4、最佳编码

(1)最佳码定义是什么?

凡是能载荷一定的信息量，且码字的平均长度最短，可分离的变长码的码字集合都可称为最佳码。

(2)最佳编码思想是什么?

将概率大的信息符号编以短的码字，概率小的符号编以长的码字，使得平均码字长度最短。

(3)最佳码的编码主要方法有哪些?

香农（Shannon）、费诺（Fano）、哈夫曼（Huffman）编码等。

信源编码有以下3种主要方法：

- (1) **匹配编码**䷀根据信源符号的概率不同，编码的码长不同：概率大的信源符号，所编的代码短；概率小的信源符号所编的代码长，这样使平均码长最短。将要讲述的香农编码、哈夫曼编码、费诺码都是概率匹配编码，都是无失真信源编码。
- (2) **变换编码**　先对信号进行变换，从一种信号空间变换为另一种信号空间，然后针对变换后的信号进行编码。
- (3) **识别编码**䷀识别编码主要用于印刷或打字机等有标准形状的文字符号和数据的编码，比如中文文字和语音的识别。䷀

第四节 变长码的编码方法

- 香农编码
- 香农-费诺-埃利斯编码
- 费诺编码方法
- 霍（哈）夫曼编码方法
- r元霍夫曼编码

1、香农编码

- 香农码的方法是选择每个码字长度 满足

$$l_i = \left\lceil \log \frac{1}{p(s_i)} \right\rceil \qquad i = 1, 2, \cdots, q$$

按照香农编码方法构造的码，其平均码长不超过上界，即 $\bar{L} \le H_r(S) + 1$

- 编码方法如下：

(1)将信源消息符号按其出现的概率大小依次排列

$$p(x_1) \geq p(x_2) \geq \cdots \geq p(x_n)$$

(2) 确定满足下列不等式的整数码长K_i：

$$-\log_2 p(x_i) \leq K_i < -\log_2 p(x_i) + 1$$

(3)为了编成唯一可译码，计算第i个消息的累加概率

$$P_i = \sum_{k=1}^{i-1} p(x_k)$$

(4)将累加概率P_i 变换成二进制数。

(5)取P_i二进数的小数点后K_i位即为该消息符号的二进制码字。

- 十进制小数转换为二进制小数

十进制小数转换成二进制小数采用"乘2取整，顺序排列法。具体做法是：用2乘十进制小数，可以得到积，将积的整数部分取出，再用2乘余下的小数部分，又得到一个积，再将积的整数部分取出，如此进行，直到积中的小数部分为零，或者达到所要求的精度为止。 然后把取出的整数部分按顺序排列起来，先取的整数作为二进制小数的高位有效位，后取的整数作为低位有效位。

表 5.2.1 二进制香农编码

x_i	$p(x_i)$	$p_a(x_j)$	k_i	码字
x_1	0.25	0.000	2	$00(0.000)_2$
x_2	0.25	0.250	2	$01(0.010)_2$
x_3	0.20	0.500	3	$100(0.100)_2$
x_4	0.15	0.700	3	$101(0.101)_2$
x_5	0.10	0.850	4	$1101(0.1101)_2$
x_6	0.05	0.950	5	$11110(0.11110)_2$

x_i	$P(x_i)$	P_i	$-\log p_2(x_i)$	K_i	码字
x_1	0.20	0	2.32	3	000
x_2	0.19	0.2	2.41	3	001
x_3	0.18	0.39	2.48	3	011
x_4	0.17	0.57	2.56	3	100
x_5	0.15	0.74	2.74	3	101
x_6	0.10	0.89	3.32	4	1110
x_7	0.01	0.99	6.64	7	1111110

2、香农-费诺-埃利斯编码

- 将香农编码中的累加概率换成修正累加概率即可得到相应的香农-费诺-埃利斯编码：
- 琊蕌
- (1) 计算出各个信源符号的修正累加概率

$$\bar{F}(s_i)=\sum_{k=1}^{i-1}p(s_k)+\frac{1}{2}p(s_i)$$

- (2) 按下式计算第i个消息的二元代码组的码长l_i

$$l_i=\left\lceil \log\frac{1}{p(s_i)}\right\rceil+1$$

- (3) 将累加概率$\bar{F}(s_i)$(十进制小数)变换成二进制小数．根据码长l_i取小数点后l_i个二进制符号作为第i个消息的码字．

x_i	$P(x_i)$	F_i	$-\log p_2(x_i)$	K_i	二进制	码字
x_1	0.10	0.05	3.34	5	00001	00001
x_2	0.19	0.195	2.41	4	00110	0011
x_3	0.15	0.365	2.74	4	01011	0101
x_4	0.17	0.525	2.56	4	10000	1000
x_5	0.18	0.7	2.48	4	10110	1011
x_6	0.20	0.89	2.34	4	11100	1110
x_7	0.01	0.995	6.64	8	11111110	11111110

