記号の出現パターンを考慮したベイズ符号の効率的アルゴリズムに関する研究

1G06H015-5　岩間大輝
指導教員　後藤正幸

1　研究背景・目的

情報通信技術の発達に伴い, 取り扱われるデータは大容量化の一途を辿っている. 大容量のデータをそのまま扱うことは, 通信路の高負荷や保存媒体の容量不足につながるため, 情報圧縮の技術が重要となっている.

情報の圧縮には確率構造の推定が必要であるが, そのような未知の情報源の推定機能を内在した方法として, ユニバーサル符号が実用化されている. ユニバーサル符号の中で, ベイズ符号 [1] とは, その符号長と理想符号長との差である冗長度がベイズ基準において最小となる符号化法である.

ベイズ符号では, 一般的に情報源アルファベットの全ての記号が出現すると仮定して推定が行われる. しかし, 現実のデータでは全ての記号が出現するとは限らない. そこで, 南茂らはアルファベット中の記号の出現パターンを考慮したベイズ符号の構成法を提案している [2]. しかし, この手法では, アルゴリズム中で繰り返し行う二項係数の乗算回数が膨大となり, 実装する上で困難となる. 本研究では南茂ら [2] で膨大になっていた二項係数の計算量の低減を目的とし, 情報源を independent identical distribution (i.i.d.) からマルコフ情報源に拡張した符号化確率の計算式の構造を利用したアルゴリズムを提案する.

2　準備

2.1　ベイズ符号

ベイズ符号は情報源の確率分布のクラスは既知であるが, そのパラメータが未知である場合を考える. 長さ n の情報源系列 $x^n = x_1x_2\cdots x_n$ に対し, パラメータの事前分布で確率モデルを平均化した混合分布を符号化確率に用いることで与えられる. ベイズ符号は, ベイズ冗長度を最小とするユニバーサル符号である [1].

情報源をマルコフ情報源, アルファベットを $A = \{a_1, a_2, \cdots, a_{|A|}\}$, 状態 s における各記号の出現確率ベクトルを $\boldsymbol{\theta}^{|A|}(s) = (\theta_1(s), \theta_2(s), \cdots, \theta_{|A|}(s))$ とし, これを全ての $s \in S$ について集めたパラメータベクトルを $\boldsymbol{\theta}$ とすると,

$$AP(x^n) = \Gamma \int P(x^n|\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{1}$$

で計算することができる [1], [2].

2.2　ベイズ符号の符号化確率の計算

状態 s における各記号の出現確率の事前確率として, 多項分布に対する自然共役事前分布である Dirichlet 分布 $P_D(\theta^{|A|}(s))$ を仮定する. このとき, $\alpha_1, \alpha_2, \cdots, \alpha_{|A|}$ を Dirichlet 分布のハイパーパラメータ, 系列 x^{t-1} における状態 s の出現回数を $N(s|x^{t-1})$, 状態 s での記号 a_i の出現回数を $N(a_i|s, x^{t-1})$ とすると, 符号化確率 $AP(x^n)$ は,

$$AP(x_t = a_i|x^{t-1}) = \frac{N(a_i|s, x^{t-1}) + \alpha_i}{N(s|x^{t-1}) - 1 + \sum_j \alpha_j}, \tag{2}$$

$$AP(x^n) = \prod_{t=1}^{n} AP(x_t|x^{t-1}), \tag{3}$$

で計算できる.

3　従来研究

南茂ら [2] は, 情報源を i.i.d. と仮定し, 情報源のアルファベットサイズ $|A|$ に比べて, 系列に実際に出現する記号数 r^* が小さい情報源について考えた. r_n を系列 x^n 内で出現した記号数とすると, $n \to \infty$ のとき $r_n \to r^*$ となる r^* を真の出現記号数と呼ぶ. このような情報源に対するベイズ符号は, 各記号の出現確率の推定だけでなく, 出現記号数の事後確率分布 $P(r|x^n)$(ただし, $r \in \{r_{x^n}, r_{x^n}+1, \cdots, |A|\}$) の推定を行い, 重み付けをする方法で与えられる. 系列 x^n で出現する記号数 r と仮定したときの x_t の事後確率分布を $Q(x_t|x^{t-1}, r)$, x^{t-1} 内で出現している記号数 $u = r_{t-1}$ とおく. このとき, 系列を x^{t-1} までを観測したもとで, x_t が出現する予測確率 $AP_D(x_t|x^{t-1})$ は次式のように計算される.

