シンボルの累積出現回数を考慮したベイズ予測アルゴリズムの提案

1X10C001-2　阿内宏武
指導教員　後藤正幸

1　研究背景と目的

近年，データマイニングの分野において様々な確率モデルの学習による予測アルゴリズムの有用性が示されてきた．なかでも，時系列データに対し，過去のデータが与えられたもとで次のデータを予測する問題は，多くの適用例を持つため広く研究されている．このような問題を扱う手法の一つに情報源符号化があり，過去の系列が与えられたもとで次の時点のデータの生起確率を算出し符号化を行う．近年，確率分布のクラスが既知の場合に，ベイズ基準のもとで冗長度を最小にする効率的なベイズ符号化アルゴリズムが提案されている．

須子ら [1] は，ベイズ符号化アルゴリズムを応用し，分類問題に対する効率的な予測アルゴリズムを提案した．この手法では，データ構造を表す決定木モデルを確率モデルとして定式化し，このモデルを用いてベイズ基準のもとで最適な予測アルゴリズムを示した．また，岩間ら [2] はベイズ符号から出力される木情報源に対しマルコフ情報源を仮定することで，著者推定問題において有効性を示した．著者推定問題は，書き手の特徴や文脈が重要視されるので，過去のデータの出現順序に従って次のデータが出現するマルコフモデルの当てはまりが良いと考えられる．

これに対し，次の時点でのデータが，出現順序ではなく，直近の一定期間内の累積出現回数に従うモデルクラスの存在も考えられる．例として，web ページの閲覧履歴データのように，ユーザーが過去に閲覧した厳密な順序よりも，直近の一定期間内に閲覧した回数に従い，次に閲覧するページが決定されるデータが挙げられる．このような時系列データに対し，従来のマルコフ情報源を仮定するベイズ符号化法は，データの出現順序まで考慮してしまうため，過度に複雑なモデルを仮定していることになり，予測精度の向上の余地があると考えられる．

本研究では，次の時点のデータが，直近の一定期間内の累積出現回数に従うモデルクラスをもつデータに適用可能なベイズ予測アルゴリズムを提案し，人工データによる実験を行い，提案アルゴリズムの有効性を示す．

2　ベイズ符号化法

ベイズ符号化法は，ベイズ最適性の下で情報源から出現するシンボルの符号化確率を効率的に計算する手法であり，二次的に木情報源モデルが生成される．本節では木情報源モデルと，効率的なベイズ符号化法 [3] について述べる．

2.1　木情報源モデル

情報源アルファベット $\mathcal{A}$ を $\mathcal{A}$=$\{a_1, a_2, \cdots, a_{|\mathcal{A}|}\}$，長さ n の系列を $x^n = x_1 x_2 \cdots x_n$ と定義する．ただし $x_t \in \mathcal{A}$ であり $(t = 1, 2, \cdots, n)$，これを t 時点でのシンボルと呼ぶ．マルコフ情報源では，時点 t から $\overline{D}$ だけ過去の系列 $x_{t-\overline{D}}$,$x_{t-(\overline{D}-1)}$,$\cdots$,x_{t-1} から一意に決まる状態 s_{t-1} によって次のシンボルの生起確率が決められている．この $\overline{D}$ 次のマルコフ情報源は，深さ $\overline{D}$ の完全 $|\mathcal{A}|$ 分木の葉ノードに各シンボルの生起確率を付与した木構造で表現できる．

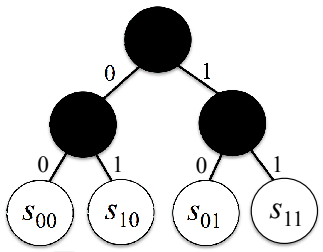

図 1．深さ $\overline{D}=2$, $\mathcal{A}$=$\{0,1\}$ の木情報源の例

深さ $\overline{D}$ のモデル集合を $\mathcal{M}$，モデル $m \in \mathcal{M}$ の状態集合を $S(m)$ とすると，図 1 の例では $S(m) = \{s_{00}, s_{10}, s_{01}, s_{11}\}$ と表せる．

2.2　効率的なベイズ符号化アルゴリズム

いま木情報源モデルの次数 $\overline{D}$ は既知であるが，真の木構造とそのパラメータが未知の場合を考える．このとき考え得る木情報源モデルは深さ $\overline{D}$ の完全 $|A|$ 分木の部分木で表せる．モデル数 $Md(\overline{D})$ は

$$Md(\overline{D}) = \begin{cases} 1 & \overline{D} = 1 \\ Md(\overline{D}-1)^{|A|}) + 1 & \text{otherwise} \end{cases} \quad (1)$$

