

# 实验三 香农编码仿真实现

## 一、 实验目的

1. 理解信源编码的意义；
2. 掌握香农编码的过程；
3. 对给定的信源使用 MATLAB 进行香农编码。

## 二、 实验原理

### 1. 香农第一定理及相关概念

香农第一定理指出了平均码长与离散信源概率之间的关系，同时也指出了可以通过编码使平均码长达到极限值，这是一个很重要的极限定理。根据香农第一定理，当选择每个码字的长度 $l_i$ 为满足下式的一个整数时，这种编码方法就是香农编码。

$$I(x_i) \leq l_i \leq I(x_i) + 1$$

其中 $I(x_i)$ 表示信源符号 $x_i$ 的信息量，即， $I(x_i) = -\log_2 p(x_i)$ 。

经过香农编码之后，码字的平均码长 $\bar{L}$ 为：

$$\bar{L} = \sum_{i=1}^n p(x_i) * l_i$$

其中， $p(x_i)$ 表示信源符号 $x_i$ 的统计概率， $l_i$ 表示信源符号 $x_i$ 经过香农编码得到的码字的长度。

香农编码的效率为：

$$\eta = \frac{\text{信源熵}}{\text{平均码长}} = \frac{H(X)}{\bar{L}}$$

### 2. 香农编码流程

香农编码严格意义上来说不是最佳码，它是采用信源符号的累计概率分布函数来分配码字。

其编码具体步骤如下：

- (1) 将  $n$  个信源符号按概率从大到小顺序进行排列，即：

$$p_1 \geq p_2 \geq \dots \geq p_n$$

- (2) 根据不等式 $-\log_2(p(x_i)) \leq l_i \leq -\log_2(p(x_i)) + 1$ ，计算信源符号 $x_i$ 对应的码字的码长（ $l_i$ 取在此范围内的整数），其中， $p(x_i)$ 表示信源 $x_i$ 的

统计概率。

(3) 根据下式计算排序后信源符号 $x_i$ 的累加概率 $P_i$ :

$$\begin{cases} P_1 = 0 \\ P_{i+1} = \sum_{j=1}^i p_j \quad i = 1, 2, \dots, n-1 \end{cases}$$

(4) 将累加概率 $P_i$ 转换成二进制小数, 取小数点后的 $l_i$ 位二进制数作为信源 $x_i$ 的码字。

香农编码的效率不高, 实用性不大, 但对其他编码方法有很好的理论指导意义。一般情况下, 按照香农编码方法编出来的码, 其平均码长不是最短的, 即不是紧致码(最佳码)。只有当信源符号的概率分布使(2)中不等式左边的等号成立时, 编码效率才能达到最高。

### 3. 香农编码示例分析

有一离散信源 X, 其概率分布如下:

$$\begin{bmatrix} X \\ p(X) \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ 0.25 & 0.15 & 0.20 & 0.05 & 0.10 & 0.25 \end{bmatrix}$$

则对该信源进行二进制香农码的编码过程如下表所示:

信源符号 $x_i$	符号概率	累加概率 $P_i$	$-\log p(x_i)$	码长 $l_i$	码字
$x_1$	$p_1 = p(x_1) = 0.25$	0	2.00	2	00
$x_6$	$p_2 = p(x_6) = 0.25$	0.25	2.00	2	01
$x_3$	$p_3 = p(x_3) = 0.20$	0.50	2.32	3	100
$x_2$	$p_4 = p(x_2) = 0.15$	0.70	2.74	3	101
$x_5$	$p_5 = p(x_5) = 0.10$	0.85	3.32	4	1101
$x_4$	$p_6 = p(x_4) = 0.05$	0.95	4.32	5	11110

以 $i = 2$ 为例,  $-\log_2 0.15 \leq l_2 < -\log_2 0.15 + 1$ , 即 $2.74 \leq l_2 < 3.74$ , 因此,  $l_2 = 3$ , 累加概率 $P_2$ 为 0.7, 变成二进制数前三位为 0.101。编程时可采用如下方法得到二进制数: 用排序后信源符号 $x_i$ 的累加概率 $P_i$ 乘以 2 后再次赋值给 $P_i$ , 如果整数部分有进位, 则 $P_i$ 转化为二进制后的小数点后第一位为 1, 否则为 0, 将新的 $P_i$ 的小数部分再次乘以 2, 同样, 如果整数部分有进位, 则转化为二进制后的小数点后第二位为 1, 否则为 0, 从而得到小数点后的第二位, 依此类推, 直到得到了满足要求的位数, 或者没有小数部分了为止, 最终即可获得所有信源符号对应的码字。

