第12卷 第1期　　电子科学学刊　　Vol. 12, No. 1
1990 年 1 月　　JOURNAL OF ELECTRONICS　　Jan., 1990

间隔编码和新近队列编码的研究*

王　继　东
(北京邮电学院)

摘要　Elias 提出的间隔编码和新近队列编码对统计特性未知的信源是良好的自适应信源编码. 本文论证了间隔编码和新近队列编码的效率以概率队列编码的效率为上界，并将 Elias 的离散无记忆信源模型下的间隔编码和新近队列编码推广到了有限状态有记忆信源.

关键词　编码;间隔编码;新近队列编码;概率

一、引　言

信源编码的目的在于压缩.好的编码方案应该使源序列编出的码序列尽可能的短.为了定量地描述编码方案的优劣,定义信源符号熵与平均符号码长之比为编码效率. 显然，寻找高效率的编码方案是信源编码研究所关心的中心内容. 在信源统计特性已知的情况下，Haffman 编码就是效率最高的实施分组码. 但是，当信源统计特性未知或者是统计特性非恒定的情况下,Haffman 码则难以适用. 好的编码应该是具有良好自适应性的通用编码. 迄今为止已有很多人对通用编码进行了研究. Ziv 和 Lempel[1-3]很早就研究了单个序列的压缩问题,他们提出的增量分段算法[2]很有特色. Davisson[4] 和 Rissanen[5-7] 在 minimax 码集和通用算术码的研究上都取得了一定的进展. 最近由 Elias[8] 提出的间隔编码和新近队列编码是两种很有吸引力的自适应通用码,尤其是对高符号熵信源,可达到很高的编码效率. 本文对间隔编码和新近队列编码进行了进一步的研究. 证明了概率队列编码的效率是上述两种编码的上界. 此外，还将这两种编码推广到了有限状态信源.

二、概率队列编码、间隔编码与新近队列编码

设离散无记忆信源 P 的符号集为 $A=\{a_1, a_2\cdots a_m\}$，各符号的概率分别为 $p(a_1)$, $p(a_2)\cdots p(a_m)$. 令 $l_H(a_i)$ 表示符号 a_i 的 Haffman 码字长度，令 $L_H(P)$ 表示信源 P 的平均 Haffman 码长. 那么,则有

*1988 年 5 月 12 日收到,同年 8 月定稿.

$$L_H(P)=\sum_i l_H(a_i)p(a_i)<H(P)+1 \tag{1}$$

式中 $H(P)$ 为信源熵.

Haffman 码的平均符号码长 $L_H(P)$ 是信源 P 按符号进行分组编码的理想平均码长. 但是,如果不知道各符号的概率值,则无法设计 Haffman 码字集. 而任何其它形式的分组编码其平均码长均高于 $L_H(P)$.

如果仅知道各符号概率的大小顺序如下式

$$p(a_1)\geqslant p(a_2)\geqslant\cdots\geqslant p(a_m) \tag{2}$$

那么可由此得出: 编码 E 为优化编码的必要条件为

$$L_E(a_i)\leqslant L_E(a_j),\qquad \text{对于 } 1\leqslant i<j\leqslant m \tag{3}$$

式中 $L_E(a_i)$ 表示符号 a_i 在编码 E 中所对应的码字.

(2)式中所描述的各符号概率之间的大小顺序我们称为概率队列. 在说明概率队列编码之前先给出一个正整数编码 e. 令 $C_e(i)$ 表示正整数 i 在编码 e 中所对应的码字, $L_e(i)$ 表示码字 $C_e(i)$ 的长度. 码字长度有如下限制:

$$\left.\begin{aligned}L_e(i)&\leqslant L_e(i+1)\\ L_e(i)&\leqslant G[\log(i)]\end{aligned}\right\} \tag{4}$$

式中 $G[\cdot]$ 为定义在$[0,\infty]$上的单调不减的上凸函数. 显然, $G[\log(\cdot)]$ 在$[1,\infty]$上也是单调不减的上凸函数,这一点在以后的证明中将要用到.

考虑如下一种编码. 对于信源符号 a_i 取编码 e 中的 $C_e(i)$ 作为其码字. 这种编码我们称其为概率队列编码,记为 PR.

