章介绍互信息。第 4 章主要介绍有关信源编码的一些基本概念和 Huffman 最优编码。第 5 章主要介绍离散信道编码定理。第 6 章介绍线性码。第 7 章介绍有关率失真理论。第 8 章介绍最大熵原理和最小鉴别信息原理。在第 9 章中对非统计意义下的信息理论（组合信息与算法信息）作了介绍，同时给出通用编码方法的代表——Lempel-Ziv 编码方法。第 10 章介绍密码学的基础知识。大部分内容附有习题。

本书适用于信息与计算科学、应用数学、概率统计、计算机科学、自动控制、通信工程等专业的本科生作教材，也可作为这些专业的研究生的学习参考，并可供有关工程技术人员自学。

出 版 说 明

1998 年，教育部颁布了经调整后的高等学校新的专业目录，从 1999 年秋季开始，各院校开始按新的专业设置进行招生。信息与计算科学专业是在这次调整中设置的，是以信息处理和科学与工程计算为背景的，由信息科学、计算科学、运筹与控制科学等交叉渗透而形成的一个新的理科专业。目前，社会对这方面的人才需求越来越多，开办这个专业的院校也越来越多。因此，系统地出版一套高质量的相关教材是一项当务之急。

由于信息与计算科学专业是一个新设的专业，有关该专业的人才培养模式、培养目标、教学计划、课程体系、教材建设等一系列专业建设问题，各院校目前正在积极地研究和探索之中。为了配合全国各类高校信息与计算科学专业的教学改革和课程建设，推进高校信息与计算科学专业教材的出版工作，在有关专家的倡议和有关部门的大力支持下，我们于 2002 年组织成立了信息与计算科学系列教材编委会，并制定了教材出版规划。

编委会一致认为，规划教材应该能够反映当前教学改革的需要，要有特色和一定的前瞻性。规划的教材由个人申报或有关专家推荐，经编委会认真评审，最后由出版社审定出版。教材的编写力求体现创新精神和教学改革，并且具有深入浅出、可读性强等特点。这一套系列教材不仅适用于信息与计算科学专业的教学，也可以作为其他有关专业的教材和教学参考书，还可供工程技术人员学习参考。

限于我们的水平和经验，这批教材在编审、出版工作中还可能存在不少的缺点和不足，希望使用本系列教材的教师、同学和其他广大读者提出批评和建议。

信息与计算科学系列教材编委会

2006 年 3 月

第二版序言

本书修订版的主要变动有如下几个方面：

1. 更正了一些错误，对部分内容的介绍更加详细。

2. 增加了意义信息和加权熵、游程编码、Kieffer-Yang 通用编码、级联信道和并联信道的信道容量等内容，使本书的内容与现代技术更加接近。

3. 增加了少量的实际应用例子，使学生能体会信息理论在实际生活中是如何应用的。

4. 修正并增加了少量的习题，并对部分习题给出了习题参考答案或提示。

很多读者认为本书理论性很强，实用编码方法介绍偏少。编者认为：本书的定位主要介绍的是信息理论的基础部分，较多的理论说明是应该的。本书不同于一般的编码理论，不可能过多地讲述编码方法。其次，这次修订我们也增加了部分新的编码方法，如游程编码、Kieffer-Yang 通用编码，并增加了几个实际的例子，以满足这些读者的需要。根据编者近几年的教学经验，第 6 章和第 10 章可以作为选学内容，可以根据学时和学生的实际情况安排讲解内容，毕竟有单独的课程或书籍来介绍它们。

另外，很多读者认为习题比较难做。实际上，每章后面的习题都可以根据该章的内容做出来，主要是做题的方法不容易想到，有一定的灵活性。这需要读者认真的、潜心的思考，不是套公式来做。这次修订也给出了部分习题的解答或提示。对于是否给出习题解答，很多专家认为，给出习题解答，往往会让部分学生懒惰下来，不认真做练习。对此，我们只是给出部分习题的提示。完整的解题过程，还是需要学生认真思考，自己得到。

本书的此次修订，得到信息与计算科学系列教材编委会、使用该教材的广大师生和武汉大学出版社的大力支持，编者在此表示由衷的感谢！

这次修订虽然主观上力求完善，但是难免还有很多不妥之处，希望读者多多指正。

石峰　莫忠息

2006 年 3 月于武汉

前　言

随着科学技术的迅速发展、知识体系的不断更新，信息的概念已在自然科学、人文与社会科学中被广泛地采用，信息理论越来越受到人们的重视。实际上，我们目前所处的时代早已被称为信息时代。香农信息理论，是一个业已成熟的科学体系，是研究信息理论的基础。