同时, 我们可以通过相应的公式计算得出如下信息:

信源信息熵:  $H(X) = -\sum_{i=1}^6 p(x_i) \log_2 p(x_i) = 2.4232 \text{ bits}$

平均码长:  $\bar{L} = \sum_{i=1}^6 p(x_i) l_i = 2.7 \text{ bits}$

编码效率:  $\eta = H(X)/\bar{L} = 0.8975$

#### 4. 实验可能用到的 Matlab Function（仅供参考）

sort(), ceil(), floor(), max(), disp()等。

### 三、 实验预习

回答以下问题

1. 给定一组离散信源概率分布如下：

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 \\ 0.20 & 0.10 & 0.30 & 0.15 & 0.25 \end{bmatrix}$$

模仿实验原理部分香农编码示例分析中的香农编码过程，以表格的形式绘制上述离散信源的香农编码过程；

2. 计算第 1 题中经过香农编码后的码字的平均码长 $\bar{L}$ 及编码效率 $\eta$ （请写出计算公式和结果）。

### 四、 实验内容（要求给出结果截图，源代码放在.m 文件中）

1. 按照实验原理中香农编码的步骤，用 MATLAB 软件编写代码实现对任意给定的离散信源进行香农编码。要求最终在控制台依次输出码字，平均码长和编码效率，并和实验预习中计算的结果进行对比。

### 五、 实验思考题

1. 当信源符号的概率分布使不等式 $-\log_2(p(x_i)) \leq l_i \leq -\log_2(p(x_i)) + 1$ 左边的等号对每一个离散信源都成立时，香农编码效率会如何？
2. 请举出两个信源概率分布的例子，使得香农编码的效率达到最大值。

# 实验四 哈夫曼编码仿真实现

## 一、 实验目的

1. 掌握哈夫曼编码的原理及编码步骤；
2. 练习 MATLAB 中哈夫曼编码函数的调用；
3. 编程实现哈夫曼编码的过程。（扩展实验）

## 二、 实验原理

### 1. 哈夫曼编码相关原理

通信的根本问题是如何将信源输出的信息在接收端的信宿精确或近似的复制出来。为了有效地复制信号，就通过对信源进行编码，使通信系统与信源的统计特性相匹配。若接收端要求无失真地精确地复制信源输出的信息，这样的信源编码即为无失真编码。即使对于一个小的时间段内，连续信源输出的信息量也可以是无限大的，所以对其是无法实现无失真编码的；而离散信源输出的信息量却可以看成是有限的，所以只有离散信源才可能实现无失真编码。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以称为最佳码。为此必须将概率大的信息符号编为短的码字，概率小的符号编为长的码字，使得平均码字长度最短。

变字长编码的最佳编码定理：在变字长码中，对于概率大的信息符号编为短字长的码；对于概率小的信息符号编为长字长的码。如果码字长度严格按照符号概率的大小顺序排列，则平均码字长度一定小于以任何顺序排列方式得到的码字长度。

哈夫曼编码就是利用了这个定理，根据信源符号的概率分布，采用不等长编码。概率大的符号，使用短的码字编码；概率小的符号，使用长的码字编码。哈夫曼编码把信源按概率大小顺序排列，并设法按逆次序分配码字的长度。在分配码字的长度时，首先将出现概率最小的两个符号相加，合成一个概率；第二步把这个合成的概率看成是一个新组合符号的概率，重复上述做法，直到最后只剩下两个符号的概率为止。完成以上概率相加顺序排列后，再反过来逐步向前进行编码。每一步有两个分支，各赋予一个二进制码，可以对概率大的编为 0 码，概率小的编为 1 码。反之亦然。

