5. 基本的な情報源符号化法

ハフマンブロック符号化は情報源符号化定理を証明するために導入したものであるが，その効率はよくない．ここでは，ハフマンブロック符号の効率の悪さについて述べた後で，平均符号長を効率的に理論限界に近づける基本的手法として，非等長情報源系列の符号化，ランレングス符号化，算術符号化を紹介する．

5.1 ハフマンブロック符号の限界

これまでの議論によれば，情報源Sの 1 次エントロピー$H_1(S)$を与える式

$$H_1(S) = -\sum_{i=1}^{M} p_i \log_2 p_i$$

は，情報源記号 a_i に対する符号長の下限$-\log_2 p_i$の平均であるとみなすことができる．ハフマンブロック符号化により，1 情報源記号あたりの平均符号長Lを，与えられた情報源Sの 1 次エントロピー$H_1(S)$にいくらでも近づけることができるが，誤差の上限はブロック長nに対して$\frac{1}{n}$である．

この性質は，実際の符号化に使用するためには全く不十分である．例えば，情報源S:⟨A: 0.002, B:0.998⟩について考えてみよう．Sの 1 次エントロピーは$H_1(S) \approx 0.02081$である．ここで，平均符号長が$H_1(S)$の 5%超となる 0.02185 以下となる瞬時符号を構成してみよう．平均符号長の$H_1(S)$からの超過量を$0.02185 - 0.02081 \approx 0.00104$以下にするためには，ブロック化の次数$n$を$\frac{1}{n} \leq 0.00104$，つまり，$n \geq 962$としなければならない．そのためには$2^{962} \approx 3.898 \times 10^{289}$個の情報源系列に対してハフマン符号を構成しなければならないことになり，現実には不可能である．

5.2 非等長情報源系列の符号化

この問題を回避する一つの策は，長さnの情報源符号を一様に符号化せず，符号化を行う情報源系列を非等長にすることである．具体的には，

(1) 情報源Sから発生する全ての系列を一意的に分解する情報源記号の系列$\{\alpha_1, \cdots, \alpha_m\}$をテンプレートとして選び，それに対してハフマン符号化を行う．

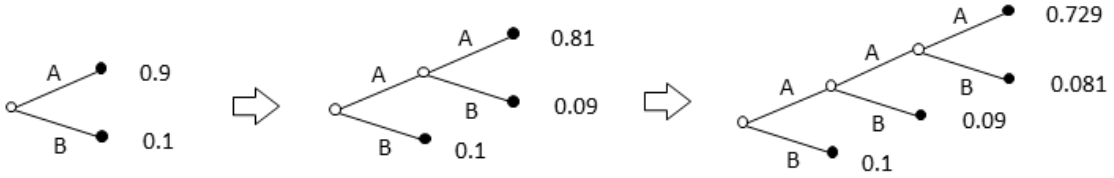
(2) ハフマン符号化に似たやり方で，$\{\alpha_1, \cdots, \alpha_m\}$の平均系列長が大きくなるようにする．

とする．

例えば，情報源記号$\{\mathrm{A}, \mathrm{B}\}$をもつ情報源$S$において，各情報源記号の発生確率を

$$P(\mathrm{A}) = 0.9, P(\mathrm{B}) = 0.1$$

としよう．まず，Sからの任意の情報源記号系列を一意的に分解できるm個の系列を，平均系列長が最大になるように選ぶ．ここでは$m = 4$としておこう．このためには，次のように最大の確率をもつ葉を伸ばし続ければよい．

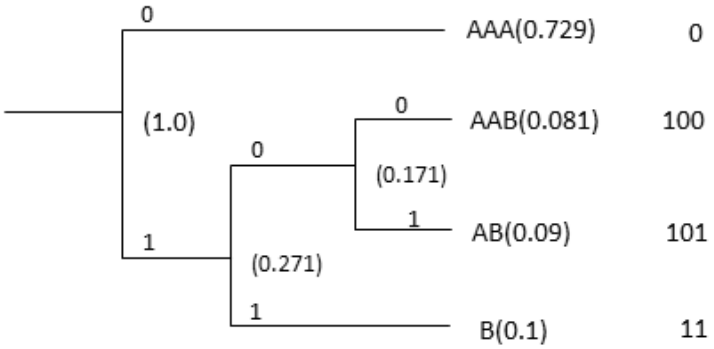


ここで，Sからの任意の情報源記号系列は，{B, AB, AAB, AAA}というテンプレートの要素の並びに一意に分解できることに注意しておこう．Sからの任意の情報源記号系列を一意に分解する情報源記号系列の大きさ 4 の集合のなかで{B, AB, AAB, AAA}は平均長が最大である．実際，{B, AB, AAB, AAA}の平均長は，

$$\bar{n} = 0.1 \times 1 + 0.09 \times 2 + 0.081 \times 3 + 0.729 \times 3 = 2.71$$

