编码基础学习

黄冬勃

2015 年 3 月 15 日

目录

1 无失真信源编码

一般而言，提高通信的有效性是以降低通信的可靠性为代价；反之，提高通信的可靠性以降低通信的有效性为代价。

1.1 单义可译码

无失真信源编码必须符合两个条件：

- 信源编码编出的每一个码字 $\omega_i(i=1,2,...,q)$，与信源 S 发出的每一种不同的符号 $s_i(i=1,2,...q)$ 一一对应；
- 每一种由 N 个信源符号组成的信源符号序列（消息），与每一种由相应的 N 个码字组成的码字序列一一对应。

当信源符号与信源码字不符合一一对应时，为奇异码

单义可译码 $\rightarrow$ 非奇异码

等长非奇异码一定是单义可译码。

1.2 非延长码

无需参考后续码符号就能即时做出译码判断的码，称为即时码。由于即时码是当即可译码，所以总是希望能编出这种即时码。从结构的角度，即时码又称为非延长码。

非延长码是单义可译码的一类子码。非延长码一定单义可译。反之，单义可译码不一定都是非延长码。信源符号 q、码符号集的码符号数 r 以及 q 个码字的码长 n_i 这三种结构参数之间，一定存在某种约束关系。

1.3 单义可译定理

非延长码的构成过程给出一个重要启示，信源编码是否具有单义可译性，与信源的信源符号数 q、码符号集 r、码字长度 $n_i(i=1,2,\cdots,q)$ 这三个编码结构参数密切相关。

定理 1.1. **设设信源 S 的符号集 $S:s_1,s_2,\cdots,s_q$；码符号集 $X:a_1,a_2,\cdots,a_r$；q 个码字长度分别为 $n_1,n_2,\cdots,n_q$。则存在单义可译码的充分必要条件是 $q,r,n_i(i=1,2,\cdots,q)$ 满足克拉夫 (Kraft) 不等式**

$$\sum_{i=1}^{q} r^{-n_i} \leqslant 1$$

1.4 平均码长与码率

无失真信源编码的单义可译问题，只是一个与结构参数 $q,r,n(i=i,2,\cdots,q)$ 有关的结构性问题，与信源符号的统计特性无关。单义可译码的 q 个码字 $w_i(i=1,2,\cdots,q)$ 之间的搭配不收任何条件的约束。单义可译是信源编码的最起码要求。对于通信工程来说，不仅要求信源编码无失真，而且要求信源编码是有效的。编码的有效性是通过平均码长和码率来定夺的。

设信源 S 的信源空间为

$$[S\cdot P]:\begin{cases} S: & s_1 & s_2 & \cdots & s_q \\ p(S): & p(s_1) & P(s_2) & \cdots & p(s_q) \end{cases}$$

且有

$$\sum_{i=1}^{q} p(s_i)=1 \tag{1.1}$$

每个信源符号所需的平局码符号数，就应该等于 q 个码字长度 $n_i(i=1,2,\cdots,q)$ 在信源 S 的概率空间 **P:**$p(s_1),p(s_2),\cdots,p(s_q)$ 中的统计平均值，即

$$\bar{n}=n_1p(s_1)+n_2p(s_2)+\cdots+n_qp(s_q)=\sum_{i=1}^{q} p(s_i)n_i \qquad \text{(码符号/信源符号)} \tag{1.2}$$

这个统计平均值 $\bar{n}$(码符号/信源符号) 称为单义可译码 $W:w_1,w_2,\cdots,w_q$ 的平均码长。

又已知信源 S 的信息熵：

$$H(S)=-\sum_{i=1}^{q} p(s_i)logp(s_i) \qquad \text{(比特/信源符号)} \tag{1.3}$$

是固定不变的。

由 eq. (1.2) 和 eq. (1.3) 可得，单义可译码 $W:w_1,w_2,\cdots,w_q$ 每一个码符号所携带的平均信息量：

$$R=\frac{H(S) \quad \text{比特/信源符号}}{\bar{n} \quad \text{码符号/信源符号}}=\frac{H(S)}{\bar{n}} \quad \text{比特/码符号} \tag{1.4}$$

R 为码率。码率越大，每一个码符号携带的平均信息量越大，W 有效性越高；R 越小，每一个码符号携带的平均信息量越少，W 有效性越低。

H(S) 是固定不变的，则平均码长 $\bar{n}$ 越大，R 越小，码有效性越低;$\bar{n}$ 越小，R 越大，码的有效性越高。

所以，要降低单义可译码的平均码长 $\bar{n}$，势必要考虑 q 个码长与 q 各概率分量的合理搭配。就是说，如果对信源编码不仅要求无失真，而且在无失真的前提下，还要求有较高的有效性的话，则在编码时不仅要使结构参数 q, r, n_i 满足 $Kraft$ 不等式，而且还要考虑 q 个码长与信源的概率空间中 q 个概率分量之间的合理搭配。要合理利用和充分挖掘信源的统计特性潜力，才能使其平均码长 $\bar{n}$ 尽量小。