经过哈夫曼编码之后，码字的平均码长 $\bar{L}$ 为：

$$\bar{L} = \sum_{i=1}^n p(x_i) * l_i$$

其中， $l_i$ 表示信源符号 $x_i$ 经过哈夫曼编码得到的码字的长度。

哈夫曼编码的效率为：

$$\eta = \frac{\text{信源熵}}{\text{平均码长}} = \frac{H(X)}{\bar{L}}$$

## 2. 哈夫曼编码流程

哈夫曼编码的具体步骤归纳如下：

- (1) 统计  $n$  个信源符号，得到  $n$  个不同概率的信息符号；
- (2) 将这  $n$  个信源符号按其概率从大到小依次排序；
- (3) 取两个概率最小的信息符号分别配以 1 和 0 两个码元，并将这两个概率相加作为一个新的信息符号概率，和未分配的信息符号构成新的信息符号序列；
- (4) 将剩余的信息符号，按概率从大到小重新进行排序；
- (5) 重复步骤(3)，将排序后的最小的两个概率相加，相加和与其他概率再排序；
- (6) 如此反复重复  $n-2$  次，最后只剩下两个概率值；
- (7) 依次从每个符号概率的第一级开始，向后寻找其在下一级的位置，按顺序记录其在每一级被分配的码元（即 0 或 1，若没有被分配码元，则不记录），直到寻找到最后一级结束，然后将记录的码元逆序排列，即可得到哈夫曼编码码字。

## 3. 哈夫曼编码示例分析

设信源共有 7 个离散符号消息，其概率如下表所示：

信源符号 $x_i$	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$
符号概率 $p(x_i)$	0.15	0.19	0.10	0.17	0.01	0.18	0.20

对信源符号进行哈夫曼编码的过程如下图所示：

信源符号 $x_i$	概率 $p(x_i)$	编 码 过 程	码 字	码长 $l_i$
$x_7$	0.20		10	2
$x_2$	0.19		11	2
$x_6$	0.18		000	3
$x_4$	0.17		001	3
$x_1$	0.15		010	3
$x_3$	0.10		0110	4
$x_5$	0.01		0111	4

以信源符号 $x_5$ 为例，根据上图哈夫曼编码流程可以看出，其在第一级概率为0.01，被分配的码元为1；其在第二级概率为0.11，被分配的码元为1；其在第三级概率为0.26，没有被分配码元；其在第四级概率为0.26，没有被分配码元；其在第五级概率为0.26，被分配的码元为1；其在第六级概率为0.61，被分配的码元为0。得到按顺序记录的码元为1110，将其逆序排列为0111，因此，信源符号 $x_5$ 对应的哈夫曼编码即为0111。

通过计算，还可以得到如下信息：

信源符号的信息熵为： $H(X) = -\sum_{i=1}^7 p(x_i) \log_2 p(x_i) = 2.6087 \text{ bits}$

该哈夫曼码的平均码长为： $\bar{L} = \sum_{i=1}^7 p(x_i) l_i = 2.72 \text{ bits}$

编码效率为： $\eta = H(X)/\bar{L} = 2.6087/2.72 = 0.95907$

#### 4. MATLAB 中 huffmandict 函数简单介绍

huffmandict 函数为已知概率分布的信源模型生成哈夫曼编解码索引表。

调用方法如下：

`[dict, avglen] = huffman (symbols, p)`

symbols 表示需要编码的离散信源符号，在这里可以用N个数字组成的一维数组代替 symbols，p 为离散信源符号的概率分布，也是一个 $1 \times N$ 的一维数组。

dict 为函数的一个返回值，是一个 $N \times 2$ 的矩阵，其中第一列存放了输入的离散信源符号，第二列存放了每个信源符号对应的哈夫曼编码的码字。avglen 为函数的另一个返回值，表示码字的平均长度。

