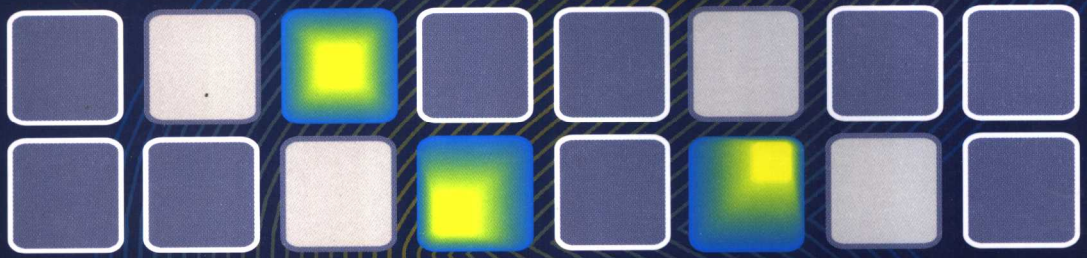
信息科学基础



信息论、编码及密码理论

辛小龙 主编

西北大学出版社
NORTHWEST UNIVERSITY PRESS

信息科学基础

——信息论、编码及密码理论

辛小龙　主编

西北大学出版社

图书在版编目(CIP)数据

信息科学基础:信息论、编码及密码理论/ 辛小龙主编. —西安: 西北大学出版社,2006. 8

ISBN 7-5604-2230- 6

I. 信... II. 辛... III. 信息技术 IV. G202

中国版本图书馆 CIP 数据核字(2006)第 110896 号

书　　名	信息科学基础 ——信息论、编码及密码理论
主　　编	辛小龙
出版发行	西北大学出版社
地　　址	西安市太白北路 229 号
邮　　编	710069
电　　话	029 - 88302590
经　　销	新华书店
印　　刷	西安华新彩印有限责任公司
开　　本	787mm × 960mm　1/16
印　　张	13. 75
字　　数	224 千字
版　　次	2006 年 8 月第 1 版　2006 年 8 月第 1 次印刷
书　　号	ISBN 7-5604-2230- 6/G · 319
定　　价	18. 00 元

内容提要

本书从信息科学的基本概念和基本方法入手，主要使用初等的数学工具，系统而又全面地介绍了信息论、编码和密码理论的基本理论、基本方法和近年来发展的新成果. 全书共分 12 章. 第 1 章通过特殊信道，描绘了信息论的核心思想. 第 2 章至第 6 章，介绍了仙农信息论的基本理论，包括信息度量、信源编码定理、重要的信源编码方法、信道编码定理、信道容量及其计算、率失真函数及其计算等内容. 第 7 章至第 9 章，介绍了编码理论基本内容，包括线性码的基本概念和编译码方法、循环码、RS 码和 Goppa 码等内容. 第 10 章至第 12 章，介绍了古典密码学、近代密码学的一些基本理论和方法，还介绍了近年发展的最新的密码系统.

本书适宜作信息与计算科学、应用数学、概率统计、计算机科学、通信工程等专业的本科生教材，也可作为这些专业研究生的学习参考，并可供有关工程技术人员参考.

前言

信息论是人们在长期通信工程的实践中，由通信技术与概率论、随机过程和数理统计相结合而逐步发展起来的一门科学. 通常人们公认信息论的奠基人是美国科学家仙农(C. E. Shannon)，他在 1948 年发表了著名的论文《通信的数学理论》，为信息论奠定了理论基础. 半个世纪以来，以通信理论为核心的经典信息论，包括编码理论与密码理论，正以信息技术为物化手段，向高精尖方向迅猛发展，并以神奇般的力量把人类社会推入到了信息时代. 随着信息理论的迅猛发展和信息概念的不断深化，信息论所涉及的内容早已超越了狭义的通信范畴，进入到了信息科学这一更广阔、更新的领域.