$$\begin{aligned} &AP_D(x_t|x^{t-1}) \\ &= \sum_{r=u}^{|A|} \binom{|A|-u}{r-u} P(r|x^{t-1}) Q_D(x_t|x^{t-1}, r). \end{aligned} \tag{4}$$

4　提案手法

4.1　本研究の着眼点

二項係数 $\binom{|A|-u}{r-u}$ は,

$$\binom{|A|-u}{r-u} = \frac{(|A|-u)!}{(r-u)!(|A|-r)!}, \tag{5}$$

で与えられるが, そのままの計算を行うと乗算回数は $2(|A|-u)$ である. よって, 式 (4) では二項係数 $\binom{|A|-u}{r-u}$ を $r = u$ から $r = |A|$ まで計算する必要があり, 式 (4) 全体での二項係数の乗算回数は $2(|A|-u)^2$ となる. この計算法を確率推定のアルゴリズムとしてそのまま用いると, アルファベットサイズ $|A|$ が大きくなったとき, 計算量が膨大となり実行が困難となる. そこで本研究では計算量の削減法を考える. ここで表記を簡単にするため, 二項係数 $\binom{|A|-u}{r-u}$ を

$$T_{r,u} = \binom{|A|-u}{r-u} \tag{6}$$

のように表すと, $u+1 \le r \le |A|$ に対し, 以下の関係が成り立つ.

$$T_{r+1,u} = \frac{|A|-r}{r-u+1} T_{r,u}. \tag{7}$$

情報源としてマルコフ情報源を仮定したとき, $r_{s,t-1}$ を系列 x^{t-1} において状態 s のもとで出現した記号数とし, $u_s = r_{s,t-1}$ とおく. 式 (4) を 式 (6) を用いて以下のように表す.

$$\begin{aligned} AP_D(x_t|x^{t-1}) &= \sum_{r=u_s}^{|A|} T_{r,u_s} P(r|x^{t-1}, s) Q_D(x_t|x^{t-1}, r, s) \\ &= T_{u_s,u_s} P(u_s|x^{t-1}, s) Q_D(x_t|x^{t-1}, r, s) \\ &+ \frac{A-u}{1} T_{u_s,u_s} P(u_s+1|x^{t-1}, s) \\ &\cdot Q_D(x_t|x^{t-1}, u_s+1, s) \\ &\vdots \\ &+ \frac{1}{|A|-u_s} T_{|A|-1,u_s} P(|A||x^{t-1}, s) \\ &\cdot Q_D(x_t|x^{t-1}, |A|, s) \end{aligned} \tag{8}$$

式 (8) より, この二項係数の計算を行うために, 直前の二項係数の値 T_{r,u_s} を保管している必要があるが, 乗算回数は $2(|A|-u_s)$ から 1 に削減することができる. 式 (8) 全体として, 乗算回数は $(|A|-u)$ となる.

4.2 提案手法のアルゴリズム

提案手法のアルゴリズムを以下に示す.

・初期設定

$Com = 1, AP_D(x_t|x^{t-1}) = 0$ とする.

・$AP_D(x_t|x^{t-1})$ の計算

for $r = u_s$ to $|A|$ **do**

$AP_D(x_t|x^{t-1}) := AP_D(x_t|x^{t-1}) + Com \times P(r|x^{t-1}, s) Q_D(x_t|x^{t-1}, r, s)$

$Com := Com \times \dfrac{|A|-r}{r-u_s+1}$

end for

4.3 計算量の評価

本研究で提案したアルゴリズムの計算量は x^{t-1} 内に出現している記号数 $r_{s,t-1}$ に依存するため計算量は確率変数となる. 乗算回数は南茂らの手法では式 (4) より, $2(|A|-u_s)^2$ 回, 提案手法では 4.1 節より $(|A|-u_s)$ 回となる. よって, 南茂らの手法と提案手法の 1 記号あたりの平均的な二項係数の計算量を $\overline{C_N}, \overline{C_P}$ は以下のように表すことができる.

$$\overline{C_N} = \frac{1}{n} \sum_{t=1}^{n} 2(|A| - r_{s,t-1})^2, \tag{9}$$

$$\overline{C_P} = \frac{1}{n} \sum_{t=1}^{n} (|A| - r_{s,t-1}). \tag{10}$$

ここで, $n \to \infty$ の極限における計算量について, 次の定理を得ることができる.

