パラメータが変動する情報源の符号化アルゴリズムに関する一考察

知識情報処理研究　　3602B054-0　平本竜馬
指　　導　松嶋敏泰教授

A note of source cording algorithm with changing parameter

by Ryuma Hiramoto

1　はじめに

現在，通信産業の発展に伴い，通信技術が必要不可欠になっている．その中で，情報を圧縮，復元するための技術を情報源符号化と言う．

情報源符号化では，情報源と呼ばれる確率モデルから出力される系列 $x_1x_2...x_n$ を他の短い系列（符号語）に変換して圧縮を行う．

情報源の確率構造について完全な情報が得られていない場合の報源符号化法は，ユニバーサル情報源符号と呼ばれる．ユニバーサル符号化の代表的な符号として，Ziv-Lempel 符号があり，コンピュータでのテキストや画像の圧縮に応用されている．

ベイズ決定理論の視点から構成されたユニバーサル符号にベイズ符号がある．具体的なアルゴリズムとしては，Context-Tree Weighting(CTW) Algorithm[1] や松嶋，平澤による符号化法 [2] がある．

Ziv-Lempel 符号が有限時点での性能を保証できないのに対し，ベイズ符号はベイズ基準の下で有限時点においても平均符号長を最小化することができる点で優れている．

ベイズ符号化法に関する研究は現在，非定常な情報源に対してもなされている．坂口，川端 [3] は過去の系列切り捨てることによりパラメータの変化に対応する有限窓 CTW 法を提案しているが，情報源モデルを定義しておらず，またどの程度の情報を切り捨てるかを決定する基準がないので最適性を示すことができない．

安田 [5] の情報源モデルは，非定常な情報源モデルを明確に定義し，その上でベイズ最適でかつ計算が容易な符号化アルゴリズムを提案した．

本研究では，安田の提案した情報源モデルにおける制約を緩和し，符号化アルゴリズムを提案し，計算量と冗長度の評価を行い考察する．

2　準備

2.1　ベイズ符号化

情報源からの長さ n のデータ系列を $x^n = x_1x_2...x_n$ とする．情報源符号化ではシンボルの生起確率を用いて符号化を行うが，確率構造が未知である場合には，符号化の際に用いる符号化確率の決定が主問題となる．ここでは，確率モデルが $P(x_t|x^{t-1},\theta^k)$ のように k 次元パラメータ θ^k で特徴づけられる，パラメトリックなモデルクラスを考える．ベイズ符号化は，ベイズ決定理論により，$P(x_t \mid x^{t-1},\theta^k)$ を θ^k の事後分布 $f(\theta^k \mid x^{t-1})$ で重みづけをした次式で表される．

$$P_c(x_t|x^{t-1}) = \int P(x_t|x^{t-1},\theta^k)f(\theta^k|x^{t-1})d\theta^k. \tag{1}$$

2.2　問題設定

時刻 t の記号 $x_t \in (0,1)$ の出現確率は，パラメータ θ_t のベルヌイ分布にしたがっていると仮定する．

$$P(x_t) = \theta_t^{1-x_t}(1-\theta_t)^{x_t}. \tag{2}$$

ここで，θ_t が毎時刻で変化する場合に，ベイズ基準の下で平均符号長を最小にする符号化確率 $P_c(x_t|x^{t-1})$ を求める．

3 従来研究

3.1 安田の研究 [5]

安田は，情報源のパラメータ変化のモデルを定義し，その情報源に対して符号化確率の計算が容易であるベイズ最適な符号化を提案した．

3.1.1 情報源の変化モデル

θ_t，θ_{t+1} は次式で構成されるものとする．

$$\theta_t = \frac{A_t}{A_t + B_t}, \tag{3}$$

$$\theta_{t+1} = \frac{A_{t+1}}{A_{t+1} + B_{t+1}}. \tag{4}$$

A_t，B_t については以下の関係を満たすものとする．

$$A_t = A_t U_t, \tag{5}$$

$$B_t = B_t V_t. \tag{6}$$

ここで，

$$U_t \sim Beta(\rho\alpha'_t, (1-\rho)\alpha'_t), \tag{7}$$

$$V_t \sim Beta(\rho\beta'_t, (1-\rho)\beta'_t). \tag{8}$$

であり，$\rho \in (0,1)$ とする．さらに α_t，β_t については，

$$\begin{cases} \alpha'_t = \alpha_t - x_t + 1, \\ \beta'_t = \beta_t - x_t. \end{cases} \tag{9}$$