下面考虑另外一种编码. 设 $X_1^N=x(1)x(2)\cdots x(N)$ 为一信源序列. 设 $x(t)$ 为当前编码符号,符号 $x(t)$ 最近的前一次出现是在 $t-k$ 处,亦即

$$k=\min_{i>0}\{i/x(t)=x(t-i)\} \tag{5}$$

将 $x(t)$ 编码为 $C_e(k)$. 那么这种编码实际上是对符号间隔的整数编码, 故称为间隔编码,记为 IN. 不难证明,如果对各信源符号设定一个初始位置,该编码是无失真可译的.

对应于上述间隔编码,如果将 $x(t)$ 不编为 $C_e(k)$ 而是编为 $C_e(k')$,其中 k' 为子序列 $x(t-k)x(t-k+1)\cdots x(t-1)$ 中不同符号的个数. 亦即

$$k'=|\{x(i)/t-k\leqslant i\leqslant t-1\}| \tag{6}$$

当设置一定的初始条件后,该编码亦是无失真可译的. 这种编码就称为新近队列编码,记为 RR.

新近队列编码的编码过程可等效为下述形式: 设有一信源符号队列, 该队列在编码过程中不断进行调整. 设 $x(t)$ 为当前编码符号, 子序列 $x(1)x(2)\cdots x(t-1)$ 为已编码序列,当前符号队列为 $\lambda(t-1)$. 设符号 $x(t)$ 在队列 $\lambda(t-1)$ 中排在第 j 位, 那么 $x(t)$ 就编码为 $C_e(j)$, 然后调整队列, 将符号 $x(t)$ 移至队列的首位, 原来队列中 1 至 $j-1$ 位的符号后移一位, 顺序数加 1. 调整后的队列记为 $\lambda(t)$. 这样就可依次进行后续符号的编码.

在上述编码过程中, 当前队列中各符号的位置是根据它们最近一次出现的先后确定

的。当然，初始队列 $\lambda(0)$ 是为了保证无失真译码而事先约定的。那么 $\lambda(0)$ 的逆序可视为一段初始序列。这样对在当前队列中排第 j 位和第 $j+1$ 位的符号 a_i 和 $a_{i'}$ 有如下解释：最近一次的 a_i 符号较最近一次的 $a_{i'}$ 符号后出现。据此将所述的队列称为新近队列，相应的编码方式称为新近队列编码。

三、间隔编码与新近队列编码的界

在新近队列编码过程中，新近队列的作用相当于对后续符号的一种预测。概率大的符号由于出现的频率高，因而排在队列前部的机会多，而小概率符号则大多数情况是排在队尾。根据前述的正整数编码特性可知：大概率符号的编码码字一般情况下较短，而小概率符号的一般较长，从而使编码达到压缩的效果。

由新近队列编码引深一步，下面考虑一个一般性的预测队列编码 E。其编码过程如下：设 $\mu(0)$ 为信源符号的初始预测队列。设 $x(t)$ 为当前编码符号，$\mu(t-1)$ 为当前预测队列。若符号 $x(t)$ 在队列 $\mu(t-1)$ 中排第 j 位，则编码为 $C_e(j)$，，然后按照某一原则将队列调整为新的预测队列 $\mu(t)$，进行后续符号的编码。

对于离散无记忆信源 P，随机变量 $X(t)$ 取符号 a_i 的概率为 $p(a_i)$。令 $R(a_i)$ 表示符号 a_i 在预测队列 $\mu(t-1)$ 中的顺序数，那么 $X(t)$ 编码的平均码长为

$$\bar{L}_E[X(t)]=\sum_{i=1}^{m}p(a_i)L_e[R(a_i)] \tag{7}$$

定理 1　预测队列编码 E 的效率以概率队列编码的效率为上界。

证明　设在编码 E 的当前预测队列 $\mu(t-1)$ 中，若有某个 a_i 使得 $R(a_i)>i$，那么一定存在一个 $i'>i$，使得

$$R(a_{i'})<R(a_i) \tag{8}$$

若将 a_i 和 $a_{i'}$ 在队列中的位置互换，得到新的队列 $\mu^{\Delta}(t-1)$。显然

$$R^{\Delta}(a_i)=R(a_{i'}),\qquad R^{\Delta}(a_{i'})=R(a_i) \tag{9}$$

式中 $R^{\Delta}(a_i)$ 表示符号 a_i 在队列 $\mu^{\Delta}(t-1)$ 中的顺序数。

对于修正的队列 $\mu^{\Delta}(t-1)$，$X(t)$ 编码的平均码长为

$$\begin{aligned}\bar{L}^{\Delta}[X(t)]&=\sum_{i}p(a_i)L_e[R^{\Delta}(a_i)]\\&=\sum_{i}p(a_i)L_e[R(a_i)]\\&\quad+[p(a_i)-p(a_{i'})](L_e[R(a_{i'})]-L_e[R(a_i)])\end{aligned}$$