定理: 状態 s の定常確率を $P^*(s)$, 状態 s での出現確率が正の値を取る記号数を r_s^* とする. $n \to \infty$ のとき, $\overline{C_N}$, $\overline{C_P}$ は以下の様に概収束する.

$$\overline{C_N} \to \sum_s 2(|A| - r_s^*)^2 P^*(s), \qquad a.s. \tag{11}$$

$$\overline{C_P} \to \sum_s (|A| - r_s^*) P^*(s), \qquad a.s. \tag{12}$$

(概証明) x^n における状態 s の出現回数を $N(s|x^n)$ は, $n \to \infty$ のとき $N(s|X^n)/n \to P^*(s), a.s.$ であり, $N(s|X^n) \to \infty, a.s.$ となる. さらに $N(s|X^n) \to \infty$ ならば, $r_{s,n} \to r_s^*, a.s.$ となる.

このとき, 式 (9), 式 (10) より, 系列 x^n の 1 記号あたりの計算量はそれぞれ以下のようになる.

$$\overline{C_N} \to \frac{1}{n} \sum_s 2(|A| - r_s^*)^2 N(s|X^n), \qquad a.s. \tag{13}$$

$$\overline{C_P} \to \frac{1}{n} \sum_s (|A| - r_s^*) N(s|X^n), \qquad a.s. \tag{14}$$

$n \to \infty$ のとき, $\frac{N(s|X^n)}{n} \to P^*(s)$ であることから, 式 (11), 式 (12) が示された. □

5 数値実験

5.1 実験条件

提案手法の有効性を確認するため計算機による数値実験を行った. $|A| = 10, r_s^* = 3$, の一次マルコフ情報源に対し, 系列長は $n = 200$, シミュレーション回数は 10^5 回とし, 平均計算量で評価を行う.

5.2 結果および考察

数値実験の結果と系列全体の平均的な計算量は, 定理で示した状態毎の計算量を状態の確率 $P^*(s)$ を重み付けた漸近計算量 $\overline{C_N}, \overline{C_P}$ に漸近的に収束することを確認した. また, 提案手法は南茂らの手法に比べ計算量が $\frac{1}{2(|A|-r_s^*)}$ となることが確認できた.

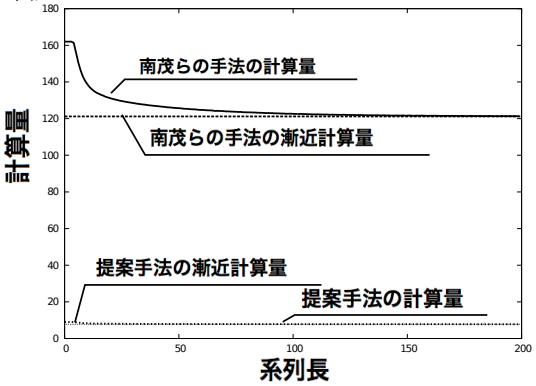

図 1. 従来手法と提案手法の計算量と漸近計算量

さらに, $|A|$ が大きくなるにつれて計算量がより大きくなるため提案手法と従来手法との差が広がり有効性が増すと考えられる.

6 まとめと今後の課題

本研究では, 記号の出現パターンを考慮したベイズ符号の計算量を削減したアルゴリズムを提案し, 漸近計算量を理論的に示した. さらに, 実際の計算量が漸近計算量に収束することを数値実験によって示した. 今後の課題はメモリ計算量を削減することが挙げられる.

参考文献

[1]T. Matsushima, H. Inazumi, and S. Hirasawa,"A class of distortionless codes designed by bayes decision theory," *IEEE Trans. Inf. Theory,* Vol. 37, No. 5, pp. 1288–1293, 1991.

[2] 南茂龍之介, 小泉大城, 松嶋敏泰,"記号パターンを考量した情報源に対するベイズ符号に関する研究," 電子情報通信学会技術報告書, IT, pp. 25–30, July 2006.