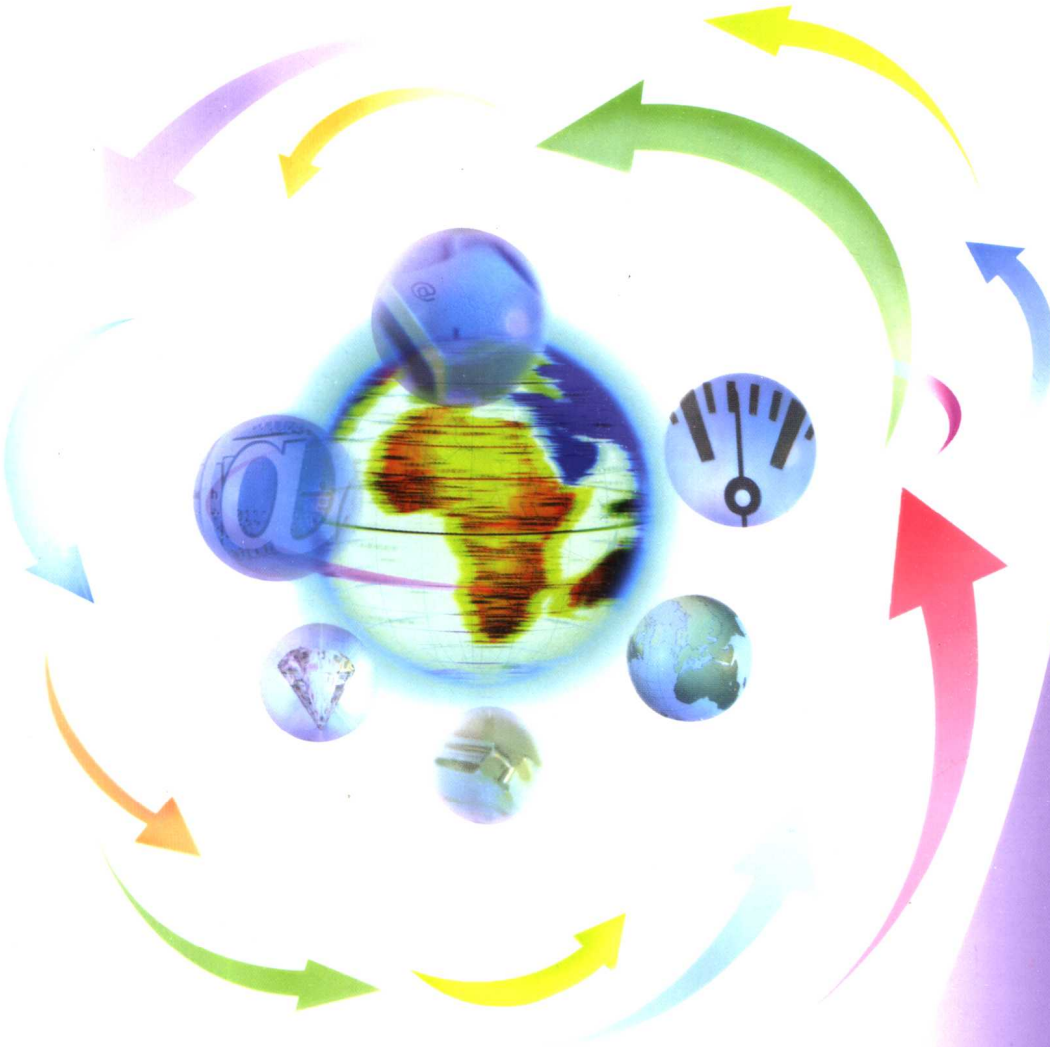
XINXI YU BIANMA JIANMING JIAOCHENG

信息与编码简明教程

主编 于 工



国防工业出版社
National Defense Industry Press

青岛科技大学教材建设基金重点资助项目

信息与编码简明教程

主编　于　工
参编　周祖荣　盖　琦
　　王泽华　闻卫军

国防工业出版社
·北京·

内 容 简 介

本书介绍了香农信息论的基本理论，重点讨论了编码原理与编码方法，并对无失真信源编码、信道编码、限失真信源编码和保密编码进行了讨论，对编码的应用以及近年来的新发展也作了简单的介绍。

本书内容全面、概念清晰、语言简练、通俗易懂、注重概念和应用、尽量减少繁杂公式的推导证明、图文并茂、讲练结合、每章均有习题与参考答案。本书适合用做信息工程、通信工程及相关专业本科生教材；也可供信息、通信、计算机、电子等有关专业的科技人员参考；同时，本书也可作为其他编码爱好者自学用书。

图书在版编目(CIP)数据

信息与编码简明教程 / 于工主编. —北京：国防工业出版社，2005.8

ISBN 7-118-04021-5

Ⅰ.信… Ⅱ.于… Ⅲ.①信息论—教材②信源编码—编码理论—教材③信道编码—编码理论—教材

Ⅳ.TN911.2

中国版本图书馆 CIP 数据核字(2005)第 074861 号

国防工业出版社出版发行

(北京市海淀区紫竹院南路 23 号)

(邮政编码 100044)

国防工业出版社印刷厂印刷

新华书店经售

*

开本 787×1092 1/16 印张 13½ 310 千字

2005 年 8 月第 1 版 2005 年 8 月北京第 1 次印刷

印数：1—4000 册 定价：19.00 元

(本书如有印装错误，我社负责调换)