因为 $p(a_i)\geqslant p(a_{i'})$，再根据(4)和(8)式得：

$$\bar{L}^{\Delta}[X(t)]\leqslant\sum_{i}p(a_i)L_e[R(a_i)]=\bar{L}_E[X(t)] \tag{10}$$

由上式可以推知，当队列按符号下标的自然顺序排列时，$X(t)$ 编码的平均码长最短。令 $R^*(a_i)$ 表示符号 a_i 在自然顺序队列 $\mu^*(t-1)$ 中的顺序数，$\bar{L}^*[X(t)]$ 表示对于队列

$\mu^*(t-1)X(t)$ 编码的平均码长. 显然

$$R^*(a_i)=i \tag{11}$$

$$\bar{L}^*[X(t)]=\sum_i p(a_i)L_e(i) \tag{12}$$

(12)式正是概率队列编码的平均每符号码长. 故定理得证.

令 $C_E[X_1^N]$ 表示信源序列 $X_1^N=x(1)x(2)\cdots x(N)$ 经过 E 编码所得到的码序列. 定义

$$\beta_E(X_1^N)=\frac{L[C_E(X_1^N)]\cdot\log|B|}{N} \tag{13}$$

式中 $L(\cdot)$ 为长度函数. B 为码符号集. 称 $\beta_E(X_1^N)$ 为序列 X_1^N 对于编码 E 的复杂度.

对信源 P 我们亦引入复杂度的概念. 令

$$\bar{\beta}_E(P)=\overline{\lim_{N\to\infty}}\sum_{X_1^N\in P}p(X_1^N)\beta_E(X_1^N) \tag{14}$$

$$\underline{\beta}_E(P)=\underline{\lim_{N\to\infty}}\sum_{X_1^N\in P}p(X_1^N)\beta_E(X_1^N) \tag{15}$$

其中 $p(X_1^N)$ 为序列 X_1^N 的概率，$\bar{\beta}_E(P)$ 和 $\underline{\beta}_E(P)$ 分别被称为信源 P 对编码 E 的复杂度上极限和复杂度下极限.

若 $\bar{\beta}_E(P)=\underline{\beta}_E(P)$ (16)

记

$$\beta_E(P)=\bar{\beta}_E(P)=\underline{\beta}_E(P)$$

称 $\beta_E(P)$ 为信源 P 对编码 E 的复杂度.

定理 2 (a) 设 $\lambda(0)$ 为新近队列编码的初始队列，$\lambda(0)$ 的逆序为间隔编码的初始序列，那么有

$$\beta_{IN}(X_1^N)\geqslant\beta_{RR}(X_1^N)\geqslant L(1) \tag{17}$$

(b) 对于离散无记忆信源 P 有

$$\beta_H(P)\leqslant\beta_{PR}(P)\leqslant\underline{\beta}_{RR}(P)\leqslant\bar{\beta}_{RR}(P)\leqslant\beta_{IN}(P)\leqslant G[H(P)] \tag{18}$$

证明 (a) 设 $X(t)$ 为当前编码符号，$C_e(k)$ 和 $C_e(k')$ 分别为 IN 编码和 RR 编码码字. 那么一定有

$$k\geqslant k'\geqslant 1 \tag{19}$$

根据(4)式则有

$$L_e(k)\geqslant L_e(k')\geqslant L_e(1) \tag{20}$$

由此可推得(17)式成立.

(c) Haffman 码是最佳的分组编码，所以(18)式中的第一个不等式自然成立. 第二个不等式为定理 1 的直接推论. 第三个不等式亦是自然成立.

根据(17)式可得

$$\bar{\beta}_{RR}(P) \leqslant \bar{\beta}_{IN}(P) \tag{21}$$

当序列长度$N\to\infty$时，IN 编码初始条件的影响可忽略不计. 那么，符号 a_i 之间的间隔为 k 的概率为

$$p(a_i)[1-p(a_i)]^{k-1} \tag{22}$$

信源 P 对编码 IN 的复杂度存在，且为

$$\beta_{IN}(P)=\sum_i p(a_i)\sum_{k=1}^{\infty} p(a_i)[1-p(a_i)]^{k-1}L_e(k) \tag{23}$$