で表される．このように木が深くなるにつれモデル数が膨大になる．そのため考え得る全てのモデルを混合した木モデルを考え，効率的な計算を行うアルゴリズムが松嶋ら [3] により提案されている．

任意の葉ノードを $s^D (0 \le D \le \overline{D})$，$s^D$ と根ノードを結ぶパス上のノード集合を $\mathcal{S} = \{s^0, s^1, \cdots, s^D\}$，$s \in \mathcal{S}$ とする．特に $t-1$ 時点で決まる葉ノードを s^D_{t-1}，s^D_{t-1} と根ノードを結ぶパス上のノード集合を $\mathcal{S}_{t-1} = \{s^0_{t-1}, s^1_{t-1}, \cdots, s^D_{t-1}\}$，$s_{t-1} \in \mathcal{S}_{t-1}$ とする．葉ノード s_{t-1} に保持されている各記号の出現確率ベクトルを $\boldsymbol{\theta}(s_{t-1})$, 葉ノード s_{t-1} の重みを $g(s_{t-1}|x^{t-1})$ とすると，符号化確率 $AP_D(x_t|x^{t-1})$ は以下のように計算することができる．

$$\begin{aligned} & AP_D(x_t|x^{t-1}) \\ &= \sum_{s_{t-1} \in \mathcal{S}_{t-1}} \int_{\boldsymbol{\theta}(s_{t-1})} P(x_t|x^{t-1}, \boldsymbol{\theta}(s_{t-1}), s_{t-1}) \\ &\qquad \times P(s_{t-1}|x^{t-1}) d\boldsymbol{\theta}(s_{t-1}) \\ &= P_C(x_t|s_{t-1} = s^0_{t-1}) \end{aligned} \quad (2)$$

$$\begin{aligned} & P_C(x_t|s_{t-1}) \\ &= \begin{cases} P(x_t|x^{t-1}, s_{t-1}) & s_{t-1} \text{ が葉ノードのとき} \\ (*) & \text{otherwise} \end{cases} \quad (3) \\ & (*) = (1 - g(s_{t-1}|x^{t-1}))P(x_t|x^{t-1}, s_{t-1}) \\ &\qquad + g(s_{t-1}|x^{t-1}) P_C(x_t|s'_{t-1}) \end{aligned}$$

$$g(s_{t-1}|x^t) = \frac{g(s_{t-1}|x^{t-1}) P_C(x_t|s'_{t-1})}{P_C(x_t|s_{t-1})} \quad (4)$$

s'_{t-1}は s_{t-1}のパス上の子ノード.

このように，混合木のただ 1 つのパスにおいて，再帰計算と重みの更新を同時に行うことにより効率的な計算を行うことができる．

3　提案手法

3.1　概要

松嶋らは，マルコフ情報源の一種である木情報源モデルを用いることで，効率的な符号化確率の計算アルゴリズムを提案した．岩間らはこのアルゴリズムを用いて，著者推定のような単語の文脈や書き手の特徴を重視するデータに対して，有効性を示している．これに対し「web ページの閲覧履歴」や「スポーツの対戦結果の時系列」のように，直近の一定期間内における累積回数を条件として，次のデータの確率が定まるモデルクラスの存在も考えられる．しかし，従来手法はマルコフ情報源を仮定しているため，過去のシンボルの出現順序関係を全て考慮してしまい，直近の一定期間内の累積回数に従うモデルクラスに対しては，過度に複雑なモデルを仮定することになり予測精度が低くなると予想される．

そこで本研究では情報源のクラスとして，次の時点のシンボルが，過去のシンボルの出現順序ではなく，直近の一定期間内の累積出現回数に従うモデルを仮定する．また，提案手法をこのようなモデルクラスに適応するために次のような木情報源モデルを提案する．

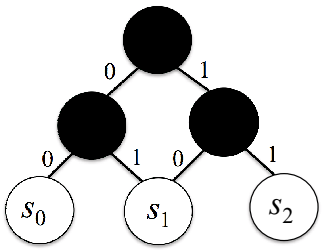

図 2. 深さ $\overline{D}=2$ の提案木情報源モデルの例

図 2 に示すような木情報源モデルは，シンボルの累積出現回数を考慮したノードを持つことができるため，一定期間内の累積出現回数を考慮した予測値を算出できる．

ここで，過去の系列 x^{t-1} により定まるパス上のノード集合を $\mathcal{S}_r(x^{t-1})$，葉ノードを $s(x^{t-1})$ と定義する．また根ノードから葉ノード $s(x^{t-1})$ に到達可能な全てのパス上のノード集合を $\mathcal{S}_n(x^{t-1})$ と定義する．従来のベイズ符号では，混合木のただ 1 つのパスを計算することで符号化確率を算出することができた．しかし提案手法では，$\mathcal{S}_n(x^{t-1})$ 内の全ノードの重みを考慮した計算を行う必要がある．すなわち，$\mathcal{S}_r(x^{t-1})$ 内と，$\mathcal{S}_n(x^{t-1})-\mathcal{S}_r(x^{t-1})$ 内で異なる再帰計算を行い予測値を算出する．