国防书店：(010)68428422 发行邮购：(010)68414474

发行传真：(010)68411535 发行业务：(010)68472764

前　言

21世纪是一个高度信息化的时代。信息与编码在电子信息工程、通信工程、计算机通信网、图像工程及数字音像等领域得到了广泛应用。随着高等教育的大力发展，几乎所有的理工类大学都开办了信息类及其相关专业，开设了有关信息理论及编码方面的课程。广大学生和教师迫切需要一本简明教程，以满足教学的需要。

信息与编码涉及的知识面较宽，既需要概率论、数论、群论等高深的数学理论为基础，又涉及通信、计算机、信号处理以及电子、电路等方面的专业知识，历来被当作研究生课程。就编码而言，包括数据压缩编码，检错、纠错编码，保密与认证编码等领域，方法众多、应用广泛。然而，作为本科生教学，计划课时一般在48学时左右，为了在较少的学时完成教学大纲规定的内容，使学生们真正学到信息论的精髓，掌握编码的基本原理和方法，应当处理好两个关系：一是深度与广度的关系；二是信息理论与编码实践的关系。

目前，国内已出版了一些专著和教材，作者从各自的领域对相应方面作了精深的论述。一部分适合于教师、科研人员及研究生的学习，另一部分适合于本科教学。根据作者多年教授此课的体会，本书在内容选材方面，注意到了广泛性与典型性相结合，教材内容覆盖面比较宽，涵盖3大类编码。由于精选典型方法介绍，达到了举一反三的功效。在理论深度方面，对数学的系统性和严密性并不苛求，能懂会用即可，强调概念，侧重方法与应用，使深奥的理论变得通俗易懂。在讲述方式上，论述概念力求准确、分析思路力求清晰、语言文字力求简练，图文并用、讲练结合，达到教师好教、学生好用的目的。

全书共分5章。第1章是信息论基础，重点介绍了香农信息理论，为后续章节奠定基础。以下4章分别对无失真信源编码、信道编码、限失真信源编码和保密编码进行分题讨论。作为必讲内容，书中对基本理论和基本编码方法进行了较详尽的论述。为了扩大知识面，书中还介绍了其他的编码方法，特别是20世纪90年代以来科技发展的新成果，可根据需要选讲或作为学生课外阅读材料。另外，配合教学的例题较多，每章后均有习题，附录给出参考答案，以便课后练习。

鉴于本人才疏学浅，错误之处在所难免，望广大读者批评指正。

于　工

2005年5月8日于青岛

目　录

绪　论

长期以来，人类进行信息交互的基本方式不外乎语言、文字和图像。19 世纪中叶，电报的发明开始了用电子技术传输文字的时代，随后出现的电话标志着现代语音通信的诞生。20 世纪上半叶，电视的发展开创了图像传输的先河。直到 20 世纪末，有线电话网和无线电话网、有线电视网和无线电视网仍然是语言和图像的主要传输途径，因特网则是目前除报刊书籍外最主要的文字传输媒体。随着数字技术的发展，3 大信息网络在数字通信的平台上融为一体的趋势日益加速。集语言、文字和图像为一体，世界范围的信息实时交互已不是梦想。一方面是通信技术一日千里地发展，技术更新的周期越来越短；另一方是人们对信息数量和质量的需求不断增长，如何更加有效、更加可靠、更加安全地传输信息，成了人们非常关注的问题。

数字通信过程中，编码是必不可少的。电报出现的同时，文字和符号的编码问题也被提上了日程，随着现代通信技术的发展，编码的作用越来越重要。为了用更短的代码表达同样多的信息，人们提出了压缩代码长度的方法，并发明了多种压缩方法和实施方案，统称为信源编码。为了及时发现并纠正信息传输中出现的错误，人们采用了各种检错和纠错技术，由此发展起来了信道编码技术，它使通信更加可靠。为了信息传输的安全，人们采用了多种保密编码和认证措施，传统的密码学在网络时代焕发出新的生命。信源编码、信道编码和保密编码这 3 大类编码是实现高效、可靠、安全通信的重要保障，是现代通信技术与信息技术不可缺少的组成部分，也是本书的主要内容。

编码，说到底是一种数学映射，通过设定的数学关系，用另一套代码来替换原来的代码。信源编码则是能够使代码变短的替换。信源编码又分为无失真信源编码和限失真信源编码两大类。离散的数字信源所发出的信息，如计算数据、重要资料、文献档案等，要求在压缩代码长度后，能够无失真地恢复原貌，这就需要进行无失真信源编码。而另一些场合下，如语音、图像、影视等，只要能满足人们的视听需求，可以容许存在一定量的失真，从而获取更大幅度的削减和压缩，这叫做限失真信源编码。它们在计算机通信、语音通信和多媒体通信中都得到了广泛应用。

信道编码也是用另一套代码来替换原来欲传输代码的数学变换，或者在欲传输的码流中有计划地插入一些新代码构成新的码流。由于这种替换或插入是按某种预定的数学规律进行的，从而能够在接收端根据设定的规律判断出传输是否有误，甚至找回正确的代码而进行自动纠错。采用信道编码能使通信的误码率大大下降，因此，它几乎在所有的数字通信系统中得到广泛应用，在磁盘、光盘以及计算机网络中更是离不开它。