所以有

$$\bar{\beta}_{IN}(P)=\beta_{IN}(P) \tag{24}$$

联系(21)式证得(18)式中的第 4 个不等式成立. 当$N\to\infty$时，符号 a_i 的平均间隔为

$$\bar{k}(a_i)=\sum_{k=1}^{\infty} kp(a_i)[1-p(a_i)]^{k-1}=\frac{1}{p(a_i)} \tag{25}$$

利用(4)式、(25)式和函数 $G[\cdot]$ 的性质，由(23)式可进一步推得：

$$\begin{aligned}\beta_{IN}(P) &\leqslant \sum_i p(a_i)\sum_{k=1}^{\infty} p(a_i)[1-p(a_i)]^{k-1}G[\log(k)]\\ &\leqslant \sum_i p(a_i)G[\log(\bar{k}(a_i))]\\ &=\sum_i p(a_i)G\left[\log\left(\frac{1}{p(a_i)}\right)\right]\\ &\leqslant G\left[\sum_i p(a_i)\log\frac{1}{p(a_i)}\right]=G[H(P)]\end{aligned} \tag{26}$$

(18)式中的最后一个不等式得证.

定理 2 证明了概率队列编码的效率是间隔编码和新近队列编码的上界；新近队列编码优于间隔编码；还证明了所有上述编码的复杂度上界为 $G[H(P)]$.

四、有限状态信源的间隔编码和新近队列编码

实践中，大量碰到的信源是有记忆信源. Elias 只考虑了无记忆信源的间隔编码和新近队列编码. 下面的讨论将这两种编码推广到了有记忆信源.

在有记忆信源当中，一类有实际意义的信源是有限状态信源. 下面的讨论也只限于此类信源. 设 $S=\{s_1,s_2,\cdots s_M\}$ 为信源状态集，函数 $f(\cdot\cdot)$ 为 $S\times A\to S$ 的状态转移函数. 亦即：

$$\forall s_j\in S,\qquad a_i\in A,\qquad \exists s_l\in S$$

使得 $f(s_j,a_i)=s_l$

设 $X_1^N=x(1)x(2)\cdots x(N)$ 为一信源序列，初始状态用 $\alpha(0)$ 表示. 该序列对应的状态序列表示为 $\alpha(0)\alpha(1)\alpha(2)\cdots\alpha(N)$. 其中 $\alpha(i)=f[\alpha(i-1),x(i)]$，$1\leqslant i\leqslant N$. $\alpha(i)$ 亦可记为 $f[\alpha(0),X_1^i]$.

对于平稳有限状态信源 P,其概率特性由一组条件分布 $\{p(a_i/s_j)/a_i\in A,\ s_j\in S\}$ 给出. 由条件分布可导出状态概率分布 $\{p(s_j)/s_j\in S\}$. 信源符号熵为

$$H(P)=\sum_{j=1}^{M}p(s_j)\sum_{i=1}^{m}p(a_i/s_j)\log\frac{1}{P(a_i/s_j)}$$

$$=\sum_{j=1}^{M}p(s_j)H(P/s_j) \tag{27}$$

式中 $H(P/s_j)$ 为状态 s_j 下的条件符号熵.

考虑一类弱大数平稳有限状态信源.当信源序列长度 $N\to\infty$ 时,状态 $s_j(1\leqslant j\leqslant M)$ 在状态序列中出现的频率以概率 1 趋近于 $p(s_j)$. 状态 s_j 下后续符号 a_i 出现的频率以概率 1 趋近于条件分布 $p(a_i/s_j)$. 用符号 Ω 表示这一类信源.

设 $P\in\Omega$,$X_1^N=x(1)x(2)\cdots x(N)$ 为信源 P 的一个序列,初始状态为 $\alpha(0)$. 将 X_1^N 按照一定规则分解为如下 M 个子序列.

$$\left.\begin{aligned}
&X_1^N/s_1=x[s_1(1)]\quad x[s_1(2)]\cdots x[s_1(n_1)]\\
&X_1^N/s_2=x[s_2(1)]\quad x[s_2(2)]\cdots x[s_2(n_2)]\\
&\cdots\\
&X_1^N/s_M=x[s_M(1)]\quad x[s_M(2)]\cdots x[s_M(n_M)]
\end{aligned}\right\} \tag{28}$$

式中 $f[\alpha(0),X_1^{s_j(i)-1}]=s_j$,对于 $1\leqslant i\leqslant n_j,\ 1\leqslant j\leqslant M$

$s_j(i_1)<s_j(i_2)$ 对于 $1\leqslant i_1<i_2\leqslant n_j,1\leqslant j\leqslant M$

上述序列的分解可简述如下:设当前状态为 $\alpha(t-1)=s_j$,后续符号 $x(t)$ 则续入 X_1^N/s_j 子序列. 由 $\alpha(t-1)$ 和 $x(t)$ 决定出下一状态 $\alpha(t)=s_{j'}$,后续符号 $x(t+1)$ 则续入 $X_1^N/s_{j'}$ 子序列. 依次类推即可将 X_1^N 展成(28)式的形式.