本教材着重介绍仙农信息理论的基本概念、基本分析方法和主要结论，其中包括了编码理论与密码理论. 全书共分 12 章. 第 1 章通过一对特殊的数学模型——二进对称信源和二进对称信道，介绍了信息理论的核心思想. 第 2 章主要针对离散信源，介绍了信息度量，以及信息熵、条件熵、联合熵、相对熵和互信息等概念，讨论了它们的一些主要性质. 第 3 章介绍了无失真信源编码，应用渐近等分性证明了信源编码定理，同时介绍了一些变长编码方法. 第 4 章针对离散信源，介绍了信道编码定理，引入了信道容量的概念并讨论了信道容量的计算. 第 5 章讨论了限失真信源编码，引入率失真函数的概念，讨论了率失真函数的计算问题. 第 6 章讨论了连续信源的信息度量，介绍了连续信源的率失

真函数、高斯信道及其信道容量等概念. 第 7、8 和 9 三章介绍了编码理论中最基本的内容. 第 7 章介绍了线性码的基本概念,讨论了伴随式译码方法,介绍了线性码及其对偶码的重量分布,给出了 Macwilliams 恒等式. 第 8 章介绍了循环码,并介绍了用循环码纠正突发错误,进而介绍了两类重要的特殊循环码——BCH 码、戈雷码. 第 9 章介绍了 RS 码和 Goppa 码的基本概念和理论. 第 10、11 章,介绍了古典密码学、近代密码学的一些重要专题,第 12 章介绍了密码学的一些最新进展.

本教材受西北大学重点课程项目资助,在此对西北大学教务处的支持表示感谢. 教材由辛小龙主编,杜凯、王伟、张静参加了教材资料的收集和教材的编写工作,其中,杜凯参加了第 2 至 4 章的编写,王伟参加了第 10、11 章的编写工作,张静参加了第 12 章的编写工作. 虽然在编写过程中参阅了大量的教科书、专著和文献,但由于编者水平所限,书中错误在所难免,敬请读者批评和指正.

编者

2006 年 7 月 30 日

目录

目录

目录

目录

第一章　绪　论

1948 年，C. Shannon 在他的题为“通信中的数学理论”一文中写道：“通信的基本问题是将某人在一个地方得到的消息精确地或大致地在另一个地方重现.”为了解决这一问题，他创立了一个完全崭新的应用数学分支，即信息论和编码理论.

在这一章中，我们将以二元对称信源和信道为模型，将信息论的思想作以介绍.

所谓二元对称信源，是指发射两个符号，即 0 和 1，且在单位时间发射 R 个符号的信源. 我们称 R 为信源的速率，并且假定 R 是连续的. 我们简称这些符号为“比特(bits)”.

二元对称信道(BSC) 是每单位时间传送一个比特的通信设施. 然而，信道并不是完全可靠的，假设它的信道原始错误概率为 p，且设 $0 \leqslant p \leqslant \frac{1}{2}$，从而，输出的比特和输入的比特不完全相同.

现在假设有两个对象，发送者和接收者. 发送者将尽可能准确地将信源输出的符号传送给接收者. 假定传送信道是二元对称信道.

现在我们考虑这样的问题：给定信源速率 R，发送者和接收者在二元对称信道上的通信的准确性有多高?对这一问题，我们最终将给出一个比较准确而且是一般性的答案，首先我们先从一些特殊情况入手.

假定 $R = \frac{1}{3}$，即信道传送的比特数比信源产生的比特数快 3 倍，所以信源输出的数据在传送前每比特可以被重复编码 3 次. 例如，如果信源输出数据的前 5 个比特是 10100，则编码后的符号串是 111000111000000. 接收者将收到每个信源数据比特的 3 个版本. 由于信道噪音，这 3 个版本或许不完全相同. 如果信道错传了信源符号串的第 2，5，6，12 和 13 位的符号，接收者将收到的字符串为 101011111001100. 不难想像，在这种情形下，接收者的最佳的译码方案是：在接收到的 3 个版本中，采取“多数占优”译码方案. 应用这种译码方案，在我们以上给出的例子中，接收者将收到的信息译为 11100，在第 2 位出错. 一般地，一个信源比特在接收中出错，仅在信道将 3 个版本中的 2 个或 3 个传错时才会出现. 所以，如果 P_e 表示比特传输错误的概率，则

$$\begin{aligned} P_e &= P\{2\text{ 个信道错误}\} + P\{3\text{ 个信道错误}\} \\ &= 3p^2(1-p) + p^3 \\ &= 3p^2 - 2p^3 \end{aligned} \tag{1.1}$$

因为 $p \leqslant \frac{1}{2}$，所以，比特传输错误概率 P_e 应小于信道原始错误概率 p. 我们 的一个简单的编码方案就改善了信道传输的可靠性. 当 p 充分小时，可靠性的改善程度非常高.