传统意义上的保密编码是通过数学映射使信息代码变得让别人不可识别，从而起到保密作用。随着网络的普及和信息化进程的加速，信息安全问题变得越来越重要，信息交换的认证问题变得比保密问题更加突出。以公开密钥体系为代表的现代保密机制应运而生，责无旁贷地承担起双重保密任务。它在计算机网络、无线通信、电子商务、军事领域广

泛应用,在日常生活中,各种密码和认证卡的使用早已非常普遍。

在总结通信技术和编码理论百年成果的基础上,1948 年,香农(Shannon)发表了两篇关于通信之数学理论的重要文章,给信息下了科学的定义,用数学理论证明了有噪信道中通过信道编码实现无失真通信的可能性与必要条件,给出了无失真信源编码的极限码长,创立了通信的信息理论。1958 年,香农又发表了关于限失真信源编码和关于保密编码的信息理论。沿着这条道路,半个世纪以来,通信技术的发展与进步日新月异,编码的改进和创新更是层出不穷,香农创立的信息理论得到人们空前的重视。

香农理论是高度概括性的理论,它揭示了信息传输的普遍原理和基本原则,给出了编码的存在定理和理论极限,对于通信系统的设计和编码方案的构建具有指导意义。有了信息理论,编码就不会盲目,工作就会心中有数,事半功倍。学习编码最好以信息论为基础,才能纲举目张、条理清楚、目标明确、原理清晰。

然而,信息论不能代替编码,正如通信原理不能代替通信电路一样。实际问题千差万别,仅有原则性的理论指导是远远不够的。编码是为解决各种实际问题而设计的算法和搭建的系统。每个编码都是具体的,必须具体问题具体分析。为了实现高效、可靠、安全的通信,编码工作者在长期的钻研和实践中,利用所能想到的各种原理,找寻种种异乎寻常的方法,摸索到了许多成功的经验,才有了今天种类繁多、方法各异、特色明显、应用广泛的编码大家族。我们学习前人的成果和他们创造的编码知识,更应当从中体会他们宝贵的钻研精神和创新思维。

本书首先简单扼要地介绍香农信息熵的基本理论,然后分别对信源编码、信道编码和保密编码基本原理和方法进行讨论。对于具有代表性的编码进行了详细的分析和讨论,要求学懂原理、学会使用;对于近年来编码的新进展,则从原理方面进行简要介绍;对于必须涉及到的艰深的数学理论,只是从应用角度做些必要的解释,不进行推导和证明。

第1章　信息论基础

信息论是指导通信和编码的基本理论，它的体系框架是建立在关于信息和信息熵的定义之上的。本章从通信的基本过程中提炼出信息的定义，以信息熵为线索，由浅入深地讨论了信息熵在无记忆离散信源、马尔科夫离散信源、离散信道的表现形式和性质特点，引出了信息传输率和信道容量的概念。最后又把这些概念推广到连续信源和波形信道中，得出了著名的香农公式。这些概念是非常重要的，是以后各章的理论基础，必须深刻理解。

1.1　通信与信息

信息(Information)作为一个专业名词，其科学定义来自通信。

通信系统由信源、信道和信宿组成。发出信息的一方为信源，接收信息的一方为信宿，传输信息的通道为信道，信息则是通信的核心。那么什么是信息呢？信源发出的语言、文字、公式、数据、声音、图像等，在通信中都叫做消息(Message)。每个消息都是具体的，其内容千千万万，形式多种多样。消息中包含着信息，但消息却不等于信息。能够替代消息并适合于在信道中传输，是连续的或脉冲的电压、电流、电磁波及光波，它们叫做信号(Signal)。信号是信息的载体，信号也不等于信息。借助通信，信宿获得的是知识、情报、机密、商情、情感交流和视听享受等，它们仍不能作为信息的确切定义。信息是消息的内涵，是信号的价值，是能使信宿获知解惑的东西。它应当是从千千万万不同形式、不同内容的消息中抽象出来的、具有共性的、可定量测度的一个量，应该有它的单位和数学表达。就如同从金银铜铁、柴米油盐中提取出质量这个物理量一样，不管各种物体的物理化学性质的有何差异、几何属性有何区别、使用价值如何不同，总可以把物体所含物质的多少定义为该物体的质量，并且用克作为质量的单位。

为了给信息下个科学的定义，应对通信过程进行更加深入地分析。通信过程发生前，信宿并不知道信源将发出什么消息，也就是说，信源发出的消息存在某种不确定性；借助通信，信宿明确了原来不明确的一些事情，增加了对信源所述事情的了解，减少了甚至消除了原来的疑问，于是认为它获得了一些信息。可见通信是消除不确定性的过程，消除掉的不确性越多，信宿获得的信息就越多，信息的多少可以用通信所消除掉的不确定性来度量。

进一步分析，信源的不确定性又来自何处呢？例如，信源发出一封信，无论采用何种文字，它总是取自某个字符集中若干字符的组合，而这个组合，在千千万万个相同数目的字符构成的各种组合中不过是一个随机样本。信源究竟选取哪个样本作为要发送的消息，存在着不确定性。可见，信源消息的不确定性归根结底来自客观事物的多样性和随机性。从这个意义上讲，信息是客观事物存在方式和运动状态的多样性和不确定性的度量。