对各子序列进行 Haffman 编码、概率队列编码、间隔编码和新近队列编码,可得到类似定理 2 的结论,所不同的只是原来的 $H(P)$ 换成了 $H(P/s_j)$. 又因为状态 s_j 出现的频率随着 N 的无限增大以概率 1 趋近于 $p(s_j)$. 所以在状态平均下,利用 $G[\cdot]$ 的特性容易得出结论:定理 2 的所有关系式对弱大数平稳有限状态信源亦成立.

对有限状态信源序列的 IN 编码和 RR 编码,其初始状态的设置是 M 组而不是无记忆信源情况下的一组. 每组初始状态对应一个子序列 $X_1^N/s_j(1\leqslant j\leqslant M)$. 在序列分解为(28)式的同时,即可进行 IN 和 RR 编码.

IN 编码和 RR 编码的实用需要具体的正整数编码 e. Elias 提出了两种二元编码方法[8]. 其中之一是将整数 i 编为 $\lfloor\log i\rfloor$ 个零后续 i 的 $\lfloor\log i\rfloor+1$ 位的二进制表示. 其中 $\lfloor x\rfloor$ 表示不大于 x 的最大整数. 对应该编码方式有

$$G(x)=1+2x \tag{29}$$

利用上述方案,对一篇英语文章[9]去掉标点符号,大小写字母归一外加一个空格符作为信源符号集,分别按无记忆信源序列和一阶马氏信源序列进行新近队列编码. 所得平均码长分别为 6.29bit 和 4.37bit. 由此可以看出对于有记忆信源采用无记忆模型下的 RR 编码和采用有记忆模型下的 RR 编码两者在压缩效果上的差别之大. 该英语文章总字符数为 6771,以其统计参数近似为信源概率参数,求得信源的一阶符号熵为 3.08bit.由

(29)式可得 $G[H(P)] = 7.16$bit. 而实际的 RR 编码平均码长远小于此值. 这说明定理 2 中的上界 $G[H(P)]$ 对 RR 编码来说相当松. 在本实验当中，一阶马氏模型下的 RR 编码效率达到 70%，这一结果是相当可喜的.

五、结　　语

IN 编码和 RR 编码对于统计特性未知的信源以及非平稳信源类的分段平稳信源是良好的具有实用意义的自适应编码. 定理 1 证明了概率队列编码的效率为预测队列编码效率的上界，从而为新近队列编码的进一步改进提供了理论依据. 间隔编码和新近队列编码应用范围的拓广肯定了他们不仅适用于无记忆信源，同时也适用于有记忆信源.

本文得到了周炯槃教授的大力支持，特此致谢.

参 考 文 献

[1] A. Lempel, J. Ziv, *IEEE Trans. on IT*, **IT-22**(1976)1, 75-81.
[2] A. Lempel, J. Ziv, *IEEE Trans. on IT*, **IT-24**(1978)9, 530—536.
[3] J. Ziv, *IEEE Trans. on IT*, **IT-24**(1978)7, 405—412.
[4] L. D. Davisson, *IEEE Trans. on IT*, **IT-26**(1980)3, 166—174.
[5] J. Rissanen, *IEEE Trans. on IT*, **IT-29**(1983)9, 656—664.
[6] J. Rissanen, *IEEE Trans. on IT*, **IT-30**(1984)7, 629—636.
[7] J. Rissanen, *IEEE Trans. on IT*, **IT-32**(1986)7, 526—532.
[8] P. Elias, *IEEE Trans. on IT*, **IT-33**(1987)1,3—10.
[9] 英语世界，1988 年，第 2 期，第 50—57 页.

STUDY OF INTERVAL AND RECENCY RANK SOURCE CODING

Wang Jidong
(Beijing University of Posts and Telecommunications, Beijing)

Abstract Interval and recency rank coding, which are invented by Elias, are good adaptive source coding schemes for independent source. The upper bound of coding efficiency of the two schemes is shown to be that of probability rank coding, and the concept of interval and recency rank coding is extended to relative sources.

Key words Coding; Interval coding; Recency rank coding; Probability