となる．

つぎに，{B, AB, AAB, AAA}に対して，次のようにハフマン符号化を行う．



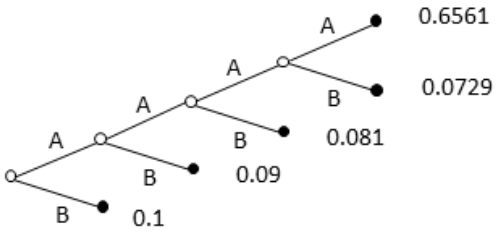
するとその結果得られる符号語{0,100,101,11}の平均符号長は

$$0.1 \times 2 + 0.09 \times 3 + 0.081 \times 3 + 0.729 \times 3 = 1.442$$

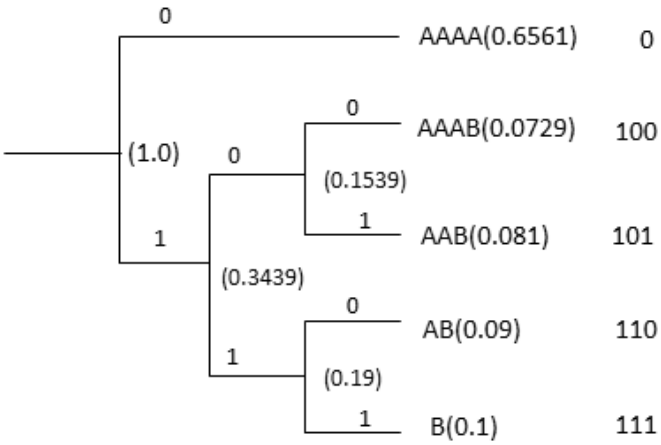
となる．

最終的な符号化器は，もとの情報源Sを，情報源記号列を{B, AB, AAB, AAA}という部分系列を出力する情報源S'と見立てる．Sからの情報源記号平均 2.71 個あたり，平均符号長 1.442 の符号を出力するので，情報源記号あたりの平均符号長は，$L = \frac{1.442}{2.71} \approx 0.532$となる．

$m = 5$とすると，テンプレート集合は，



となる．ここで得られたテンプレート集合{B, AB, AAB, AAAB, AAAA}の平均長は，3.429 である．このテンプレート集合に対してハフマン符号化を行うと，



となる．その平均符号長は 1.6878 である．従って，情報源記号平均 3.439 個あたり，平均符号長 1.6878 の符号を得る．従って情報源記号あたりの平均符号長は，$\frac{1.6878}{3.439} \approx 0.491$となる．　$H(S) \approx 0.469$に対して 4.7%増しとなり，かなり改良されたと言える．

問題　これは，何次のハフマンブロック符号を行ったのと同等の効果か？

5.3 ランレングス符号化法

ランレングス符号化法では，上に示した方式を少し単純化したものであり，同じ記号が連続するランレングス(run length)を用いて符号化する．例えば，Aのランレングス（ただし，最大値 4）によって符号化するのであれば，

B ⇒　0 回
AB ⇒　1 回
AAB ⇒　2 回
AAAB ⇒　3 回
AAAA ⇒　4 回

というテンプレートを用いる．情報源記号列

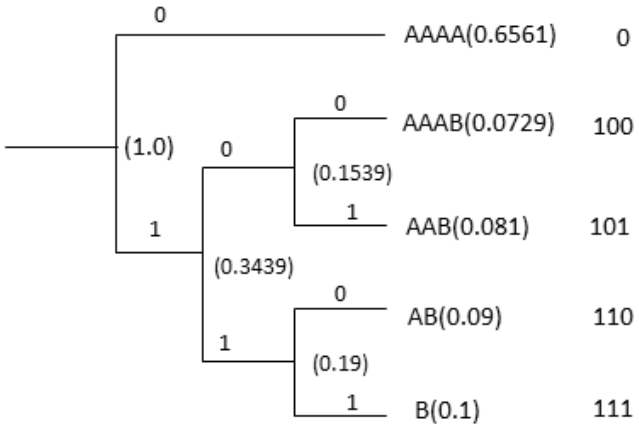
ABBAAAAAAABAAAB

については，

AB · B · AAAA · AAB · AAAB

と分解したのちに符号化を行う．

ランレングスハフマン符号化法は，ランレングスをさらにハフマン符号化したものである．例えば，



という作業を行い，$\mathrm{AAAA, AAAB, AAB, AB, B}$というランレングスに対して，$0,100,101,110,111$という符号語をそれぞれ対応付ける．

$P(\mathrm{A}) = 1-p, P(\mathrm{B}) = p, p < 1-p$，つまり，Aが連続して発生する傾向にあると仮定し，上のようにして実現されるランレングス符号化法の平均符号長を計算してみよう．

$\mathrm{B, AB, AAB, \cdots, A}^{N-1}$と，長さ$N-1$までのAのランレングスを符号化すれば，その平均長$\bar{n}$は，

$$\begin{aligned}\bar{n} &= \sum_{i=0}^{N-2}(i+1)p_i + (N-1)p_{N-1} \\ &= \sum_{i=0}^{N-2}(i+1)(1-p)^i p + (N-1)(1-p)^{N-1} \\ &= \frac{1-(1-p)^{N-1}}{p}\end{aligned}$$