定理 1.2. **设设离散无记忆信源 *S* 的信息熵为 *H(S)*，码符号集 *X* 的码符号数为 *r*，则单义可译码 *W* 的平均码长**

$$\bar{n} \geqslant \frac{H(S)}{logr} \tag{1.5}$$

定理 1.3. **设离散无记忆信源 *S* 的信息熵为 *H(S)*，码符号集 *X* 的码符号数为 *r*。若信源编码的平均码长**

$$\bar{n} < \frac{H(S)}{logr}$$

则用码符号集 *X* 对信源 *S* 的信源编码不可能单义可译。

推论 1.1. **设离散无记忆信源 *S* 的信息熵为 *H(S)*，码符号集 *X* 的码符号数为 *r*，则单义可译码 *W* 的码率 *R*（信息单位/码符号）$\leqslant logr$，当且仅当信源编码 *W* 是最佳码时，*R=logr*。若要码率 $R > logr$，则信源码 *W* 一定不能单义可译。**

定理 1.4. **设离散无记忆信源 *S* 的信息熵为 *H(S)*，码符号集 *X* 的码符号数为 *r*。用码符号集 *X* 对信源 *S* 编出的单义可译码的平均码长**

$$\bar{n} < \frac{H(S)}{logr} + 1 \tag{1.6}$$

推论 1.2. **设离散无记忆信源 *S* 的信息熵为 *H(S)*，码符号集 *X* 的码符号数为 *r*。用码符号集 *X* 对信源 *S* 进行无失真信源编码的平均码长 $\bar{n}$ 满足**

$$\frac{H(S)}{logr} \leqslant \bar{n} < \frac{H(S)}{logr} + 1 \tag{1.7}$$

其码率 *R* 满足（根据 *eq.* (1.4)）

$$\frac{logr}{1 + \frac{logr}{H(S)}} < R \leqslant logr \tag{1.8}$$

该推论说明，用含有 r 种码符号的码符号集 X，对信息熵为 H(S) 的离散无记忆信源 S 进行无失真信源编码，**其平均码长 $\bar{n}$ 在 $\left[\frac{H(S)}{logr}, \frac{H(S)}{logr} + 1\right]$ 中取值，码率 R 在 $\left[\frac{logr}{1 + \frac{logr}{H(S)}}\right]$ 中取值。**

1.5 信源扩展与数据压缩

信源 $S: s_1, s_2, \cdots, s_q$ 发出的消息，往往不是信源 S 的单个符号 $s_i (i = 1, 2, \cdots, q)$，而是由单个符号 s_i 组成的某一序列。若馨苑 S 发出的消息由 N 个符号组成，则每一条消息都可看作信源 S 的 N 次扩展信源 $\mathbf{S} = (S_1, S_2, \cdots, S_N)$ 的某一个“符号”$\alpha_i = (s_{i1} s_{i2} \cdots s_{iN})$(其中: $s_{i1} s_{i2} \cdots s_{iN} \in s_1, s_2, \cdots, s_q; i1, i2, \cdots, iN = 1, 2, \cdots, q; i = 1, 2, \cdots, q^N$)。若在构造单义可译码时，不把信源符号 s_i 作为编码对象，而直接把消息 $\alpha_i = (s_{i1}, s_{i2}, \cdots, siN)(i = 1, 2, \cdots, q^N)$ 作为编码对象，是一个完整的码字 w_i 不对应单个信源符号 s_i，而直接对应一个消息 α_i，使码字 w_i 与 α_i 一一对应。这样的编码方法，是每个信源符号 s_i 所需要的平均码符号数，即平均码长进一步下降，码率进一步提高。

定理 1.5. 设离散无记忆信源 S 的信息熵为 $H(S)$，码符号集 X 的码字符号数为 r。若用码符号集 X 中的码符号对无记忆信源 S 的 N 次扩展信源 $S^N = S_1S_2\cdots S_N$ 进行单义可译编码，则当扩展次数 N 足够大 $(N \to \infty)$ 时，单义可译码的平均码长 $\bar{n}$ 可无限地接近下限值 $H(S)/logr$，即有

$$\lim_{N\to\infty} \bar{n} = \frac{H(S)}{logr} \tag{1.9}$$

根据以上 theorem 1.5，若不把信源单个符号作为编码对象，而直接把信源的 N 次扩展信源的单个"符号"作为编码对象，是单义可译码的码字一一对应，泽当扩展次数 N 足够大时，信源的每一个信源符号所需的平均码符号数，即平均码长可无限接近于下界值，单义可译码的码率可无限接近与 logr。接近的程度随着扩展次数的增加而增加。编码的有效性将明显提高。

单义可译码的平均码长的减少，表明每传递一个信源符号所需传递的码符号数随之减少。这表明，采用扩展信源的手段，可以达到数据压缩的目的。当然，这要付出相应的代价，码字数将从 q 增加到 q_N，当 q 和 N 相当大时，编码将变得相当复杂，其复杂程度同样随着扩展次数 N 的增加而明显地增大。