现在我们容易看到，当信道传输消息时，信源符号被重复的次数越多，可靠性就越高. 那么，如果 $R = 1/(2n+1)$，则我们将每个信源输出的比特在信道传送前重复 $2n+1$ 次并且应用“多数占优”译码方案. 从而我们可以得最终的比特错误概率 $P_e^{(2n+1)}$ 的一个表达式：

$$\begin{aligned} P_e^{(2n+1)} &= \sum_{k=n+1}^{2n+1} P\{2n+1\text{ 位被发送的符号串中第 } k \text{ 位出错}\} \\ &= \sum_{k=n+1}^{2n+1} \binom{2n+1}{k} p^k (1-p)^{2n+1-k} \\ &= \binom{2n+1}{n+1} p^{n+1} + p \text{ 的次数更高的项.} \end{aligned} \tag{1.2}$$

如果 $n > 1$，上式当 $p \to 0$ 时，比 $n = 1$ 时更快地趋于 0. 所以在这个相对弱的意义上，较长的重复编码方案要比较短的重复编码方案要好. 然而，我们可以给出更强的断言：对给定的信道原始错误概率 $p < \frac{1}{2}$ 的一个二元对称信道，当 $n \to \infty$ 时，$P_e^{(2n+1)} \to 0$，即，应用重复编码方案，信道的可靠性可以按照我们的期望进行改善. 我们应用弱大数定理来证明这一断言. 假定在以上信道上 N 个比特被传送，则对任意 $\varepsilon > 0$，由弱大数定理有

$$\lim_{N\to\infty} P\left\{\left|\frac{\text{信道错误个数}}{N} - p\right| > \varepsilon\right\} = 0 \tag{1.3}$$

换句话说，对充分大的 N，所收到的比特出现错误的分数 $f_e^{(N)}$ 和 p 无限接近. 从而，我们对 $P_e^{(2n+1)}$ 有以下估计：

$$\begin{aligned} P_e^{(2n+1)} &= P\{f_e^{(2n+1)} \geqslant \frac{n+1}{2n+1} = \frac{1}{2} + \frac{1}{4n+2}\} \\ &= P\{f_e^{(2n+1)} > \frac{1}{2}\} \end{aligned}$$

$$= P\{|f_e^{(2n+1)} - p| > \frac{1}{2} - p\}.$$

所以，当 $n \to \infty$ 时，由(1.3)式有 $P_e^{(2n+1)} \to 0$. 由此我们得出结论：即使信道有较大的噪音，当 R 充分小时，仍然有可能使得总体错误概率充分的小.

以上我们讨论了信源速率 $R < 1$ 时的情况. 那么当 $R > 1$ 时，通信的准确度如何呢?我们以下来讨论后一种情况.

假定 $R > 1$，则信道仅能传送信源输出比特的 $\frac{1}{R}$，而需要接收者去猜测其余的比特，这种猜测就像投掷一个质地均匀的硬币，正面和反面的概率均为 $\frac{1}{2}$. 对于这个不是非常清楚的译码方案，容易计算最终比特错误概率是

$$P_e = \frac{1}{R} \times p + \frac{R-1}{R} \times \frac{1}{2}$$

$$= \frac{1}{2} - (\frac{1}{2} - P)/R \tag{1.4}$$

另外，对某些 $R > 1$，一个稍有改进的译码方法是可行的. 不妨设 $R = 3$. 这时，信道单位时间所能传送的信源比特仅为信源发出的比特的 $\frac{1}{3}$. 所以送信者将信源发出的符号串每3位分一组，在每一组中采用“多数占优”方法发射一个比特. 比如，信源发射字符串为 101110101000101，则发送者将传送符号串 11101 通过信道. 接收者仅仅将所收到的每一个比特重复三次. 假如信道在第 2 位出错，接收者将收到 10101，他扩充这个信息为 111000111000111，从而出现 5 比特错误. 一般地，我们可以计算出最终比特错误概率为