不确定性就是随机性，随机事物是用概率统计方法描述的。概率大的事物，出现的可能性大，不确定性就小；概率最大为 1，概率百分之百就是肯定要发生的事物，其不确定性为 0。反之，概率小的事物，出现的可能性小，不确定性就大；概率为 0，就是不会出现的事物，其不确定性无穷大。注意：数字通信中，数据一般用二进制表示，为方便起见，对数的底取为 2，这时自信息的单位叫比特(bit)，简称 b。本书中，除非有特别声明之处，均默认对数的底取为 2，例如，$\log^p$ 表示以 2 为底 p 的对数。因此，如果某事物出现概率为 p，则可定义为

$$I = \log(1/p) = -\log p \tag{1.1}$$

为该事物的不确定度，也称为该事件的自信息。

之所以取对数，是考虑到我们的定义应当符合以下事实：互不相关事件同时出现的概率是各事件单独出现概率之积，而总的自信息却应当是各事件自信息之和，即

$$I = \log(1/(p_1p_2)) = \log(1/p_1) + \log(1/p_2) = -\log p_1 - \log p_2 = I_1 + I_2 \tag{1.2}$$

自信息不过是不确定度的代名词，通信所消除掉的不确定度才是信宿所获得的信息量。设信源发出消息 X 的先验概率为 $p(x)$，则信源消息的的先验不确定度为 $-\log[p(x)]$；经信道传输完成通信过程后，如果是理想信道，信宿无失真地收到了这个消息，信源所发消息的不确定度被全部消除了，信宿获得信息的量值就等于信源所发消息的不确定度。

然而，一般情况下消息在传输过程中会受到噪声干扰，信宿收到消息可能有失真、有差错。尽管如此，信宿毕竟得到了一些关于原发消息的描述，对原发消息的不确定性一般会比通信前减少。不确定度的减小就意味着确定了一些事情，其自信息减少的数量即为信宿所获得的信息。但因这个量值是建立在概率理论基础之上的，计算只能具有统计上的意义。设信宿收到的消息为 Y，在已收到 Y 的条件下按统计规律可计算出关于消息 X 的后验概率为 $p(x|y)$，由此所定义的后验不确定度 $-\log[p(x|y)]$ 就代表了收到 Y 后关于消息 X 仍存在的不确定度。统计平均而言，后验不确定度的平均值 $\mathrm{E}[-\log\{p(x|y)\}]$ 小于先验不确定度的平均值 $\mathrm{E}[-\log\{p(x)\}]$。二者之差就是信宿获得的平均信息量。

1.2 离散信源

1.2.1 信息熵

1. 信息熵的定义

信源 X 从 m 个符号 $\{a_1, a_2, a_3, \cdots, a_m\}$ 组成的字符集 A 中随机选取字符来发送信息，这样的信源属于离散信源。

设信源取各字符的概率分别为 $p_1, p_2, p_3, \cdots, p_m$；可用概率矢量

$$\boldsymbol{p}(X) = (p_1, p_2, p_3, \cdots, p_m) \tag{1.3}$$

来描述。由式(1.1)的定义，信源发出各字符的自信息分别为

$$I_i = -\log p_i \quad (i = 1,2,\cdots,m)$$

所以信源发出单个字符的平均自信息为

$$H(X) = \sum_{i=1}^{m} p_i I_i = -\sum_{i=1}^{m} p_i \log p_i \tag{1.4}$$

式中，$H(X)$叫做信源的(单符号)信息熵，也叫先验熵，它反映信源平均发送每个符号的不确定度。

信源发送的符号一旦通过理想信道传输而被信宿无失真地接收后，这些不确定度就被完全消除了，信宿从每个信源符号中平均可获得量值为 $H(X)$的信息量。从这个意义上讲，信源熵也代表了平均每个信源符号所能承载的信息量。

需要指出的是，除了信源发信的先验概率外，通信过程中还会涉及到多种概率，每一种概率都代表一种不确定度，都可以定义相应的信息熵。例如，由联合概率 $p(xy)$ 可定义联合熵 $H(XY) = -\sum\sum p(xy)\log p(xy)$；由条件概率 $p(x \mid y)$ 可定义条件熵 $H(X \mid Y) = -\sum\sum p(xy)\log p(x \mid y)$，它们在后面的学习中会经常遇到。

【例 1】 信源发出 A、B、C、D4 个符号，其概率分别为 3/8、1/4、1/4、1/8。(1)求各符号的自信息和信源信息熵；(2)求信源发出符号序列 *ABAABCDBADCADCBDABCABABAACAAADCABCDACBABCABCABAACBDDCAAABC* 时，序列总自信息和平均每符号自信息。

解：(1) 已知各符号概率 $p(A) = 3/8, p(B) = 1/4, p(C) = 1/4, p(D) = 1/8$，

根据公式(1.1)，各符号的自信息分别为

$$I(A) = \log 8/3 = 1.415(\text{b})$$

$$I(B) = I(C) = \log 4 = 2(\text{b})$$

$$I(D) = \log 8 = 3(\text{b})$$

根据公式(1.4)，信源信息熵为

$$H(X) = 3/8I(A) + 1/4I(B) + 1/4I(C) + 1/8I(D) = 1.906(\text{b/符号})$$

(2) 因符号序列中含有 23 个 A、14 个 B、13 个 C、7 个 D，

序列总自信息为

$$I = 23I(A) + 14I(B) + 13I(C) + 7I(D) = 107.546(\text{b})$$

平均每符号自信息为

$$H = 107.546/57 = 1.89(\text{b/符号})$$

讨论：信息熵 $H(X)$是信源发送大量符号的统计平均结果，反映信源的性质；而给定序列的平均自信息 H 只代表某次通信的结果，它只对这个样本序列有意义。随机样本偏离统计平均值是正常现象。

