普通高等教育“十五”国家级规划教材

信息论与编码

仇佩亮　编著

高等教育出版社

普通高等教育“十五”国家级规划教材

信息论与编码

仇佩亮　编著

高等教育出版社

内容提要

本书是“十五”国家级规划教材。信息论和编码是研究信息传输和信息处理过程中一般规律和具体实现的一门应用科学，是现代信息科学和技术工程的基础理论。本书是在吸取了国内外经典教材的优点，结合作者教学经验的基础上编写而成。本书写得深入浅出，既保持理论的完整性、系统性，又概念清楚、易读好懂，同时介绍了信息论的新发展。教材主要介绍Shannon信息理论和相关的编码技术。内容包括如下11章：绪论、熵和互信息、离散无记忆信源的无损编码、信道、信道容量及信道编码定理、率失真理论和保真度准则下的信源编码、受限系统和受限系统编码、线性分组纠错编码、循环码、卷积码、Turbo码与迭代译码、多用户信息论与多用户编码。

本书适合作为高等院校电子信息类专业的高年级本科生和研究生教材，对于从事信息科学和技术领域工作和研究的人员也极具参考价值。

图书在版编目（CIP）数据

信息论与编码 / 仇佩亮编著. —北京：高等教育出版社，2003.12（2004重印）

ISBN 7-04-013047-5

Ⅰ.信… Ⅱ.仇… Ⅲ.①信息论-高等学校-教材②编码理论-高等学校-教材 Ⅳ.O157.4

中国版本图书馆CIP数据核字(2003)第095841号

出版发行	高等教育出版社	购书热线	010-64054588
社　　址	北京市西城区德外大街4号	免费咨询	800-810-0598
邮政编码	100011	网　　址	http://www.hep.edu.cn
总　　机	010-82028899		http://www.hep.com.cn
经　　销	新华书店北京发行所		
印　　刷	高等教育出版社印刷厂		
开　　本	787×960　1/16	版　　次	2003年12月第1版
印　　张	32.5	印　　次	2004年7月第2次印刷
字　　数	610 000	定　　价	40.10元

本书如有缺页、倒页、脱页等质量问题，请到所购图书销售部门联系调换。

目　录

第1章 绪论

信息论是应用近代概率统计方法来研究信息传输、交换、存储和处理的一门学科，也是源于通信实践发展起来的一门新兴应用科学。

信息是系统传输、交换、存储和处理的对象，信息载荷在语言、文字、数据、图像等消息之中。在信息论中，信息和消息是紧密相联的两个不同概念。同样一个消息，比如一张当日的报纸，对于不同的人从中可获得的信息是不一样的；同样的天气预报“明天有雨”，对于干旱地区和雨量充沛地区来说其信息含量也不一样。一张纸写上几个字成为一封家信，对于收信者是家书抵万金，但对旁人可能是废纸一张。因此信息是一种奇妙的东西，它是有别于物质和能量的一种存在。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。信息的重要性是毋庸置疑的。控制论创始人维纳说过：“要有效地生活，就要有足够多的信息”。目前社会上流行一些提法，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”……这些提法充分说明了人们对信息重要性的认识。

信息的度量是信息论研究的基本问题。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。存在许多种关于信息度量的定义，但至今最为成功，也是最为普及的信息度量是由信息论创始人香农(Shannon)在他的光辉著作《通信的数学理论》[17]中提出的，是建立在概率模型上的信息度量。他把信息定义为“用来消除不确定性的东西”。既然信息与不确定性相联系，因此用概率的某种函数来描述不确定性是自然的，所以香农用

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

来度量事件 A 的发生所提供的信息，其中 $P(A)$ 为事件 A 的概率。这个定义与人们的直觉经验相吻合。如果一个随机试验有 N 个可能结果或者说一个随机消息有 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