$$P_e = \frac{1}{4} \times (1 - p) + \frac{3}{4} \times p$$

$$= \frac{1}{4} + \frac{1}{2}p \tag{1.5}$$

这个值小于 $R = 3$ 时，“投掷硬币”译码方案所得到的概率为 $\frac{1}{3} + \frac{1}{3}p$. 对其它的一些 R 的整数值，我们也可以用这个方案来编译码(见习题).

截至目前我们所考虑的编译码方案并不是毫无意义，但还是平凡的方案. 以下我们给出一个非平凡例子.

假定 $R = \frac{4}{7}$，那么，当信道传送信源发出的每4个比特符号时，就有时间发送额外3比特符号通过信道. 我们将用如下规则来选择这些附加比特：如果4个信源比特被记做 x_0, x_1, x_2, x_3，则额外的比特，或一致校验位，记为 x_4, x_5, x_6，被以下方程所确定：

$$\left.\begin{aligned} x_4 &= x_1 + x_2 + x_3 (\mathrm{mod}2) \\ x_5 &= x_0 + x_2 + x_3 (\mathrm{mod}2) \\ x_6 &= x_0 + x_1 + x_3 (\mathrm{mod}2) \end{aligned}\right\} \tag{1.6}$$

比如，$(x_0, x_1, x_2, x_3) = (0\ 1\ 1\ 0)$，则 $(x_4, x_5, x_6) = (0\ 1\ 1)$，并且完整的7位码字0 1 1 0 0 1 1 将被传送通过信道.

现在我们考虑给出一个译码算法，即接收者要从接收到的有错误的7位码字，来估计信源所发出的4位信息. 我们先将一致校验方程(1.6)重新写成以下形式：

$$\left.\begin{aligned} x_1 + x_2 + x_3 + x_4 &= 0 \\ x_0 + x_2 + x_3 + x_5 &= 0 \\ x_0 + x_1 + x_3 + x_6 &= 0 \end{aligned}\right\} \tag{1.7}$$

(方程组(1.7)中的算法都是模2下计算.) 如果我们定义一致校验矩阵 H 为：

$$H = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}$$

则16个可能的码字 $X = (x_0, x_1, x_2, x_3, x_4, x_5, x_6)$ 中的每一个码字都适合矩阵方程

$$HX^T = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \tag{1.8}$$

假定 $X = (x_0, x_1, \cdots, x_6)$ 是被传送的码字，$Y = (x_0 + z_0, x_1 + z_1, \cdots, x_6 + z_6)$ 是接收向量，即 $Y = X + Z$，其中 $Z = (z_0, \cdots, z_6)$ 表示错误格式.

作为接收者，仅仅知道 Y，然而他想通过 Y 知道 X. 为此，他可以计算向量 $S = (s_0, s_1, s_2)$，满足以下条件：

$$\begin{aligned} S^T &= HY^T \\ &= H(X + Z)^T \\ &= HX^T + HZ^T \end{aligned}$$

$$= HZ^T \tag{1.9}$$

我们称向量 S 为接收向量 Y 的伴随式(Syndrome). 伴随式中某一分量为0,表示接收向量 Y 适合对应的一致校验方程;某一分量为1表示 Y 不适合对应的一致校验方程. 由(1.9)式可知,伴随式不依赖于所传送的码字,而仅依赖于错误格式 Z. 然而,因为 $X = Y + Z$,那么接收者如果知道 Z,也将知道 X. 从而,他可以将问题转化为找 Z. 方程 $S^T = HZ^T$ 表明 S^T 是 Z 中分量为1的那些分量所对应的 H 矩阵中的列之和,即

$$S^T = Z^0\begin{pmatrix}0\\1\\1\end{pmatrix} + Z^1\begin{pmatrix}1\\0\\1\end{pmatrix} + \cdots + Z^6\begin{pmatrix}0\\0\\1\end{pmatrix} \tag{1.10}$$