#### 5. 实验可能用到的 Matlab Function 或数据类型（仅供参考）

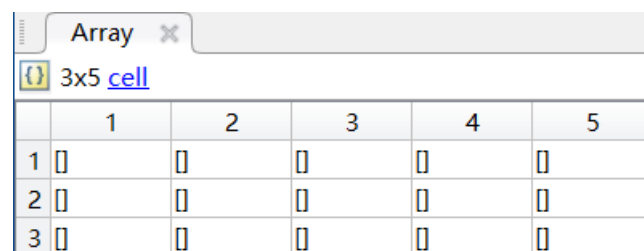
huffmandict(), find(), 元胞数组（cell）等。

下面用一个示例对元胞数组（cell）进行简单的介绍：

(1) 初始化一个大小为 $3 \times 5$ 的元胞数组 Array；

`Array = cell(3,5);`

运行结果如下图所示：



	1	2	3	4	5
1	[]	[]	[]	[]	[]
2	[]	[]	[]	[]	[]
3	[]	[]	[]	[]	[]

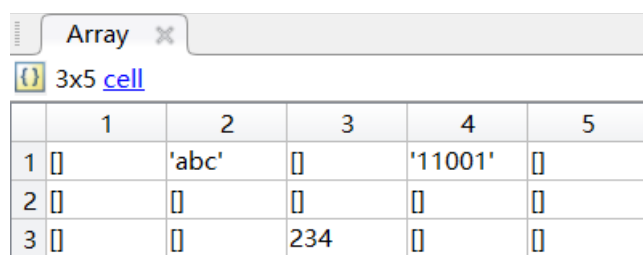
(2) 在 Array 中坐标分别为(1,2)，(1,4)和(3,3)的位置存入字符串“abc”，字符串“11001”和数字 234；

`Array{1,2} = 'abc';`

`Array{1,4} = '11001';`

`Array{3,3} = 234;`

运行结果如下图所示：



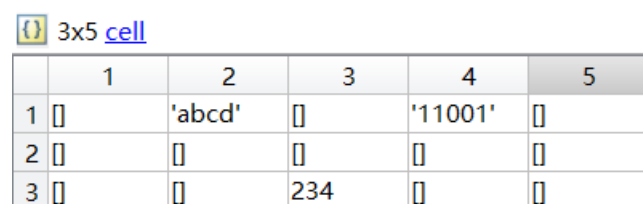
	1	2	3	4	5
1		'abc'		'11001'	
2					
3			234		

可以发现，元胞数组 `Array` 中的每个位置都可以存放不同类型的数据，且存入的字符串长度可以不相同。

(3) 在 `Array` 中坐标为(1,2)的位置中的字符串后面追加字符 ‘d’ ；

`Array{1,2} = [Array{1,2}, 'd'];`

运行结果如下图所示：



	1	2	3	4	5
1		'abcd'		'11001'	
2					
3			234		

(4) 分别读取元胞数组 `Array` 中坐标为(1,4)和(2,3)位置的数据，运行结果如下图所示：

```
>> disp(Array{1,4});  
11001  
>> disp(Array{2,3});  
>>
```

### 三、 实验预习

回答以下问题

1. 给定一组离散信源概率分布如下：

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ 0.21 & 0.10 & 0.30 & 0.09 & 0.25 & 0.05 \end{bmatrix}$$

计算上述离散信源的哈夫曼编码（概率大的编为 0 码，概率小的编为 1 码），写出信源 $x_1$ 至 $x_6$ 经过哈夫曼编码得到的码字。

2. 计算第 1 题中经过哈夫曼编码后的平均码长 $\bar{L}$ 及编码效率 $\eta$ （请写出计算公式和结果）。

#### 四、 实验内容（要求给出结果截图，源代码放在.m文件中）

1. 使用 MATLAB 软件自带的系统函数 `huffmandict` 编写代码实现对任意给定的离散信源进行哈夫曼编码。要求最终在控制台依次输出码字，平均码长和编码效率，并和实验预习中计算的结果进行对比；
2. 按照实验原理中给定的哈夫曼编码的流程，使用 MATLAB 软件编写代码实现对任意给定的离散信源进行哈夫曼编码。要求最终在控制台依次输出码字，平均码长和编码效率，并和实验预习中计算的结果进行对比。  
（扩展实验）

#### 五、 实验思考题

1. 分别以实验三和实验四中实验预习部分给定的离散信源分布概率作为输入，运行得到二者的香农编码的码字和哈夫曼编码的码字，对比两种编码方式的平均码长与编码效率。
2. 试写出两个信源分布，其哈夫曼编码的效率为 1。