信息论基础

李　莹
liying2009@ecust.edu.cn

第五章：无失真信源编码

一、信源编码的相关概念

二、定长码及定长信源编码定理

三、变长码及变长信源编码定理

四、变长码的编码方法

1. Kraft不等式和McMillan不等式

- Kraft定理：即时码存在的充要条件是 $\sum_{i=1}^{q} r^{-l_i} \le 1$
- McMillan定理：唯一可译码存在的充要条件是 $\sum_{i=1}^{q} r^{-l_i} \le 1$

- 任何一个唯一可译码均可用一个相同码长的即时码来代替。
- 上述定理是存在性定理:
  - ✓ 当满足Kraft（或McMillan）不等式时，必然可以构造出即时码（或唯一可译码），否则不能构造出即时码（或唯一可译码）。
  - ✓ 该定理不能作为判断一种码是否是即时码（或唯一可译码）的判断依据。

信源符号s_i	符号出现的概率	码1	码2	码3	码4
s_1	1/2	0	0	1	1
s_2	1/4	11	10	10	01
s_3	1/8	00	00	100	001
s_4	1/8	11	01	1000	0001

2. 变长唯一可译码判别方法

步骤：

1.构造F_1 ：考察 C中所有码字，如果一个码字是另一个码字的前缀，则将后缀作为F_1 中的元素。

2.构造 $F_n(n>1)$ ：将C 与 F_{n-1}比较。如果 C 中有码字是F_{n-1}中元素的前缀，则将相应的后缀放入F_n 中；同样 F_{n-1} 中若有元素是 C 中码字的前缀，也将相应的后缀放入 F_n 中。

3.检验 F_n ：

1)如果 F_n 是空集，则断定码 C 是唯一可译码，退出循环；

2)反之，如果 F_n 中的某个元素与 C中的某个元素相同，则断定码 C 不是唯一可译码，退出循环。

3)如果上述两个条件都不满足，则返回步骤2。

例5.4：

C	F_1	F_2	F_3	F_4	F_5
a	d	eb	de	b	ad
c	bb	cde			$bcde$
ad					
abb					
bad					
deb					
$bbcde$					

结论：F_5中包含了C中的元素，因此该变长码不是唯一可译码。

问题：判断 C={1,10,100,1000}是否是唯一可译码?

3.紧致码平均码长界限定理

- 平均码长

$$\overline{L} = \sum_{i=1}^{q} p(s_i) l_i$$

码符号/信源符号

- 对于给定的信源和码符号集，若有一个唯一可译码，其平均长度 $\overline{L}$ 小于所有其他唯一可译码，则称这种码为**紧致码**或**最佳码**。

- **紧致码平均码长界限定理：**

设离散无记忆信源的熵为$H(S)$，用r 个码符号进行编码，则总可找到一种无失真信源编码，构成唯一可译码，使其平均码长满足：

$$\frac{H(S)}{\log r} \le \bar{L} < \frac{H(S)}{\log r} + 1$$

定理5.6　紧致码平均码长界限定理

$$\frac{H(S)}{\log r} \le \bar{L} < \frac{H(S)}{\log r} + 1$$

$$H(S) - \bar{L}\log r \le 0$$

$$= -\sum_{i=1}^{q} p(s_i)\log p(s_i) - \sum_{i=1}^{q} p(s_i) l_i \log r$$

$$= -\sum_{i=1}^{q} p(s_i)\log p(s_i) - \sum_{i=1}^{q} p(s_i)\log r^{l_i}$$

$$= -\sum_{i=1}^{q} p(s_i)\log p(s_i) + \sum_{i=1}^{q} p(s_i)\log r^{-l_i}$$

$$= \sum_{i=1}^{q} p(s_i)\log \frac{r^{-l_i}}{p(s_i)}$$

$$\le \log\left[\sum_{i=1}^{q} p(s_i)\frac{r^{-l_i}}{p(s_i)}\right]$$

$$= \log\left(\sum_{i=1}^{q} r^{-l_i}\right) \le \log 1 = 0$$

$$\bar{L} \ge \frac{H(S)}{\log r}$$

平均码长=下限值时

$$p(s_i) = r^{-l_i} \quad (i = 1, 2, \ldots, q)$$

$$l_i = -\log_r p(s_i)$$

$$\frac{H(S)}{\log r} \le \overline{L} < \frac{H(S)}{\log r} + 1$$

$$-\log_r p(s_i) \le l_i < -\log_r p(s_i) + 1$$

$$-p(s_i)\log_r p(s_i) \le p(s_i) l_i < -p(s_i)\log_r p(s_i) + p(s_i)$$

$$\sum_{i=1}^{q} -p(s_i)\log_r p(s_i) \le \sum_{i=1}^{q} p(s_i) l_i < \sum_{i=1}^{q} -p(s_i)\log_r p(s_i) + \sum_{i=1}^{q} p(s_i)$$