$$\begin{cases} \alpha_{t+1} = \rho\alpha'_t, \\ \beta_{t+1} = \rho\beta'_t. \end{cases} \tag{10}$$

のように α'_t，β'_t を介して更新されてゆく．この情報源は θ_1 のハイパパラメータである α_1，β_1 と ρ によって決定される．ρ は θ_t の変動の大きさに関係する定数であり，この値が 1 に近づくほど定常に近づき，1 で定常となる．α_1 と β_1，ρ は既知である．

3.1.2 事前分布の設定

A_t，B_t の事前分布については γ 分布を仮定する．

$$f(A_t) = \frac{1}{\Gamma(\alpha_t)} A_t^{\alpha_t - 1} e^{-A_t}, \tag{11}$$

$$f(B_t) = \frac{1}{\Gamma(\beta_t)} B_t^{\beta_t - 1} e^{-B_t}. \tag{12}$$

これは θ_t の事前分布として次式で表されるベータ分布を仮定しているのと等価である．

$$f(\theta_t) = Beta(\alpha_t, \beta_t). \tag{13}$$

3.1.3 符号化確率

以下の補題により，符号化確率を計算する．

補題 時刻 1 から $t-1$ までの記号 x^{t-1} を用いたときの時刻 t の記号 x_t のベイズ最適な符号化確率は以下で表される．

$P(x_t \mid x^{t-1})$

$$= \int P(x_t \mid x^{t-1}, \theta_t) f(\theta_t \mid x^{t-1}) d\theta_t \tag{14}$$

$$= \int \theta_t^{1-x_t} (1-\theta_t)^{x_t} f(\theta_t \mid x^{t-1}) d\theta_t \tag{15}$$

$$= \begin{cases} E[\theta_t \mid x^{t-1}] & x_t = 0 \\ 1 - E[\theta_t \mid x^{t-1}] & x_t = 1 \end{cases} \tag{16}$$

$$= \begin{cases} \frac{\sum_{k=1}^{t-1} \rho^{t-k}(1-x_k) + \rho^{t-1}\alpha}{\sum_{k=1}^{t-1} \rho^{t-k} + \rho^{t-1}(\alpha+\beta)} & x_t = 0 \\ \frac{\sum_{k=1}^{t-1} \rho^{t-k} x_k + \rho^{t-1}\beta}{\sum_{k=1}^{t-1} \rho^{t-k} + \rho^{t-1}(\alpha+\beta)} & x_t = 1. \end{cases} \tag{17}$$

ここで，α，β は事前分布 $f(\theta_1)$ の初期パラメータとする．

3.1.4 安田の提案の利点

安田の方法では，ρ, α_t, β_t だけを記憶し，毎時刻 (9),(10) の計算を行うだけで，3.1.3 節の補題により

$$E[\theta_t | x^{t-1}] = \frac{\alpha_t}{\alpha_t + \beta_t}. \tag{18}$$

が求まり，符号化確率を容易に計算することができる．

3.2 従来法の問題点

安田では情報源の変化モデルが明確に示されており，A_t, B_t に自然共役な事前分布を設定したことによって，ベイズ最適な符号化確率を容易に計算できる．しかし，実用化の面から見た場合，観測することのできない ρ を既知としていることが問題である．