长期以来，信息论的教材均是为通信工程或概率统计等专业的学生而编写的，需要读者具有较多的相关专业的知识。为使其他专业的学生能尽早掌握信息理论的基本原理和方法，也使信息论的思想、原理和方法在更为广泛的范围内得到推广和应用，本书尽量用较少的概率统计知识和通信工程知识，尽可能以离散的情形来讨论有关问题。

本教材着重介绍香农信息理论的基本概念、基本分析方法和主要结论，同时，还介绍组合信息和算法信息知识。由于香农（Claude Shannon）提出的信息论是一种基于统计意义上的信息理论，这一理论对通信技术的发展产生了持久而又深刻的影响，但它对信息技术的其他某些方面，如人工智能、机器学习等，则指导作用很少，所以人们对更为广泛意义下的信息的研究一直没有停止。到目前为止，较为成熟的研究成果有：E. T. Jaynes 在 1957 年提出的最大熵原理的理论；S. E. K. Kullback 在 1959 年首次提出后又被 J. S. Shore 等人发展了的鉴别信息和最小鉴别信息原理的理论；A. N. Kolmogorov 在 1965 年提出的关于信息度量定义的三种方法——概率法、组合法和计算法；A. N. Kolmogorov 于 1968 年阐明并被 J. Chaitin 在 1987 年系统发展了的关于算法信息的理论。这些成果大大丰富了信息理论的概念、方法和应用范围。首先，它把信息的统计定义进一步推广并对非统计意义下的信息给出了一种度量。其次，信息度量的意义已不再局限于信源编码和信道编码，而是已经系统地发展成为信息处理的一种准则，如最大熵原理和最小鉴别信息原理在目前的估计理论中占据着重要的地位。基于此，本书对这些基本概念和成果也作了较为详细的介绍。

传统的信息论包含三个方面：信息论基础、编码理论、密码理论。由

于信息论的基本原理和方法与编码理论息息相关，我们在介绍信息理论的基本知识时，处处涉及编码理论，虽然在第5章的最后提到了重复码和Hamming码，但那仅仅只是为了加深读者对有关编码的感性认识，在第6章我们专门对较为成熟的线性码作了简单而又系统的介绍。由于现代密码学是基于信息理论的，所以我们对密码学作了一点介绍。

全书共分10章。第1章主要介绍香农信息理论的基本概念、基本内容、发展简史以及通信系统模型。第2章讨论信源的特点和分类，信息的度量，以及信息熵、条件熵、联合熵等。第3章介绍互信息。第4章主要介绍有关信源编码的一些基本概念和Huffman最优编码。第5章主要介绍离散信道编码问题、信道编码定理，并对Hamming码作了简单的介绍。第6章系统地从理论上对线性码作了介绍。第7章介绍有关率失真理论。第8章介绍两个应用得非常广泛的最大熵原理和最小鉴别信息原理，它们是估计理论中两个重要的原理，也是两个重要的方法。非统计意义下的信息理论（组合信息与算法信息）在第9章中作了介绍，同时给出通用编码方法的代表——Lewpel-Ziv编码方法。第10章介绍密码学的基础知识。

本书前5章和第7章是信息理论的基本内容，力求系统、全面地介绍有关信息的基本概念，如信源和信宿、信道和编码的基本概念和理论。作者认为本书第2章所讲述的一些基本概念和问题是最基本的，特别是熵的概念，初学者应该细心研读，准确理解。这一章内容掌握得好，其余部分的内容也就容易理解和掌握了。学习这部分内容的数学基础要求是学习过数学分析和概率论的基本内容。当然，最好对随机过程的一般理论和分析方法也有所了解。

第6章、第8～第10章的内容是专题研讨部分，读者可根据需要选读有关内容。作者力求反映国内外在这些专题研究的主要成果。这部分内容主要是为了扩大读者的知识面。教师在讲授信息论理论基础课程时可只作简略介绍或只选讲其中部分内容。

全书第1～第5章、第7～第9章由石峰编写，第6章和第10章由莫忠息编写。虽然在编写过程中参考了大量的教科书和有关专著，但由于编者的学识有限，书中错误在所难免，敬请读者批评和指正，编者不胜感激。