3、费诺编码方法

- **编码步骤：**

（1）将信源消息符号按其出现的概率大小依次排列：

$p(x_1) \geq p(x_2) \geq \ldots \geq p(x_n)$。

（2）将依次排列的信源符号按概率值分为两大组，使两个组的概率之和近于相同，并对各组赋予一个二进制码元“0”和“1”。

（3）将每一大组的信源符号进一步再分成两组，使划分后的两个组的概率之和近于相同，并又赋予两个组一个二进制符号“0’和“1”。

（4）如此重复，直至每个组只剩下一个信源符号为止。

（5）信源符号所对应的码字即为费诺码。

表 5.14　费诺编码

信源符号	概率	第 1 次分组	第 2 次分组	第 3 次分组	第 4 次分组	码字	码长
s_1	0.2	0	0			00	2
s_2	0.19		1	0		010	3
s_3	0.18			1		011	3
s_4	0.17	1	0			10	2
s_5	0.15		1	0		110	3
s_6	0.10			1	0	1110	4
s_7	0.01				1	1111	4

该码的平均码长为

$$\overline{L}=\sum_{i=1}^{7}p(s_i)l_i$$
$$=0.20\times 2+0.19\times 3+0.18\times 3+0.17\times 2+0.15\times 3+0.10\times 4+0.01\times 4$$
$$=2.74\ \text{码符号/信源符号}$$

信息传输率为

$$R=\frac{H(S)}{\overline{L}}=\frac{2.61}{2.74}=0.953\ \text{比特码符号}$$

4、哈夫曼编码方法

- 哈夫曼(***Huffman***)编码是一种效率比较高的变长无失真信源编码方法。

- **哈夫曼编码 步骤：**

（1）将n个信源消息符号按其出现的概率大小依 次排列，

$$p(x_1) \geq p(x_2) \geq \ldots \geq p(x_n)$$

（2）取两个概率最小的字母分别配以0和1两码元，并将这两个概率相加作为一个新字母的概率，与未分配的二进符号的字母重新排队。

（3）对重排后的两个概率最小符号重复步骤（2）的过程。

（4）不断继续上述过程，直到最后两个符号配以0和1为止。

（5）从最后一级开始，向前返回得到各个信源符号所对应的码元序列，即相应的码字。

哈夫曼编码练习

表 5.11　霍夫曼编码

信源符号s_i	概率$p(s_i)$	编码过程					码字 w_i	码长 l_i
		S_1	S_2	S_3	S_4	S_5		
						0.61		
					0.39	0.39		
				0.35	0.35			
			0.26	0.26	0.26			
s_1	0.20	0.20	0.20	0.20			10	2
s_2	0.19	0.19	0.19	0.19			11	2
s_3	0.18	0.18	0.18				010	3
s_4	0.17	0.17	0.17				011	3
s_5	0.15	0.15					000	3
s_6	0.10	0.11					0010	4
s_7	0.01						0011	4

该哈夫曼码的平均码长为

$$\bar{K}=\sum_{i=1}^{7} p(x_i)K_i=2.72\text{码元/符号}$$

信息传输速率

$$R=\frac{H(X)}{\bar{K}}=\frac{2.61}{2.72}=0.9596\text{比特/码元}$$

由此可见，哈夫曼码的平均码长最小，消息传输速率最大，编码效率最高。

例[5.4.2] 单符号离散无记忆信源 $\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{Bmatrix} x_1, & x_2, & x_3, & x_4, & x_5, \\ 0.4 & 0.2 & 0.2 & 0.1 & 0.1 \end{Bmatrix}$ ，用两种不同的方法对其编二进制哈夫曼码。