【例 2】 二元信源的概率矢量可写为 $p = (x, 1-x)$，画出信息熵随概率 x 的变化曲线 $H(x)$。

解：$H(x) = -x\log x - (1-x)\log(1-x)$

x	0	0.1	0.2	0.3	0.4	0.5
$H(x)$	0	0.47	0.72	0.88	0.97	1.0

注意两点：(1)$x = 0$ 时，$-x\log x$ 为 $0\cdot\infty$ 的形式。由洛必达法则可确定：

$$\lim_{x\to 0}(-x\log x) = \lim_{x\to 0}\frac{\log x}{-\frac{1}{x}} = \lim_{x\to 0}\frac{\frac{1}{x}}{\frac{1}{x^2}}\lim_{x\to 0}x = 0$$

(2) 如图 1.1 所示,因为当 $x \geqslant 0.5$ 时 $1-x \leqslant 0.5$,所以 $H(x)$以 $x=0.5$为对称轴,在 $x=0.5$ 处取极大值。

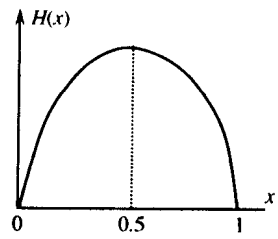


图 1.1 二元信源的信息熵

2. 信息熵的基本性质

(1) 非负性 由于概率值为 $0 \leqslant p_i \leqslant 1$ 的数,当对数的底大于 1 时,必有 $\log(p_i) \leqslant 0$,因此恒有

$$H(X) \geqslant 0 \tag{1.5}$$

(2) 对称性 m 个 p_i 任意交换顺序,信息熵 $H(X)$数值不变。其含义之一是,熵只取决于所有符号的概率分布方式,而与具体哪个符号具有哪个概率无关。其含义之二是,既然熵只取决于概率分布方式,那么与符号叫什么名字是没有关系的,无论符号是$(a_1、a_2、a_3)$、(红、绿、蓝)还是(晴、雨、风),只要概率分布一样,信息熵就一样。这正是编码能传承信息的原因。

(3) 极值性 信源符号等概分布时信息熵达到极大值,即

$$H(X) = \log m$$

证明:设信源符号概率分布矢量为 $\boldsymbol{p}(X) = (p_1, p_2, p_3, \cdots, p_m)$,另构造一随机矢量 $\boldsymbol{y} = (y_1, y_2, y_3, \cdots, y_m)$,满足 $y_i = 1/p_i (i = 1, 2, \cdots, m)$。

数学上有一个定理,若 $f(\boldsymbol{y})$是随机矢量 $\boldsymbol{y}$ 的∩型凸函数,则随机函数 $f(y_i)$的统计平均值不大于随机变量 y_i统计平均值的函数,这叫做詹森不等式。具体写出来为

$$\mathrm{E}[f(y_i)] \leqslant f(\mathrm{E}[y_i]) \tag{1.6}$$

即

$$\sum_{i=1}^{m} p_i f(y_i) \leqslant f\left(\sum_{i=1}^{m} p_i y_i\right) \tag{1.7}$$

取 $f(y_i)$为 $\log(y_i)$,则上式变为

$$\sum_{i=1}^{m} p_i \log(y_i) \leqslant \log\left(\sum_{i=1}^{m} p_i y_i\right)$$

代入 $y_i = 1/p_i$ 得

$$\sum_{i=1}^{m} p_i \log\left(\frac{1}{p_i}\right) \leqslant \log\left(\sum_{i=1}^{m} p_i \frac{1}{p_i}\right) = \log(m)$$

即

$$H(X) \leqslant \log m$$

而等概信源指信源符号的概率分布为

$$p_1 = p_2 = p_3 = \cdots = p_m = 1/m \tag{1.8}$$

的信源,由式(1.4)关于信息熵的定义,不难求出等概信源的信息熵为

$$H_0 = \log m \tag{1.9}$$

m 是字符集中字符总数。因此

$$H(X) \leqslant H_0 \tag{1.10}$$

等概信源是不确定性最大的信源。不等概信源发出一些符号比另一些符号概率大,这表明发送符号时具有某种倾向性,倾向性出现就意味着随机性受限,所以不等概信源的不确定性要比等概信源小。从例 2 中的二元信源已清楚地看到这个规律。

3. 码率和信息率

定义单位时间信源发送符号的个数叫做码率,也叫波特率,用 R_B(每秒符号数)表示。

而每个符号平均只有量值为 $H(X)$的自信息,因此,又定义单位时间信源发送信息的量值叫做信息率,也叫比特率,用 R_b(每秒比特数)表示。显然

$$R_b = R_B H(X)$$

许多情况下,信源符号概率未知或不需要知道,往往用平均每符号的最大信息荷载量(即最大熵)来衡量问题,二元信源的最大熵是 log2 = 1b,这就是我们常把 1 位二进码叫做 1b 的原因。多元(M 元)信源,其最大熵为 log$M$;然而一般总是选取符号数目为 $M = 2^k$,比如八进制、十六进制符号等都比较常见。于是每个 M 元符号的最大信息荷载量就是 kb。