接收者一旦计算出 S,则他的任务就是求解关于 Z 的方程 $S^T = HZ^T$. 注意到方程(1.10)中7个未知量仅有3个方程,从而对任何 S,将有16个可能的解 Z. 然而这也是一个进步,我们将128个可能性减少为16个. 那么如何从剩余的16个可能性中选择出正确答案呢?比如,假定 $Y = (0\,1\,1\,1\,0\,0\,1)$ 被接收,则计算出伴随式 $S = (1\,0\,1)$,而 Z 的16个候选者为

0 1 0 0 0 0 0	0 0 1 0 0 1 1
1 1 0 0 0 1 1	0 0 0 1 0 1 0
0 0 0 0 1 0 1	0 1 1 1 0 0 1
0 1 1 0 1 1 0	1 0 1 0 0 0 0
0 1 0 1 1 1 1	1 0 0 1 0 0 1
1 0 0 0 1 1 0	1 1 1 1 0 1 0
1 1 1 0 1 0 1	0 0 1 1 1 0 0
1 1 0 1 1 0 0	1 0 1 1 1 1 1

因为原始比特错误概率 $p < \frac{1}{2}$,错误格式 Z 的分量中1的个数越少,它越接近于真实的错误格式. 在上例中,具有最小重量的错误格式只有一个,即 $Z = (0\,1\,0\,0\,0\,0\,0)$. 从而,接受者对 Z 的最好估计为 $Z = (0\,1\,0\,0\,0\,0\,0)$;对被传送的码字的最好估计是 $X = Y + Z = (0\,0\,1\,1\,0\,0\,1)$,最后得到信源所发出的符号序列为(0 0 1 1).

当然,在以上例子中,我们并不是靠运气,因为我们能证明:对任何伴随式 S,都存在 $HZ^T = S^T$ 的惟一一个解,它的重量为0或1. 事实上,如果 $S = (0\,0\,0)$,则 $Z = (0\,0\,0\,0\,0\,0\,0)$ 是我们所期望的解. 如果 $S \neq (0\,0\,0)$,则 S^T 必作为矩阵 H 的某一列出现. 如果 S^T 为矩阵 H 的第 i 列,则错误格式 $Z = (0,\cdots,0,1,0,\cdots,0)$ 就是方程 $HZ^T =$

S^T 的惟一一个具有最小重量的解，其中 Z 的第 i 个分量为 1，其余分量为 0.

按照这个方案，我们可以给出一个译码算法. 我们称由 H 确定的线性码为汉明码. 现在我们给出汉明码译码步骤：

1. 计算伴随式 $S^T = HY^T$.

2. 如果 $S = 0$，设 $\hat{Z} = 0$；转到 4.

3. 如果 $S \neq 0$，在矩阵 H 中找到和 S 相对应的列，设为第 i 列；设 $\hat{Z}$ = 第 i 个分量为 1，其余分量为 0 的向量.

4. 设 $\hat{X} = Y + \hat{Z}$.（这就是译码器对传送码字 X 的估计）.

5. 输出 $(\hat{x}_0, \hat{x}_1, \hat{x}_2, \hat{x}_3)$，即向量 X 的前四位.（这就是译码器对原始信源符号序列的估计.）

当然，由这个算法产生的向量 $\hat{Z}$ 可能不等于实际错误格式 Z. 然而，如果信道至多产生一个错误，即 Z 的重量为 0 或 1，则由以上讨论知 $\hat{Z} = Z$. 从而，汉明码是一个纠正单个错误的码. 事实上，容易看到：以上译码算法不能够正确翻译出原始传送的码字 X 的充要条件是信道产生两个或更多的错误. 从而，如果 P_E 表示码字传输错误的概率 $P\{\hat{X} \neq X\}$，则

$$\begin{aligned} P_E &= \sum_{k=2}^{7} \binom{7}{k} p^k (1-p)^{7-k} \\ &= 21p^2 - 70p^3 + \cdots \end{aligned}$$