- **方法一**：合并后的新符号排在其它相同概率符号的后面。

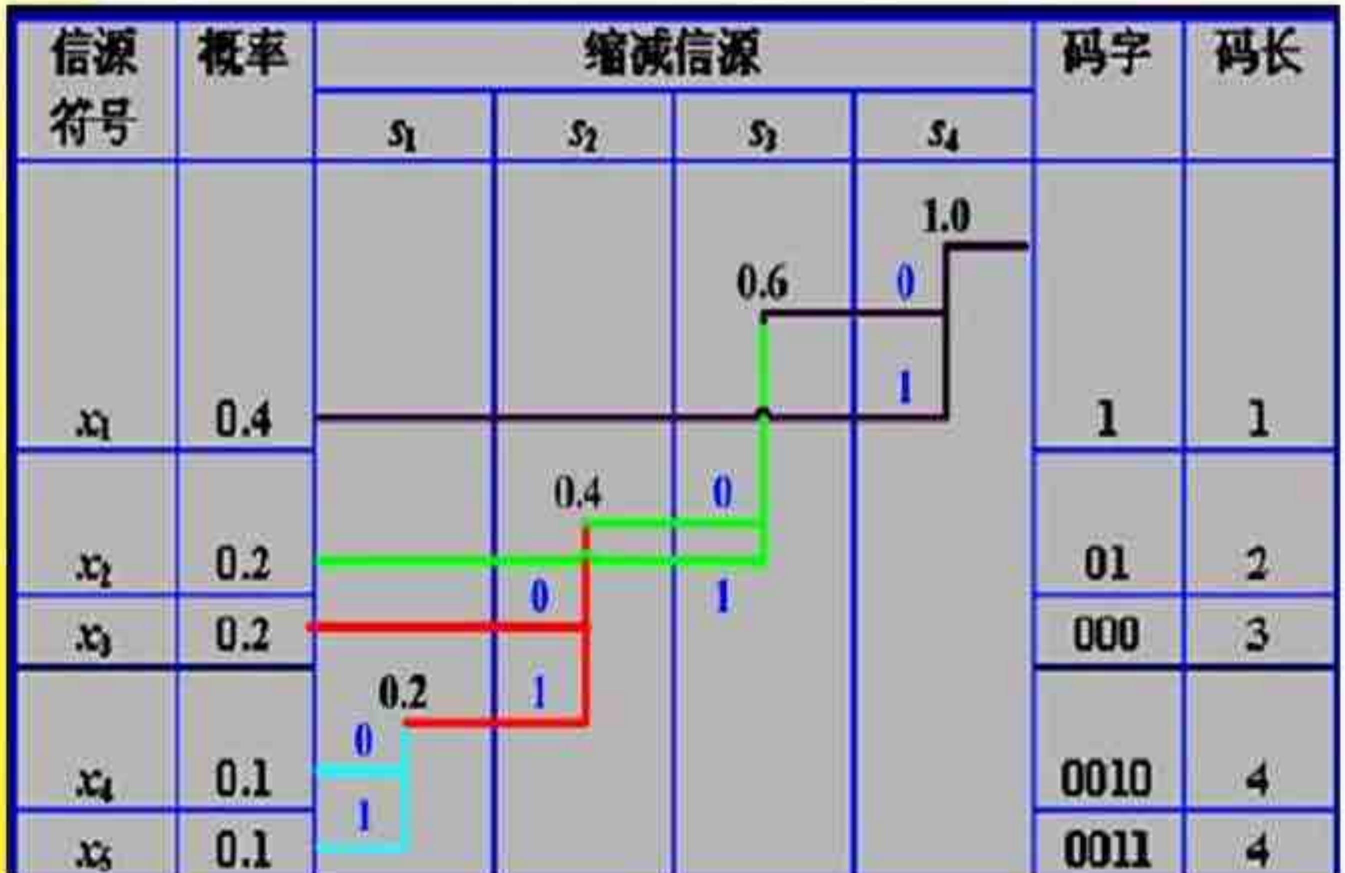

图5.4.3 二进制哈夫曼编码（编法一）

- **方法二**：合并后的新符号排在其它相同概率符号的前面。

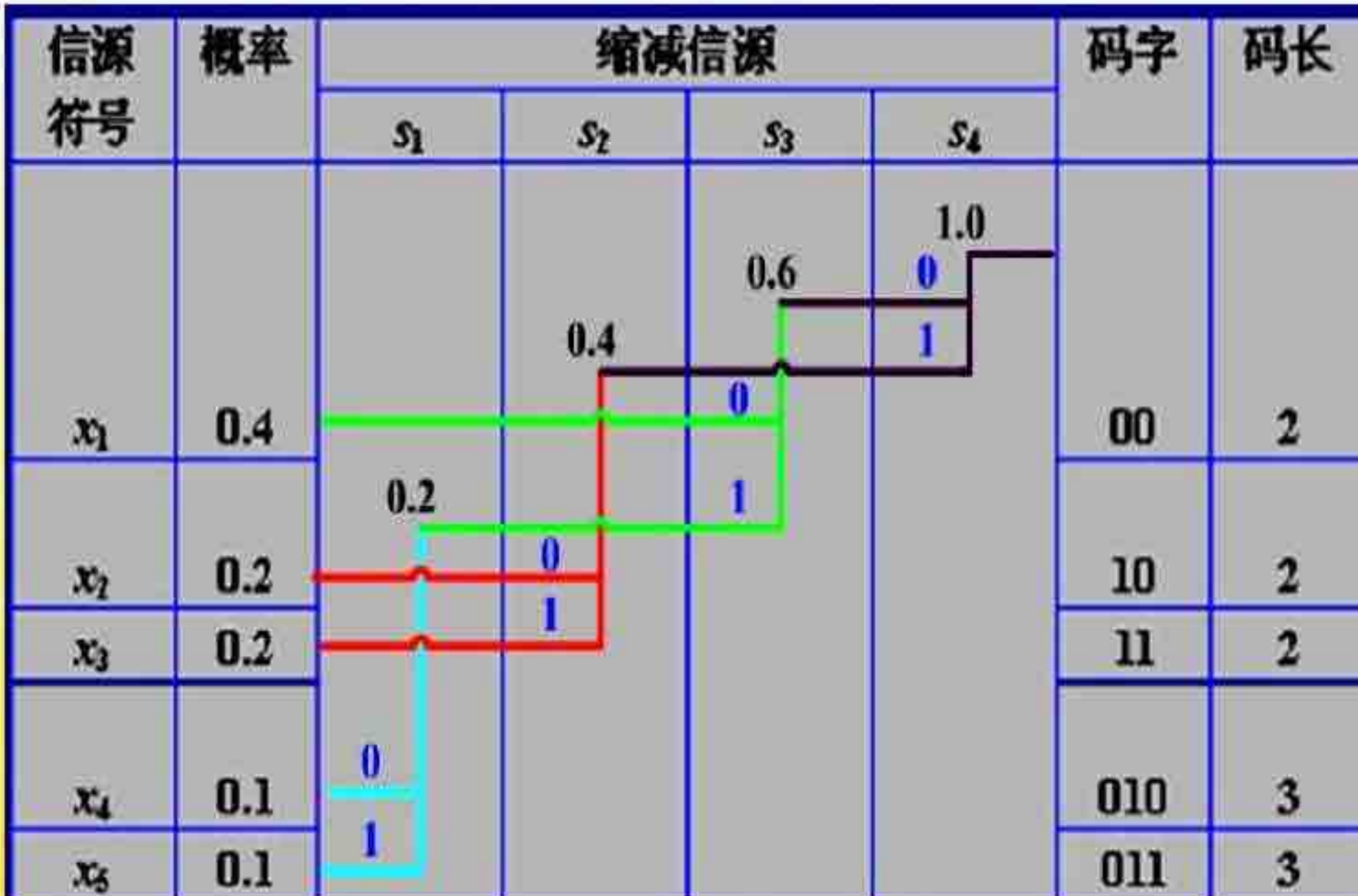

图 5.4.5 二进制哈夫曼编码（编法二）

- 单符号信源编二进制哈夫曼码，编码效率主要决定于信源熵和平均码长之比。对相同的信源编码，其熵是一样的，采用不同的编法，得到的平均码长可能不同。平均码长越短，编码效率就越高。
- 编法一的平均码长为

$$\overline{K}_1=\sum_{i=1}^{5}p(x_i)k_i=0.4\times1+0.2\times2+0.2\times3+(0.1+0.1)\times4=2.2(\text{比特/符号})$$

- 编法二的平均码长为

$$\overline{K}_2=\sum_{i=1}^{5}p(x_i)k_i=(0.4+0.2+0.2)\times2+(0.1+0.1)\times3=2.2(\text{比特/符号})$$

- 可见 $\overline{K}_1=\overline{K_2}$，本例两种编法的平均码长相同，所以编码效率相同。

讨论：哪种方法更好？

- 定义码字长度的方差σ²：长度k_i与平均码长 $\overline{K}$ 之差的平方的数学期望，即

$$\sigma^2 = E[(k_i - \overline{K})^2] = \sum_{i=1}^{n} p(x_i)(k_i - \overline{K})^2$$

- 编法一码字长度方差：

$$\sigma_1^2 = 0.4(1-2.2)^2 + 0.2(2-2.2)^2 + 0.2(3-2.2)^2 + (0.1+0.1)(4-2.2)^2 = 1.36$$