となる．

これらの系列をハフマン符号化するときの平均符号長L_Nは，

$$L_N < -\sum_{i=0}^{N-1} p_i \log_2 p_i + 1$$

となる．ここで，

$$p_i = \begin{cases} p(1-p)^i & i = 0, \cdots, N-2 \\ (1-p)^i & i = N-1 \end{cases}$$

であることに注目すると，

$$L = \frac{L_N}{\bar{n}} < H(S) + \frac{1}{\bar{n}}$$

が得られる．

例えば，$p = 0.001, N = 2^{10}, n = \log_2 N = 10$とすれば，$H(S) \approx 0.0114$である．この問題に対して，普通にフマンブロック符号化すると，

$$L = \frac{L_N}{n} < H(S) + \frac{1}{n}$$

であるので，平均符号長の上限は，

$$L < H(S) + \frac{1}{n} \approx 0.0114 + \frac{1}{10} = 0.1114$$

となる．

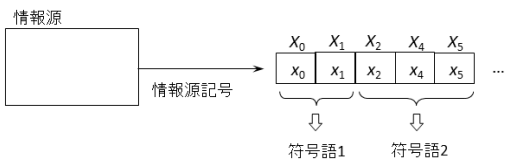
一方，ランレングスハフマン符号化を適用すれば 1 情報源記号あたりの平均符号長は，

$$L < H(S) + \frac{1}{\bar{n}} \approx 0.0114 + \frac{1}{1-(1-p)^{2^{10}-1}} \approx 0.0114 + 0.00156 = 0.01296$$

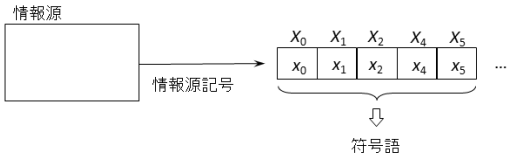
となる．0.00156が下限からの増分となる．

4.7 算術符号

通常の符号化では符号化の結果は符号語の列となるが，算術符号化では，情報源系列全体を一つの符号語に符号化する．



通常の符号化の枠組み



算術符号化の枠組み

情報源Sは，$\{\mathrm{A},\mathrm{B}\}$を情報源記号とし，

$$P(\mathrm{A}) = 1-p, P(\mathrm{B}) = p, p < \frac{1}{2}$$

となる記憶のない 2 元情報源（Aが発生しやすい）とする．Sの（n次 2 元）算術符号化は次のように行う．

(1) Sから発生する長さnの情報源系列（2^n通り）に0～2^n-1の番号を付す．

(2) 第i番目の系列をb_i，その発生確率を$P(b_i)$とするとき，b_0～b_{i-1}の発生確率の和：

$$C(b_i) = \begin{cases} 0 & i = 0 \\ \sum_{j=0}^{i-1} P(b_i) & i = 1, \cdots, 2^n - 1 \end{cases}$$

を「b_iの累積確率」と呼ぶ．

(3) $C(b_i)$について

$$0 = C(b_0) < C(b_1) < \cdots < C(b_{2^n-1}) < 1$$

が成立する．

(4) 情報源系列b_iの累積確率 を他と区別できる 2 進数で最小桁数表示したものを符号語として受信者に通報する．

ここで，$C(a_i)$の値は，

$$\begin{cases} P(\lambda) = 1 \\ C(\lambda) = 0 \\ P(x\mathrm{A}) = (1-p)P(x) \\ P(x\mathrm{B}) = pP(x) \\ C(x\mathrm{A}) = C(x) \\ C(x\mathrm{B}) = C(x) + P(x\mathrm{A}) \end{cases}$$

に基づいて計算する．例えば，　$P(\mathrm{A}) = 0.9, P(\mathrm{B})$=0.1 とすると，

$$C(\mathrm{AAA}) = C(\mathrm{AA}) = C(\mathrm{A}) = C(\lambda) = 0$$

$$C(\mathrm{AAB}) = C(\mathrm{AAA}) + P(\mathrm{AAA}) = 0.9^3 = 0.729$$

$$C(\mathrm{ABA}) = C(\mathrm{AB}) = C(\mathrm{A}) + P(\mathrm{AA}) = 0 + 0.9^2 = 0.81$$

$$C(\mathrm{ABB}) = C(\mathrm{AB}) + P(\mathrm{ABA}) = C(\mathrm{A}) + P(\mathrm{AA}) + P(\mathrm{ABA}) = 0.891$$

$$C(\mathrm{BAA}) = C(\mathrm{BA}) = C(\mathrm{B}) = P(\mathrm{A}) = 0.9$$

$$C(\mathrm{BAB}) = C(\mathrm{BA}) + P(\mathrm{BAA}) = 0.9 + 0.081 = 0.981$$

$$C(\mathrm{BBA}) = C(\mathrm{BB}) + P(\mathrm{BBA}) = 0.9 + 0.9 \times 0.1 = 0.99$$

$$C(\mathrm{BBB}) = C(\mathrm{BBA}) + P(\mathrm{BBA}) = 0.99 + 0.1^2 \times 0.9 = 0.999$$

となる．