1.2.2 无记忆信源的信息熵

实际通信过程中,信源发送消息往往不是单个符号,而是符号序列。当字符组成序列(如句子或文章)时,会出现两个新问题:一是随着序列的伸延,信源选取字符的概率是否随着时间改变;二是序列前后字符之间是否统计相关。对于第一个问题,假设所讨论的信源是平稳信源,即信源选取字符的概率不随时间改变。对于第二个问题,分两种情况来讨论:字符之间不存在统计关联的信源叫做无记忆信源;字符之间存在统计关联的信源叫做有记忆信源。

设离散平稳无记忆信源发出长度为 N 的符号序列 $X_1X_2\cdots X_N$,每位符号都是从字符集 $A = \{a_1, a_2, a_3, \cdots, a_m\}$ 中随机选取字符,各字符被抽取的概率分别为 $p_1, p_2, p_3, \cdots, p_m$。由于序列 X 中各字符间统计无关,所以序列 $X_1X_2\cdots X_N$ 出现的概率为

$$p(x_1x_2\cdots x_N) = p(x_1)p(x_2)\cdots p(x_N) \tag{1.11}$$

因此,序列 $X_1X_2\cdots X_N$ 的信息熵为

$$H(X_1X_2\cdots X_N) = \sum_{x_1}\sum_{x_2}\cdots\sum_{x_N} p(x_1x_2\cdots x_N)\log p(x_1x_2\cdots x_N) \tag{1.12}$$

将式(1.11)代入式(1.12)得

$$\begin{aligned} H(X_1X_2\cdots X_N) = & \sum_{x_1}\sum_{x_2}\cdots\sum_{x_N} p(x_1)p(x_2)\cdots p(x_N)\cdot \\ & [\log p(x_1) + \log p(x_2) + \cdots + \log p(x_N)] = \\ & \sum_{x_1} p(x_1)\log p(x_1) + \sum_{x_2} p(x_2)\log p(x_2) + \cdots + \sum_{x_N} p(x_N)\log p(x_N) = \\ & H(X_1) + H(X_2) + \cdots + H(X_N) \end{aligned}$$

又由于是平稳信源,$H(X_1) = H(X_2) = \cdots = H(X_N)$,都等于单符号信息熵 $H(X)$,所以有

$$H(X_1X_2\cdots X_N) = N \cdot H(X) \tag{1.13}$$

定义信源序列的平均符号熵为

$$H_N = H(X_1X_2\cdots X_N)/N \tag{1.14}$$

则无记忆信源的平均符号熵等于单符号信息熵,即

$$H_N = H(X) \tag{1.15}$$

1.2.3 有记忆信源的信息熵

有记忆信源指序列各位符号之间统计相关,信源先后发出的符号之间受条件概率

$$p(x_1x_2\cdots x_N) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1x_2)\cdots p(x_N \mid x_1x_2\cdots x_{N-1}) \tag{1.16}$$

的约束。将式(1.16)代入式(1.12)中,不难推导出

$$H(X_1X_2\cdots X_N) = H(X_1) + H(X_2 \mid X_1) + H(X_3 \mid X_1X_2) + \cdots + H(X_N \mid X_1X_2\cdots X_{N-1}) \tag{1.17}$$

式中 $H(X_1) = -\sum_{x_1} p(x_1)\log p(x_1)$

$$H(X_2 \mid X_1) = -\sum_{x_1}\sum_{x_2} p(x_1x_2)\log p(x_2 \mid x_1)$$

$$H(X_3 \mid X_1X_2) = -\sum_{x_1}\sum_{x_2}\sum_{x_3} p(x_1x_2x_3)\log p(x_3 \mid x_1x_2)$$

$$\cdots$$

$$H(X_N \mid X_1X_2\cdots X_{N-1}) = -\sum_{x_1}\sum_{x_2}\cdots\sum_{x_N} p(x_1x_2\cdots x_N)\log p(x_N \mid x_1x_2\cdots x_{N-1})$$

为各阶条件熵。可以证明,条件熵不大于无条件熵,强条件熵不大于弱条件熵,即

$$H(X_1) \geqslant H(X_2 \mid X_1) \geqslant H(X_3 \mid X_1X_2) \geqslant \cdots \geqslant H(X_N \mid X_1X_2\cdots X_{N-1}) \tag{1.18}$$

这里仅证明 $H(X_1) \geqslant H(X_2 \mid X_1)$,后面的不等式部分不再证明,道理与此相同。

因为

$$H(X_2 \mid X_1) = -\sum_{x_1}\sum_{x_2} p(x_1x_2)\log p(x_2 \mid x_1) = -\sum_{x_1}\sum_{x_2} p(x_1)p(x_2 \mid x_1)\log p(x_2 \mid x_1)$$