定理 1.6. 设各态经历有记忆离散信源 S 的极限熵为 H_∞，码符号集 X 的码符号数为 r。若用码符号集 X 中的码符号对信源 S 的 N 次扩展信源 S=$S_1S_2\cdots\S_N$ 进行单义可译编码，则当扩展次数 N 足够大 $(N \to \infty)$ 时，单义可译码的平均码长 $\bar{n}$ 可无限接近于 $H_\infty/logr$，即有

$$\lim_{n\to\infty} \bar{n} = \frac{H_\infty}{logr} \tag{1.10}$$

推论 1.3. 设各态历经的 m 阶 Markov 信源 m 阶条件熵为 $H(S_{m+1}/S_1S_2\cdots S_N)$，则用码符号数为 r 的码符号集 X 对信源 S 稳定后的没一条消息进行单义可译编码，其平均码长

$$\bar{n} = \frac{1}{logr} \cdot H(S_{m+1}/S_1S_2\cdots S_m) \tag{1.11}$$

各态历经的 m 阶 *Markov* 信源 S 的极限熵：

$$H_{\infty m} = H(S_{m+1}/S_1S_2\cdots S_m) \tag{1.12}$$

这个推论 corollary 1.3告诉我们，对各态历经的 m 阶 *Markov* 信源 S 这样一种特殊的有记忆信源来说，当信源稳定后，用含有 r 种不同码符号的码符号集 X，对 m 阶 *Markov* 信源的消息进行单义可译编码时，其平均码长 $\bar{n}$ 可达到下界值 $\left\{\frac{1}{logr}H(S_{m+1}/S_1S_2\cdots S_m)\right\}$。有：

$$H(S_k/S_1S_2\cdots S_{k-1}) \leqslant H(S_{k-1}/S_1S_2\cdots S_{k-2}) \tag{1.13}$$

. 各态历经的 m 阶 *Markov* 信源 S 的记忆长度 m 越大，单义可译码的平均码长 $\bar{n}$ 就可越小，其数据压缩的程度就越高，码率 R 就越大。

综上所述，在进行无失真信源编码时，可以采用扩展信源的手段，达到压缩数据的目的。对有记忆信源来说，扩展的程度越高，压缩的效果越好，编码的有效性越高。

1.6 无失真信源编码定理

无失真信源编码通信系统的信息传输速率（简称速率）为

$$\xi = \frac{R_t}{H(S)} \quad \left(\frac{\text{比特/秒}}{\text{比特比特/信源符号}} = \frac{\text{信源符号信源符号}}{\text{秒}}\right) \tag{1.14}$$

$$R_t = \frac{R}{t} \quad \left(\frac{\text{比特/码符号}}{\text{秒/码符号}} = \frac{\text{比特}}{\text{秒}}\right) \tag{1.15}$$

速率 ξ 可作为无失真信源编码通信系统的有效性的衡量标准。

定理 1.7. **设离散无记忆信源 S 的信息熵为 H(S)，输入符号集为 X 的无噪离散信道的信道容量为 C_t(比特/秒)。若 ε 是大于零的任意小的数，则以 X 为码符号集的信源 S 的单义可译码在无噪离散信道上的信息传输速率**

$$\xi \leqslant \left[\frac{C_t}{H(S)} - \varepsilon\right] \tag{1.16}$$

定理 1.8. *设各态历经有记忆信源 S 的极限熵为 H_∞，输入符号集为 X 的无噪离散信道的信道容量为 C_t(比特/秒)。若 ε 是大于零的任意小的数，则以 X 为码符号集的信源 S 的单义可译码在无噪离散信道上的信息传输速率*

$$\xi \leqslant \left[\frac{C_t}{H_\infty} - \varepsilon\right] \tag{1.17}$$

这个定理称为有记忆信源的无失真信源编码定理。

信道容量 C_t(比特/秒) 是离散无噪信道本身的特征参量（由输入符号 r 决定)，对给定的离散无噪信道来说，C_t 是一个固定不变的量。另一方面，各态历经有记忆信源 S 的极限熵 H_∞，总是小于（或等于）离散无记忆信源 S 的信息熵 *H(S)*，即总有

$$H_\infty \leqslant H(S) \tag{1.18}$$

对同一个给定的信道容量为 C_t（比特/秒）的离散无噪信道来说，有

$$\lim_{N\to\infty} \xi_{\text{无记忆}} = \frac{C_t}{H(S)} \leqslant \lim_{N\to\infty} \xi_{\text{有记忆}} = \frac{C_t}{H_\infty} \tag{1.19}$$

其中 ξ 和 ξ 分别表示有记忆信源 S 和离散无记忆信源 S 的无失真信源编码的信息传输速率。eq. (1.19) 表明，在采用扩展信源的方法来提高单义可译码有效性的过程中，考虑信源发出符号之间的统计依赖关系，比不考虑信源发出符号之间的统计依赖关系时的有效性要高。

1.7 霍夫曼 (Huffman) 码