- 编法二码字长度方差：

$$\sigma_2^2 = (0.4+0.2+0.2)(2-2.2)^2 + (0.1+0.1)(3-2.2)^2 = 0.16$$

- 可见：第二种编码方法的码长方差要小许多。意味着第二种编码方法的码长变化较小，比较接近于平均码长。
  - 第一种方法编出的5个码字有4种不同的码长；
  - 第二种方法编出的码长只有两种不同的码长；
  - 显然，第二种编码方法更简单、更容易实现，所以更好。

结论：在哈夫曼编码过程中，对缩减信源符号按概率由大到小的顺序重新排列时，应**使合并后的新符号尽可能在靠前的位置**，这样可使合并后的新符号重复编码次减少，使短码得到充分利用。

哈夫曼编码的特点

哈夫曼码是用概率匹配方法进行信源编码。

它有两个明显特点：

◆ 一是哈夫曼码的编码方法保证了概率大的符号对应于短码，概率小的符号对应于长码，充分利用了短码；

◆ 二是缩减信源的最后二个码字总是最后一位不同，从而保证了哈夫曼码是即时码

这两个特点，保证了哈夫曼码是最佳码

5、r元霍夫曼编码

二进制霍夫曼码的编码方法很容易推广到 r 进制的情况，只是编码过程中构成缩减信源时，每次都是将 r 个概率最小的信源符号合并

为了充分利用短码，使霍夫曼码的平均码长最短，必须使最后一个缩减信源恰好有 r 个信源符号. 因此对于 r 元霍夫曼编码，信源 S 符号个数 q 必须满足 $q=(r-1)\theta+r$. $\boxed{\theta}$ 表示信源缩减次数. 如果不满足上式，则可以在最后增补一些概率为 0 的信源符号

例：对如下单符号离散无记忆信源编三进制哈夫曼码。

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{Bmatrix} x_1, & x_2, & x_3, & x_4, & x_5, & x_6 & x_7 & x_8 \\ 0.4 & 0.18 & 0.1 & 0.1 & 0.07 & 0.06 & 0.05 & 0.04 \end{Bmatrix}$$

对香农码、费诺码、哈夫曼码特点进行归纳

- 香农码、费诺码、哈夫曼码都考虑了信源的统计特性，使经常出现的信源符号对应较短的码字，使信源的平均码长缩短，从而实现了对信源的压缩；

香农编码方法特点：

- 由于b_i总是进一取整，香农编码方法不一定是最佳的；
- 由于第一个消息符号的累加概率总是为0，故它对应码字总是0、00、000、0…0的式样；
- 码字集合是唯一的，且为即时码；
- 先有码长再有码字；
- 对于一些信源，编码效率不高，冗余度稍大，因此其实用性受到较大限制。

费诺编码特点：

- 概率大，则分解的次数小；概率小，则分解的次数多。这符合最佳编码原则。
- 码字集合是唯一的。
- 分解完了，码字出来了，码长也有了。
- 因此，费诺编码方法又称为子集分解法。
- 费诺编码方法比较适合于每次分组概率都很接近的信源，特别是对每次分组概率都相等的信源进行编码时，可达到理想的编码效率。

- **哈夫曼编码特点：**
- 由于哈夫曼编码总是以最小概率相加的方法来“缩减”参与排队的概率个数，因此概率越小，对缩减的贡献越大，其对于消息的码字也越长；
- 最小概率相加的方法使得编码不具有唯一性，尤其是碰到存在几个消息符号有着相同概率的情况，将会有多种路径选择，亦即具有多种可能的代码组集合；
- 尽管对同一信源存在着多种结果的哈夫曼编码，但它们的平均码长几乎都是相等的，因为每一种路径选择都是使用最小概率相加的方法，其实质都是遵循最佳编码的原则，因此哈夫曼编码是最佳编码。
- 哈夫曼编码是一种最佳编码，实现也不困难，因此到目前为止它仍是应用最为广泛的无失真信源编码之一。

总 结

- 香农码、费诺码、哈夫曼码都考虑了信源的统计特性，使经常出现的信源符号对应较短的码字，使信源的平均码长缩短，从而实现了对信源的压缩；
  - 香农码有系统的、惟一的编码方法，但在很多情况下编码效率不是很高；
  - 费诺码和哈夫曼码的编码方法都不惟一；
  - 费诺码比较适合于对分组概率相等或接近的信源编码，费诺码也可以编m进制码，但m越大，信源的符号数越多，可能的编码方案就多，编码过程就越复杂，有时短码未必能得到充分利用；
  - 哈夫曼码对信源的统计特性没有特殊要求，编码效率比较高，对编码设备的要求也比较简单，因此综合性能优于香农码和费诺码。