编　者

2006年3月

目　录

第1章 概　论

当今，信息已经成为现代社会的一项重要资源．信息的要领在自然科学和社会科学中均已被广泛地采用．研究信息的产生、获取、检测、传输、处理、识别及其应用的信息科学技术，在近几十年里得到了迅速发展．目前社会上流行的一些提法，如“信息、材料、能源是现代科学的三大支柱”、“信息、物质、能量是构成一切系统的三大要素”等，充分说明了人们对信息的重要性的认识．

1.1　信息理论的基本内容

信息科学是研究信息的产生、获取、度量、变换、传输、处理、识别及其应用的一门科学，也是源于通信实践发展起来的一门新兴应用科学．它所研究的问题是带有根本性的、基础性的问题，所给出的方法也是具有普遍性的、令人信服的、可以解决实际问题的方法，所得结论是严谨的、经得起考验的结论．

信息理论的基本问题是信源和信宿、信道以及编码问题．

信息的获取或产生主要依赖于信息源，简称**信源**．信源大致可分为三大类：一是自然信源，包括来自于物理、化学、天体、地学、生物等方面的自然的信息．获取信息的主要工具是传感器和传感设备，其种类繁多，形式不一，不胜枚举，主要有物理型（热、光、磁、电、声、力）传感器、化学型（气体、化合等）传感器和生物型（神经、感觉、嗅觉、视觉、听觉、触觉等）传感器．二是社会信源，包括政治、军事、管理、金融、商情以及各种情报等．采集信息主要靠社会调查，利用统计方法加以整理．三是知识源．古今中外记录下来的知识和专家的经验都蕴含大量的

信息.

在信息理论中，信息和消息是紧密相关的两个不同的概念. 一般认为，消息是信息的载体，如语言、文字、各种符号、声音、图片等，而信息蕴含在消息之中. 同一个消息，比如说当天新闻联播的一篇报道，不同的人从中获取的信息是不一样的；一封家书，对于收信人而言可抵万金，但对旁人来说可能是废纸一张. 因此信息是一个奇妙的东西，它是有别于物质和能量的一种存在. 信息可由一个人掌握，也可由多人所知晓. 信息的本质和它的科学定义是当前科学界，乃至哲学界热衷研究的课题，但是它的重要性是毋庸置疑的.

信息的核心问题是它的度量问题. 从目前的研究来看，要对通常意义下的信息给出一个统一的度量是困难的. 至今最为成功的，也是最为普及的信息度量是由信息论创始人 Shannon 在他的光辉著作《通信的数学理论》中提出的、建立在概率统计模型上的信息度量. 他把信息定义为“用来消除不确定性的东西”. 用概率的某种函数来描述不确定性是自然的，所以，Shannon 用

$$I(A) = -\log P(A)$$

来度量事件 A 发生所提供的信息量，称之为事件 A 的自信息，其中 $P(A)$ 为事件 A 发生的概率. 这个定义与人们的直觉经验相吻合. 如果一个随机实验有 N 个可能的结果或一个随机消息有 N 个可能值，若它们出现的概率分别为 $p_1, p_2, \cdots, p_N$，则这些事件的自信息的平均值

$$H = -\sum_{i=1}^{N} p_i \log p_i$$

作为这个随机实验或随机消息所提供的平均信息也是合理的. H 也称为**熵**，是借用统计物理中的一个名词. 因为在物理学中，熵是描述系统的不规则性或不确定性的一个物理量.

由于信息论是源于通信实践发展起来的一门新兴应用科学，故通信系统的基本模型也是信息理论的基本模型，该模型如图 1.1 所示.

信源是产生消息(或消息序列)的源. 消息通常是符号序列或时间函数. 消息取值服从一定的统计规律，所以信源的数学模型可以是离散的随机序列或连续的随机过程.

所谓编码，就是用符号来表达信息，即进行信源编码；还需要将符号转换成信道所要求的信号，即进行信道编码. 译码是编码的反变换.

信源编码器把信源产生的消息变换成数字序列. 在不允许编码失真的情况下，信源编码器的目的是在保证能从其输出数字序列并能准确无误地

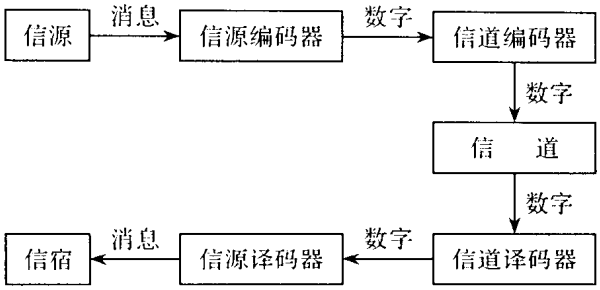


图 1.1 基本的通信系统模型

恢复输入消息序列的前提下，提高输出数字序列的效率，也就是保证在不失真的条件下，对输入消息序列进行压缩. 或者，在允许失真的情况下，信源编码器的目的是对给定的信源，在保证消息平均失真不超过某个给定允许值 D 的条件下，对输入消息序列进行压缩.