4 提案法

4.1 問題設定

安田の情報源モデルにおいて，ρ が未知の場合に符号化確率 $P_c(x_t \mid x^{t-1})$ を求める．

4.2 ベイズ最適符号化確率の定式化

ρ の事前分布を $P(\rho)$ とした場合のベイズ最適な符号化確率は以下ので表される．
$P_c(x_t \mid x^{t-1})$

$$= \int\int P(x_t \mid x^{t-1}, \theta_t) f(\theta_t \mid x^{t-1}, \rho) \cdot P(\rho \mid x^{t-1}) d\theta_t d\rho. \quad (19)$$

4.3 尤度関数

尤度関数は次式で表される．

$$\begin{aligned} L(\rho) &= \prod_{t=1}^{n} P(x_t|x^{t-1}, \rho) \\ &= \prod_{t=1}^{n} \int P(x_t \mid x^{t-1}, \theta_t) f(\theta_t \mid x^{t-1}) d\theta_t \\ &= \prod_{t=1}^{n} \int \theta_t^{1-x^t} (1-\theta_t)^{x_t} f(\theta_t \mid x^{t-1}) d\theta_t \\ &= \prod_{t=1}^{n} \left(\frac{\sum_{k=1}^{t-1} \rho^{t-k}(1-x_k) + \rho^{t-1}\alpha}{\sum_{k=1}^{t-1} \rho^{t-k} + \rho^{t-1}(\alpha+\beta)} \right)^{1-x_t} \\ &\quad \cdot \left(1 - \frac{\sum_{k=1}^{t-1} \rho^{t-k}(1-x_k) + \rho^{t-1}\alpha}{\sum_{k=1}^{t-1} \rho^{t-k} + \rho^{t-1}(\alpha+\beta)} \right)^{x_t}. \end{aligned} \quad (20)$$

尤度関数は，毎時点でパラメータの異なるベルヌーイ分布の積になっており，指数型ではない．このため共役な事前分布を構成することができず，解析的に積分を解くことが不可能である．

ここで，真の $\rho = 0.996, \alpha_1 = 0.7, \beta_1 = 0.3, n = 10000$ のときの対数尤度関数は図 1 で表される．

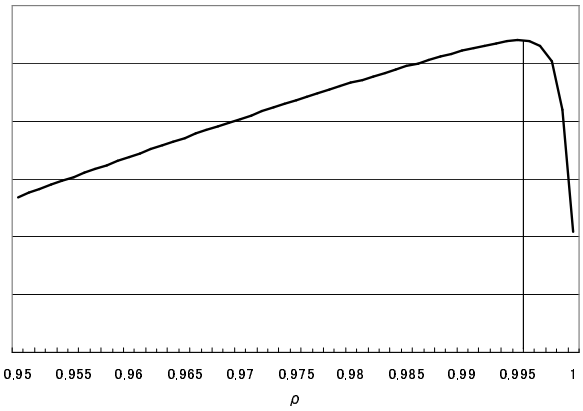

図 1: ρ の対数尤度

4.4 提案法

符号化関数を直接積分することができないため，事前分布を，等間隔でサンプリングした C 個の $\rho_i = i/C \quad (i = 1, ..., C)$ で代表させ，この事後確率の比で重みづけすることで積分計算を近似する．近似した符号化確率は次式で表される．
$P_c(x_t|x^{t-1})$

$$= \sum_{i=1}^{C} P(x_t \mid x^{t-1}, \theta_t) f(\theta_t|x^{t-1}, \rho_i) \cdot \frac{P(\rho_i|x^{t-1})}{\sum_{i=1}^{C} P(\rho_i|x^{t-1})}. \quad (21)$$

ここで，各 ρ_i に対し，安田のアルゴリズムを適用することにより，α_t と β_t を用いて ρ_i の事後確率の計算が容易に行うことができる．
$P(\rho_i|x_t) \propto P(x_t|x^{t-1}, \rho_i) P(\rho_i \mid x^{t-1})$

$$= \left(\frac{\alpha_t}{\alpha_t + \beta_t} \right)^{1-x_t} \left(1 - \frac{\alpha_t}{\alpha_t + \beta_t} \right)^{x_t} P(\rho_i \mid x^{t-1}). \quad (22)$$