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

作为这个随机试验或随机消息所提供的平均信息也是合理的。H 也称为熵，这是借助于统计物理学中的一个名词。事实上熵作为信息的代名词也是由 20 世纪伟大的数学家、物理学家冯·诺依曼向香农建议的。在物理学中熵是描述系统的不规则性或不确定性程度的一个物理量。

信息论所研究的通信系统基本模型如图 1.1.1 所示。

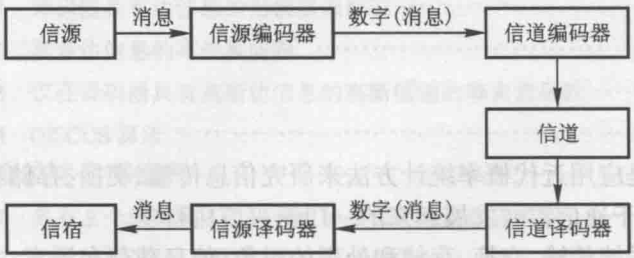

图 1.1.1 通信系统的基本模型

信源是产生消息(或消息序列)的源。消息通常由符号序列或时间函数组成。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一个离散的随机序列或连续的随机过程。

信源编码器把信源产生的消息变换成数字序列。对无损信源编码来说，信源编码器的目的是在保证能从其输出数字序列中无错误地恢复出输入消息序列的前提下，减少输出数字序列的速率，也就是保证在不失真的条件下对输入消息序列进行压缩。在允许失真的情况下，信源编码的目的是对给定信源，在保证消息平均失真不超过某给定允许值 D 的条件下，尽量减少输出数字序列的速率。

信道在实际通信系统中是指传输信号的媒介或通道，如架空明线、电缆、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、解调器等归入信道，并把系统中各部分的噪声和干扰都归入信道中。在信道的输入、输出模型中，根据噪声和干扰的统计特性，用输入、输出的条件概率(或称转移概率)来描述信道特性。

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。信道编码器的最主要作用是要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，即消息的归宿。

信息论解决了通信中的两个基本问题。首先对于信源编码，信息论回答了“达到不失真信源压缩编码的极限(最低)编码速率是多少?”这一问题。香农的答复是这个极限速率等于该信源的熵 H。事实上香农认为每个随机过程，不管是音乐、语言、图像，都有一个固有的复杂性，该随机过程不能被无失真地压缩到

该固有复杂性之下，这个固有复杂性就等于该随机过程的熵。信息论对通信解决的第二个问题是关于信道编码方面的。在香农以前，人们都认为增加信道的信息传输速率总要引起错误概率的增加，认为要使错误概率为零，则传输速率只能为零。但香农却出人意料地证明，只要信息传输速率小于信道容量 C，传输的错误概率可以任意地小，反过来如果超过信道容量，则传输的错误是不可避免的。对每个信道可以根据它的噪声干扰特征计算出它的容量 C。

香农信息论与信息编码技术是两个密不可分的学科领域，或者说它们是信息科学的两个不同方面。香农信息论指出了通信中信源编码和信道编码的极限速率。香农利用随机编码方法，证明了当码长趋于无限时，存在一种编码方式，能够达到这个理论上的极限速率。香农所使用的证明方法在理论上极为漂亮，但实际上无法实现。香农的证明方法是一种“存在性”证明方法。这种方法在计算上是不可实现的。对于实际的通信专家和编码专家来说必须去寻找有效的、可实现的编码方法。借助于电子科学技术的发展，无论对于信源编码，还是对于信道编码，目前都有许多具有实用价值的编、译码方案，它们的性能正逐步向香农指出的极限逼近。

编码理论工作者和通信工程师所追求的目标不仅仅是要寻找达到香农理论极限的编码方法，更重要的是要寻找可以实现的编码方法，因此，编码的复杂性是放在首位考虑的因素。在无损压缩编码中，早期的 Huffman 编码被认为是最优的变长度压缩编码方法，但是它的复杂性随着码长的增大急剧增加，所以对于大的码长来说 Huffman 编码是不实际的。20 世纪 70 年代开始的算术编码，虽然按平均码长来说不是最佳的，但它是一个在线的算法，计算复杂性随码长线性增加，因此，算术码是一种实用的码。有人认为算术码的提出标志着无损压缩编码的一个突破。