利用詹森不等式

$$\sum_{i=1}^{m} p_i f(y_i) \leqslant f\left(\sum_{i=1}^{m} p_i y_i\right)$$

现在视 p_i 为 $p(x_1)$,y_i 为 $p(x_2 \mid x_1)$,$f(y_i)$ 为 $-y_i\log y_i$,求和是对 i(即 x_1)进行,则不等式左边

$$\sum_i p_i f(y_i) = -\sum_{x_1} p(x_1)p(x_2 \mid x_1)\log p(x_2 \mid x_1)$$

右边

$$f\left(\sum_i p_i y_i\right) = -\left(\sum_i p_i y_i\right)\log\left(\sum_i p_i y_i\right)$$

但

$$\sum_i p_i y_i = \sum_{x_1} p(x_1)p(x_2 \mid x_1) = \sum_{x_1} p(x_1x_2) = p(x_2)$$

所以

$$f\left(\sum_i p_i y_i\right) = -p(x_2)\log p(x_2)$$

代入得

$$-\sum_{x_1} p(x_1)p(x_2 \mid x_1)\log p(x_2 \mid x_1) \leqslant -p(x_2)\log p(x_2)$$

两边再对 j(即 x_2)求和,即得

$$H(X_2 \mid X_1) \leqslant H(X_2)$$

对于平稳离散信源,$H(X_1) = H(X_2)$,代入上式,即得到证明结果

$$H(X_1) \geqslant H(X_2 \mid X_1) \tag{1.19}$$

式中等号对应无记忆信源。

仍定义信源序列的平均符号熵为 $H_N = H(X_1X_2\cdots X_N)/N$,在式(1.16)中,分别取序列长度为 1,2,…,N,并利用式(1.17),可得

$$H(X_1) = H_1 = H(X)$$

$$H_1 \geqslant H_2 \geqslant H(X_2 \mid X_1)$$

$$H_2 \geqslant H_3 \geqslant H(X_3 \mid X_1X_2)$$

$$\vdots$$

$$H_{N-1} \geqslant H_N \geqslant H(X_N \mid X_1X_2\cdots X_{N-1})$$

由此得到以下结论:

(1) 序列的平均符号熵不大于单符号信息熵

$$H_N \leqslant H(X_1) = H_1 \tag{1.20}$$

对无记忆信源取等号。

(2) 随着序列长度 N 的增大,各符号之间的相互制约增加,不确定程度将会减小,使有记忆信源的平均符号熵呈减小趋势。连同 $H_0 = \log m$ 表示无记忆等概信源的信息熵,可写为

$$H_0 \geqslant H_1 \geqslant H_2 \geqslant H_3 \geqslant \cdots \geqslant H_{N-1} \geqslant H_N \tag{1.21}$$

(3) 随着序列长度 N 的增大,平均符号熵 H_N 与同阶条件熵 $H(X_N \mid X_1X_2\cdots X_{N-1})$的差别会越来越小,当 $N \to \infty$ 时有

$$H_\infty = \lim_{N\to\infty} H_N = \lim_{N\to\infty} H(X_N \mid X_1X_2\cdots X_{N-1}) \tag{1.22}$$

它叫做极限熵,代表实际信源的信息熵。

(4) 对于有记忆信源,只有极限熵才是严格意义上的信源平均符号熵,而 $H_1, H_2, H_3, \cdots, H_N$都是近似的平均符号熵。这是因为有限长序列不能包含有记忆信源的全部关联,它只能看作是从无穷长的消息序列中截取的一段,计算序列熵时必然忽略了该序列与序列外的符号的关联,称之为"截短近似"。

【例 3】 某有记忆信源发出符号时相邻符号的条件概率为 $p(0|0) = 0.8$, $p(0|1) = 0.1$,求它的最大熵 H_0、平均符号熵 H_1、H_2、H_3 和极限熵 H_∞。

解:(1) 最大熵 H_0

假设信源等概且无记忆,才能达到最大熵,即

$$H_0 = \log m = \log 2 = 1(\text{b/符号})$$

(2) 平均符号熵 H_1

先计算信源发送单个符号 $X=(0,1)$的概率 $p(X)$。

由 $p(0|0)=0.8$ 知 $p(1|0)=0.2$;由 $p(0|1)=0.1$ 知 $p(1|1)=0.9$。

利用
$$p(x_2)=\sum_{x_1}p(x_1x_2)=\sum_{x_1}p(x_1)p(x_2\mid x_1)$$

当 $x_2=0$ 时
$$p(0)=p(0)p(0\mid 0)+p(1)p(0\mid 1)=0.8p(0)+0.1p(1)$$

即
$$0.2p(0)=0.1p(1)$$

连同归一化条件
$$p(0)+p(1)=1$$

解得
$$p(0)=1/3;p(1)=2/3$$

于是有
$$H_1=H(X)=-(1/3)\log(1/3)-(2/3)\log(2/3)=0.9183(\text{b/符号})$$

(3) 平均符号熵 H_2

信源发出长度为 2 的序列 $X_1X_2=(00,01,10,11)$,其概率可由 $p(x_1x_2)=p(x_1)p(x_2|x_1)$计算,即
$$p(X_1X_2)=\left(\frac{1\times0.8}{3},\frac{1\times0.2}{3},\frac{2\times0.1}{3},\frac{2\times0.9}{3}\right)=(0.2667,0.0667,0.0667,0.6)$$
$$H(X_1X_2)=-0.2667\log0.2667-0.0667\log0.0667-0.0667\log0.0667-0.6\log0.6=1.4716$$
$$H_2=H(X_1X_2)/2=0.7359(\text{b/符号})$$