5 提案法の性能

5.1 計算量の評価

提案アルゴリズムでは，C 個の ρ_i それぞれの事後確率の推定，パラメータ推定に安田のアルゴリズムを適用するため，安田のアルゴリズムの C 倍の計算量になる．

5.2 冗長度の評価

冗長度の評価をシミュレーションにより明らかにする．情報源は 3.1.1 節の変化モデルを持つと仮定する．評価基準は個別系列に対する 1 記号あたりの冗長度を次式で定義する．

$$R(x^t) = \frac{1}{t}\left(\log_2 \frac{1}{P_c(x^t)} - \log_2 \frac{1}{P_a(x^t)}\right). \quad (23)$$

ここで $P_a(x^n)$ は系列の真の出現確率であるとし，

$$P_c(x^t) = \prod_{k=1}^{t} P_c(x_k|x^{k-1}). \quad (24)$$

とする．

5.2.1 シミュレーション条件

実験　ρ の事前分布を一様分布と仮定する．情報源のパラメータを $\rho = 0.996$，初期値 $\theta_1 = 0.6(\alpha_1 = 0.6, \beta_1 = 0.4)$ とする．
提案法と，真の ρ が既知の場合で，時刻 1 から 30001 までの一記号当たりの冗長度を比較する．C の値は 10,20,50,100,1000 の 5 パターンとする．試行は時刻 30001 までの冗長度の 100 回平均をプロットした．

5.3 シミュレーション結果

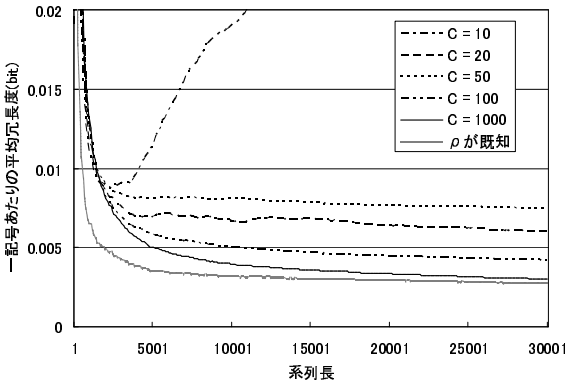

図 2: 一記号あたりの冗長度

6 考察

6.1 計算量について

計算量は C にのみ依存するので，ニュートン法による ρ の推定など，系列長とともに計算量が増大する手法に比べ，計算量を抑えることができる．

6.2 シミュレーションについて

図 2 より提案法において，C が大きいほど一記号当たりの平均冗長度が小さい．これは，区間を小さくとることで，数値積分の精度が高まるからだと言える．

7 まとめ

本研究では，安田の情報源モデルにおける ρ が未知の場合に対して，安田のアルゴリズムを利用することにより，計算量の増加を抑えた符号化アルゴリズムを提案した．

今後の課題としては，ρ が小さい場合にはパラメータ θ_t が 0 または 1 に吸収されてしまう現象，また，尤度関数が特徴的な形状をしているなどの性質があるので，これを用いて更なる冗長度の低下，計算量削減を目指したい．

参考文献

[1] F.M.J.Willems，Y.M. Shtarkov and T.J. Tjalkems，”The Context-Tree Weighting Method:Basic Properties”，$IEEE, Trans. Inform. Theory$，Vol.41, No.3, May, 1995, pp. 653-664.

[2] T. Matsushima and S. Hirasawa, ”A Bayes coding argorithm for FSMX sources”，$In Proc. Int. Symp. of Information Theory$, Sep., 1991, pp.1288-1293.

[3] 坂口 浩章，川端 勉，”有限窓を用いた文脈木重み付け法”，信学論 A, 1997,pp.2155-2163.

[4] 松嶋 敏泰，”帰納・演繹推論と予測-決定理論による学習モデル-”，第 1 回 情報理論的学習理論ワークショップ予稿集，1998，pp. 1-8.

[5] 安田 豪毅,”パラメータが時間変化する情報源とその符号化に関する一考察”，信学技報, Vol.101，No.177, 2001，pp.25-30.