众所周知，自然界的信号都是连续的，无论是语音信号、图像信号或各种传感信号，不可能用有限比特不失真地表示它们。因此问题在于如何设计一种编码方法，使其在给定的许可失真范围内，用最少的比特表示它们，或者说如何用给定的比特数来表示这个连续信号，使失真最小。这就是保真度意义下的压缩编码。几十年来，在这方面已发展了许多成功的实用压缩编码方法，比如矢量量化、预测编码、变换编码、子带编码等技术，其中许多技术已成为国际标准，例如 CCITT 中关于语音压缩和图像压缩的标准。正是由于这些有损压缩编码技术的应用使得语音、图像信号的码率可以成十倍甚至上百倍地降低，同时使由压缩编码引起的信号质量下降不为人类感官所觉察。这些编码技术是当前各种多媒体技术的核心。

信道编码也就是通常所说的纠错编码，是另一大类信息编码技术。这类编码的目的在于检测或纠正传输中的错误，提高信息在传输中的可靠性。纠错编

码中最早的 Hamming 码[55]是几乎与香农信息论同时被提出来的。早期纠错码研究集中在线性分组码,采用的数学工具是矩阵理论。到 20 世纪 60 年代,由于以有限域理论为主的抽象代数工具的引入使线性分组码的研究突飞猛进。循环码,特别是 BCH 码、RS 码等的研究,不仅为线性编码的研究打下坚实的基础,而且由于代数构造的引入使得译码复杂性大为下降。20 世纪 70 年代以后基于概率译码的序贯编码理论,特别是卷积码,获得了极大的发展。20 世纪 70 年代,纠错编码技术首先在宇宙飞船、深空通信中获得了成功应用,这极大地鼓舞了纠错编码研究者。今天由于微电子技术的发展,使以前难以实现的复杂译码算法在超大规模芯片中得到实现,从而使得纠错编码成为通信系统中不可缺少的一部分。以前由数学家们研究的技术,如 RS 码、Viterbi 算法等,已成为通信工程师的口头禅。纠错编码的成功刺激研究者寻找性能更优越的码,例如代数几何码、Turbo 码、低密度校验码(LDPC)等。这些码的性能已非常接近香农的极限。同时调制技术与纠错编码的结合,信源编码与信道编码的结合会产生一些性能更好的传输技术。相信随着科学的发展和需求的增长,新的、更好的码会不断涌现。

多个信源利用多个发信机和多个收信机在通信网络上进行信息传输会产生许多新的问题,如相互干扰、相互协作、相互叠加、反馈等。在多用户工作条件下的通信极限问题是单用户信息论的推广,称为多用户信息论或网络信息论。多用户信息论与单用户情况一样,主要研究两类问题,即多用户信源压缩编码和多用户信息在网络信道的传输。多用户信息编码中发展起来的许多思想,如叠加编码、嵌入编码、逐次抵消、时域灌水、脏纸上写字、信息细化和边信息应用等,已经在理论和实际上得到了应用。有理由认为新一代的无线网络通信必须从多用户信息论中吸取思想精华,才可望获得新的突破。

香农信息论源于通信实践。它对通信领域的成功应用使得香农理论被称为通信的数学理论。但香农理论的思想、方法,甚至某些结论已渗透到许多其他学科中。

• 统计数学　香农理论本身就是一种数学理论,它与随机过程中 Ergodic(各态历经)理论有密切关系。香农编码定理的基本核心——渐近等分原理(AEP),实际上就是某种形式的大数定律。因此利用熵、互信息等概念来研究 Ergodic 系统是非常有效的。另外,用相对熵作为随机分布之间的距离,在假设检验中、在大偏离理论中均有很好的应用。利用相对熵可以有效估计差错概率指数。