信道在实际通信系统中是指传输信号的媒介或通道，如电缆、光缆、电离层、人造卫星等. 在信息论的模型中也把发送端和接收端的调制、解调器等归入到信道中，并把系统中各部分的噪声和干扰都归入到信道中. 在信道的输入、输出模型中，根据噪声和干扰的统计特征，用输入、输出的条件概率（或称转移概率）来描述信道特征.

信道编码器把信源编码输出的数字序列变换成适合于信道传输的由信道入口符号组成的序列. 其主要作用是对其输出序列提供保护，以抵抗信道噪声和干扰.

信道译码器和信源译码器分别是信道编码和信源编码的反变换，信宿是消息的接收者.

总之，编码是把信息变换成信号的方法、措施，而信道则是传送、存储信号的具体的物理设施，译码是编码的反变换. 这些问题的讨论是结合通信系统模型的研究进行的. 从信源得来的信息，经过编码后进入信道. 信道是系统的关键部分，它将信源的输出输入系统，然后再将输出信号经加工后送给用户，前者称为**编码**或**调制**，后者称为**解码**（**译码**）或**解调**. 正是在研究通信系统的基础上，1948 年，Shannon 建立了信息理论的基础. 他利用随机编码的方法证明，在含噪的信道中，利用适当的编码器和适当的解码器就可以得到近乎无误的通信. 然而这一非构造性的证明并没有告诉我们如何设计这样的编码器和解码器，也没有告诉我们它们是如何复杂. 从那以后，科学家们做了许多解决这些实际问题的工作，但迄今为

止还是没有完全解决，人们仍在继续研究解决这些问题的答案. 因而，信息理论仍处在不断的发展之中，并且起着越来越大的指导作用.

本书讲述的信息理论，其大部分内容与通信科学的信息理论的基本内容大体一致，原因在于我们是以通信系统的基本模型来描述信息科学中的基本理论的.

1.2 信息理论的发展简史

信息理论基础的建立，一般来说，开始于 Shannon 研究通信系统时所发表的论文. 随着研究的深入与发展，信息论具有了较为宽广的内容.

从历史上看，信息论的形成是两部分人共同努力的结果，一部分是通信工程方面的学者，另一部分是统计数学家. 这两部分人虽然研究的是同一个领域的问题，但是他们感兴趣的方面和侧重点不一样. 这种情况从信息论的产生开始一直保持到现在.

信息理论与通信理论是分不开的，通信理论的发展正是信息理论发展的基础. 通信中最重要的问题就是信息量问题和传送信息的速度问题. 1267 年 Roger Bacon（罗杰·培根）提出了利用所谓“共振针”进行远距离的通信. 16 世纪 Gilbert Porta（吉尔伯特）提出了共鸣电报. 1746 年，英国 Watson（沃森）在 2 英里的电线上，传送了电信号. 1876 年 Graham Bell（贝尔）发明了电话. 1925～1927 年，引入了电视，出现宽频带问题及远距离传送图像时的相位问题和噪声问题. 噪声问题一直是一个大问题，研究这方面问题的先驱者有：Einstein（爱因斯坦，1905 年），Schottky（肖特克，1918 年），Johnson（约翰逊，1928 年）和 Nyquist（奈奎斯特，1928 年）. 1922 年，John Carson（约翰·卡森）研究了调频信号的频带，提出“调频信号”非窄带而是宽带，他认为“许多问题中往往包含有基本的谬误”，但到 1936 年，Armstrong（阿姆斯特朗）公布了他的调频试验的结果，指出：“在一组载波中可将最强的一个波分离出来，许多很靠近的载波不致互相干扰.”这样就改正了卡森的错误结论，于是调频技术得到发展.

1924 年，美国的奈奎斯特和德国的 Küpfmüller（屈普夫米勒）同时提出这样的定理：在速率一定的情况下传输电报信号需要一定的频带. 证明了信号传输速率与信道带宽成正比. 经过 4 年后，Hartley（哈特利）将它写成公式. 设由 N 个符号组成一条消息，这 N 个符号又是从 S 个符号中选取出来的，因此可有 S^N 条可能的消息. Hartley 定义信息量为