我们用 $P_e^{(i)}$ 表示第 i 个分量出错的概率 $P\{\hat{x}_i \neq x_i\}$，则对 $0 \leqslant i \leqslant 6$，有

$$\begin{aligned} P_e^{(i)} &= 9p^2(1-p)^5 + 19p^3(1-p)^4 + 16p^4(1-p)^3 \\ &\quad + 12p^5(1-p)^2 + 7p^6(1-p) + p^7 \\ &= 9p^2 - 26p^3 + \cdots \end{aligned} \tag{1.11}$$

比较(1.11)和(1.1)，我们可以看到，对二元对称信道，当原始错误概率非常小时，汉明码在信源速率 $R = 4/7 = 0.571$ 时与比较粗糙的重复编码方案在信源速率 $R = 1/3 = 0.333$ 时所完成的功能是相当的.

对 $R = 7/4$，我们可以转变送信者和接收者的角色，而利用(7,4)汉明码来进行通信. 送信者先将信源序列 7 位分一组，然后用以上的译码算法将每一组 7 位的序列减少为每一组 4 位，然后发送这 4 位信号序列通过信道. 接收者利用一致校验规则(1.6)将 4 位信号序列加上其它 3 位，译成 7 位接收序列. 在这个方案下，最后每个分量错误概率 $P_e^{(i)} = P(\hat{x}_i \neq x_i)$ 与 i 有关，但平均错误概率 $P_e = \left(\sum_{i=0}^{6} P_e^{(i)}\right)/7$ 被下式

给出：

$$P_e = \frac{1}{8}(1-p)^4 + \frac{53}{28}(1-p)^3 p + 3(1-p)^2 p^2 + \frac{59}{28}(1-p)p^3 + \frac{7}{8}p^4$$

$$= \frac{1}{8} + \frac{39}{28}p + \cdots \tag{1.12}$$

对无噪二元对称信道，即$p = 0$，这个结果是非常好的. 比如，利用“投掷硬币”估计错误概率计算方案(1.4). 当$R = 7/4$时，可计算出$P_e = 3/14 = 0.214$，它是大于用以上方法得到的错误概率.

现在，我们针对一个误传概率为$p = 0.1$的特殊二元对称信道BSC来将以上的讨论作以总结. 对应于我们迄今所讨论的每一个通信模式，在xoy平面上放置一个点(x,y)，具有$x = R$，且$y = P_e$，这里R为码率，P_e为误差概率，即图1.1中显示的. 我们将不断地建立这样的通信模型并放一个点在图1.1上. 我们的最终目标是希望了解哪些点能够达到而哪些点不能达到. 令人难以置信的是，这样的目标已经被Shannon所达到. 在给出Shannon的结果前，我们结合错误概率P_e，将编码模型中的码率R用新的公式来表示.

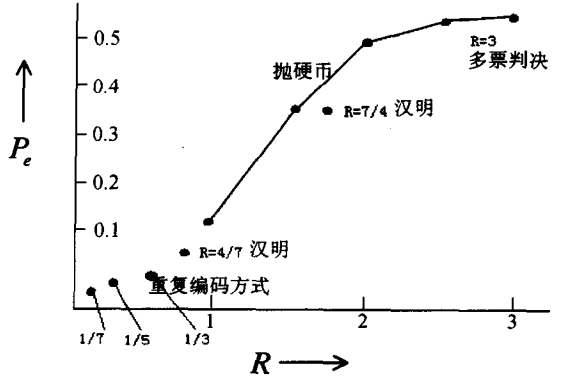


图1.1

图1.2给出了一个(n,k)码的通信模式. 在这个模型中，信源序列被进行k长分组，每一个k长信源序列U被编码为一个n长码字X，通过信道传送X，接收端接收的向量为Y. 译码器将带有噪音的n长码字Y翻译为k长向量V，它是原始信源序列U的估计值. 这种通信系统的码率为$R = k/n$；错误概率被定义为