• 计算机科学　计算机和通信是密不可分的,计算能力受制于计算部件之间的通信能力,同时通信能力又受制于计算能力,所以计算和通信是一对双螺旋,信息论的每一步发展直接影响计算科学的发展。且不说各种信源编码、信道

编码、存储编码技术的发展如何直接推动计算机技术的发展，即使计算机中“最佳随机数发生”这么一个简单问题也被证明与最佳信源编码等价。在计算科学中数据串的 Kolmogorov 复杂性被定义为利用通用计算机打印出这个数据串并停机所需的最短二元程序的长度。可以证明一个随机源所输出数据序列的 Kolmogorov 复杂性等于该随机源的香农熵，从而 Kolmogorov 复杂性理论与香农信息论建立了联系。

• 哲学和科学方法论　最大熵准则或最大信息原则是许多科学研究中常用的准则，实践证明这个准则是有效的、合理的。信息论赋予最大熵准则以明确的内涵。最大熵准则和最小描述长度准则都是一种科学的方法论，在信息论中可找到它们的联系。这给予相信“最简单的解释是最好的”信条的人们一个科学的佐证。

另外，信息论的思想和方法还在经济、生物等方面获得应用，已产生了“信息经济学”、“信息生物学”等边缘学科。因此，人们深信信息论的学习有助于对其他学科的研究，同时其他相关学科的研究也会促进信息论的发展。比如量子力学理论与经典信息论的结合已产生了目前发展迅速、前途不可限量的量子信息论、量子编码理论和量子计算理论等。完全可以相信这些理论是属于 21 世纪的工程科学理论，它们将对 21 世纪新科技产生巨大的作用。

第2章 熵和互信息

在本章，介绍信息论中的两个最重要的概念，即信源的熵和互信息，同时介绍它们的一些性质。信源的熵是用来刻画信源发出的消息(随机变量)的平均不确定性，而两个随机变量之间的互信息则表示一个随机变量对另一个随机变量所提供的信息量。因为对于2个随机变量来说，知道了其中一个的取值往往可以减少另一个随机变量的不确定性，这种不确定性的减少被认为是一个随机变量对另一个随机变量提供了互信息。

在本章中首先讨论事件的自信息和事件之间的互信息，然后介绍离散随机变量的平均自信息，即熵和平均互信息；进而把熵和互信息概念推广到连续随机变量和随机过程的情况。

2.1 随机变量的熵和互信息

通常的通信传输系统或信息处理系统可以用图2.1.1来表示。其中方框黑盒子中所谓的信息处理系统可以是某种确定性的处理，如线性滤波、放大等，也可以是某种不确定的处理，例如叠加上某种随机噪声、干扰，或者是经历某种随机衰落、失真等。一般用2个随机变量 X 和 Y 来表示它们的输入和输出。

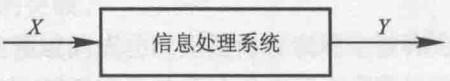

图2.1.1 通信传输系统或信息处理系统

对于随机变量 X，可以用3个量组成的三元组 $\{X,\mathcal{X},q(x)\}$ 来描述它，其中 X 是随机变量的名字(本书中一般用英文大写表示随机变量，用小写字母表示随机变量的取值)，$\mathcal{X}$ 表示随机变量 X 的取值范围，$q(x)$ 表示 X 的概率分布。这样的三元组也称为概率空间。

如果 X 和 Y 是离散随机变量，它们的取值范围为

$$\mathcal{X}=\{x_k;\quad k=1,2,\cdots,K\}$$

$$\mathcal{Y}=\{y_j;\quad j=1,2,\cdots,J\}$$

对每个 $x_k\in\mathcal{X}$，X 取值为 x_k 的概率为