$$\frac{H(S)}{\log r} \le \overline{L} < \frac{H(S)}{\log r} + 1$$

4.无失真变长信源编码定理

- 香农第一定理（变长无失真信源编码定理）：

设离散无记忆信源的熵为 $H(S)$，它的N次扩展信源为 S^N，对扩展信源 S^N进行编码。总可以找到一种编码方法，构成唯一可译码，使平均码长满足：

$$\frac{H(S)}{\log r} \leq \frac{\bar{L}_N}{N} < \frac{H(S)}{\log r} + \frac{1}{N}$$

- 当 $N \to \infty$时，有 $\lim_{N\to\infty} \frac{\bar{L}_N}{N} = H_r(S)$

证明:

$$\frac{H(S)}{\log r} \le \bar{L} < \frac{H(S)}{\log r} + 1$$

$$S^N = S_1 S_2 \cdots S_N$$

$$\frac{H(S^N)}{\log r} \le \bar{L}_N < \frac{H(S^N)}{\log r} + 1$$

$$\frac{H(S^N)}{N \log r} \le \frac{\bar{L}_N}{N} < \frac{H(S^N)}{N \log r} + \frac{1}{N}$$

$$\frac{H(S)}{\log r} \le \frac{\bar{L}_N}{N} < \frac{H(S)}{\log r} + \frac{1}{N}$$

$$\lim_{N \to \infty} \bar{L} = \frac{H(S)}{\log r}$$

$$\mathbf{S} = S_1 S_2 \cdots S_N$$

- 把香农第一定理推广到一般离散信源，有

$$\frac{H_\infty}{\log r} \le \frac{\bar{L}_N}{N} < \frac{H_\infty}{\log r} + \frac{1}{N}$$

$$H_\infty \le H(S)$$

并且 $$\lim_{N\to\infty} \frac{\bar{L}_N}{N} = \frac{H_\infty}{\log r}$$

$$R=\frac{H(S)}{\overline{L}}\le\frac{H(S)}{\frac{H(S)}{\log r}}=\log r$$

信息传输率(码率)

$$R=H(X)=\frac{H(S)}{\overline{L}}\quad\frac{bit/\text{信源符号}}{\text{码符号/信源符号}}=bit/\text{码符号}$$

编码效率

$$\eta=\frac{H_r(S)}{\overline{L}}\quad\frac{r\text{进制单位/信源符号}}{\text{码符号/信源符号}}$$

有效性 $\leftrightarrow \overline{L}$

例5.5　设离散无记忆信源$\begin{bmatrix} X \\ P \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \\ \frac{3}{4} & \frac{1}{4} \end{bmatrix}$，

求 R、η及扩展信源的R、η 。

解：$H(X) = \frac{1}{4}\log_2 4 + \frac{3}{4}\log_2 \frac{4}{3} = 0.81$　比特/符号

假定信源序列的长度为L=1,也用二元编码，其即时码如下表所示。

符号	符号概率	即时码
x_1	3/4	0
x_2	1/4	10

编码效率 $\eta_1 = \frac{0.811}{1.25} \times 100\% = 64.88\%$

输出的信息率为

$R_1 = 0.6488$ 比特 / 码元

假定信源序列的长度为L=2,也用二元编码，其即时码如下表所示。

序列	序列概率	即时码
x_1x_1	9/16	0
x_1x_2	3/16	10
x_2x_1	3/16	110
x_2x_2	1/16	111

这个码的码字平均长度

$$\overline{K_2}=\frac{9}{16}\times1+\frac{3}{16}\times2+\frac{3}{16}\times3+\frac{1}{16}\times3=\frac{27}{16}$$ 二元码符号 / 信源序列

单个符号的平均码长

$$\overline{K}=\frac{\overline{K_2}}{2}=\frac{27}{32}$$ 码元符号/符号

编码效率 $$\eta_2=\frac{32\times0.811}{27}\times100\%=96.1\%$$

输出的信息率为

$\mathbf{R_2=0.\ 961}$ 比特 / 码元符号

将信源序列的长度增加，L＝3或L=4，对这些信源序列X进行编码，并求出其编码效率为

$$\eta_3 = 98.5\%$$

$$\eta_4 = 99.1\%$$

信息传输率分别为：

R_3＝0．985比特／码元符号

R_4＝0．991比特／码元符号