(4) 平均符号熵 H_3

信源发出长度为 3 的序列 $X_1X_2X_3=(000,001,010,011,100,101,110,111)$,其概率可由
$$p(x_1x_2x_3)=p(x_1)p(x_2|x_1)p(x_3|x_1x_2)=p(x_1)p(x_2|x_1)p(x_3|x_2)$$
计算。由于已知信源符号只存在两两相关,故上式中 $p(x_3|x_1x_2)=p(x_3|x_2)$。计算结果为
$$p(X_1X_2X_3)=(0.2133,0.0533,0.0067,0.06,0.0533,0.0133,0.06,0.54)$$
所以
$$H(X_1X_2X_3)=2.0249$$
$$H_3=H(X_1X_2X_3)/3=0.6750(\text{b/符号})$$
显然,符合 $H_0\geqslant H_1\geqslant H_2\geqslant H_3$ 的规律。

(5) 极限熵 H_∞

二阶条件熵
$$H(X_2\mid X_1)=-\sum_{x_1}\sum_{x_2}p(x_1x_2)\log p(x_2\mid x_1)=$$
$$-0.2667\log0.8-0.0667\log0.2-0.0667\log0.1-0.6\log0.9=0.5533(\text{b/符号})$$

三阶条件熵
$$H(X_3\mid X_1X_2)=-\sum_{x_1}\sum_{x_2}p(x_1x_2x_3)\log p(x_3\mid x_1x_2)$$

代入数据后仍求得

$$H(X_3 | X_1X_2) = 0.5533(\text{b/符号})$$

实际上，由于已知信源符号只存在两两相关，$N \geqslant 2$ 以后，条件熵不再随着序列长度而变化，所以 $N \to \infty$ 时有

$$H_\infty = \lim_{N\to\infty} H_N = \lim_{N\to\infty} H(X_N | X_1X_2\cdots X_{N-1}) = 0.5533(\text{b/符号})$$

1.2.4 马尔科夫信源

1. 马尔科夫信源的定义和数学模型

考虑到有记忆信源符号之间的关联，总是邻近符号之间关联较强，距离远的符号之间关联较弱。于是，人们提出“马尔科夫信源”模型：假设距离小于或等于 N 的符号之间存在关联，距离大于 N 的符号之间不存在关联，N 为关联长度。关联长度 $N=2$ 时，符号两两相关，叫做一阶马尔科夫信源；关联长度 $N=3$ 时，符号三三相关，叫做二阶马尔科夫信源；关联长度为 $N+1$ 时，叫做 N 阶马尔科夫信源。

首先必须强调，马尔科夫信源模型不等于截短近似模型，截短近似只处理 N 个符号，马尔科夫信源所处理的乃是一个无穷长的链。马尔科夫信源假设关联长度为 N，并非是从序列中截取 N 个符号，虽然在该模型中只有邻近的 N 个符号有关联，更远的符号没关联，但是通过近邻关联，近邻又连接近邻，使整个无穷长链都连锁起来。即使是 $N=2$ 的一阶马尔科夫信源，它也通过两两相关把所有的符号连在一起，统统加以处理。马尔科夫近似比截短近似的优越之处在于它合理地假设了关联的程度，而没有硬性地切断关联。

为了从理论上处理无穷长链的问题，引入了“状态”的概念：把排在指定符号前面与它相关联的 $N-1$ 个符号定义为该符号的状态。由于所有符号都取自 m 个符号的集合，长度为 $N-1$ 的符号串共有 $L = m^{N-1}$ 种不同的形式，就是说状态数目共有 L 个。只要知道 L 个状态下分别发出 m 个符号的条件概率，无限长序列中的全部关联就都清楚了。于是求解这个无穷长的互相连锁的链的复杂问题就被简化为求解有限个状态决定有限个符号的简单问题。

2. 状态转移与稳态概率

对于给定的马尔科夫信源，各状态下发出符号的概率应是确定的。然而，随着序列的延伸，当前符号的位置在不断前移，相应的状态也在不断变化，这种变化称为状态转移。新状态由老状态去掉最前头的符号，再接上新发出的符号构成，所发的符号确定了，转移成什么状态也就被确定了。

【例 4】 二阶马尔科夫信源已知条件概率 $p(0|00) = p(1|11) = 0.8$，$p(0|01) = p(1|10) = 0.6$，求它的状态符号依赖关系和状态转移概率。

解：二阶马尔科夫信源相关长度为 3，每个符号与它前面 2 个符号有关联。状态由 2 符号组成，共有 4 个状态。分别为 $E_1 = 00$，$E_2 = 01$，$E_3 = 10$，$E_4 = 11$，已知的条件概率为

$$p(0|E_1) = p(1|E_4) = 0.8; p(0|E_2) = p(1|E_3) = 0.6。$$

根据归一化条件可求出另外 4 个状态符号依赖关系为

$$p(1|E_1) = p(0|E_4) = 0.2; p(1|E_2) = p(0|E_3) = 0.4;$$

$E_1 = 00$ 发出 0 变为 $\boxed{0}\,00 = E_1$，其状态转移概率 $P(E_1|E_1) = p(0|E_1) = 0.8$；