3.2 提案アルゴリズム

以下に提案アルゴリズムを示す．

Step1) $s=s(x^{t-1})$ とし，式 (5) を計算する．

$$P_C(x_t|s)=P(x_t|x^{t-1},s) \tag{5}$$

Step2) $\mathcal{S}_n(x^{t-1})$ 内にある s の全ての親ノード s' に対し式 (6) を計算する．ノードが $\mathcal{S}_r(x^{t-1})$ 内ならば重みを式 (7) で更新する．

$$P_C(x_t|s')=\begin{cases}(*) & s'\in\mathcal{S}_r(x^{t-1})\\ (**) & s'\in\mathcal{S}_n(x^{t-1})-\mathcal{S}_r(x^{t-1})\end{cases} \tag{6}$$

$$\begin{aligned}(*)=&(1-g(s'|x^{t-1}))P(x_t|x^{t-1},s')\\&+g(s'|x^{t-1})\sum_{i=1}^{|\mathcal{A}'|}P_C(x_t|s'^{(i)})\\(**)=&g(s'|x^{t-1})\sum_{i=1}^{|\mathcal{A}'|}P_C(x_t|s'^{(i)})\end{aligned}$$

$s'^{(i)}$ は $\mathcal{S}_n(x^{t-1})$ 内の s' の i 番目の子ノード

$|\mathcal{A}'|$ は s' の $\mathcal{S}_n(x^{t-1})$ 内における子ノード数

$$g(s'|x^t)=\frac{g(s'|x^{t-1})\sum_{i=1}^{|\mathcal{A}'|}P_C(x_t|s'^{(i)})}{P_C(x_t|s')} \tag{7}$$

Step3) s' が根ノードのとき，予測値を式 (8) により算出する．そうでなければ $s=s'$ として Step2 へ．

$$AP_D(x_t|x^{t-1})=P_C(x_t|s') \tag{8}$$

□

4 実験

提案手法の有効性を示すために，人工データによる実験を行う．

4.1 実験方法

次の時点のシンボルが，直近の一定期間内の累積出現回数に従って出現する人工データを発生させる．情報源アルファベットは $\mathcal{A}=\{0,1\}$, 真の木情報源モデルの深さは $\hat{D}=5$ に設定した．また，予測誤差の理論値が 0.2 になるように真のモデルのパラメータを設定した．

比較手法として，従来のベイズ符号化法を用いる．学習系列長は 10 から 90 までは 10 刻み，100 から 900 までは 100 刻みとした．また長さ 100 のテスト系列に対して，学習は行わずに予測値を算出し，シンボルの予測誤差を測る．このデータを各学習系列長に対し 100 セット用意し，予測誤差の平均を実験結果とする．また，比較手法，提案手法とも木の深さを $\overline{D}=5$ に設定した．

4.2 実験結果と考察

図 3 に実験結果を示す．横軸は学習系列長，縦軸は予測誤差を表す．

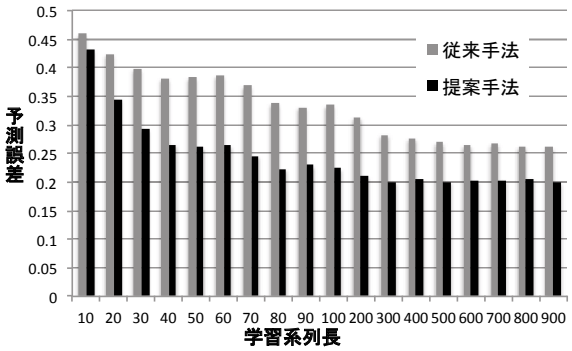

図 3. 実験結果

図 3 より，全ての学習系列長において提案手法が従来手法より低い予測誤差を示し，学習系列長が増加するにつれ理論値に近づいていることが分かる．この結果から，過去のシンボルの出現順序ではなく，直近の一定期間内の累積出現回数に従うモデルクラスに対し，提案アルゴリズムの有効性が示せた．また短い学習系列において差が顕著に表れた．提案手法では木の深さの線形オーダーで葉ノードが増加するのに対し，従来手法では木の深さの指数オーダーで葉ノードが増加する．そのため従来手法では，パラメータ数が過度に多いモデルになってしまい，学習データへのオーバーフィッティングが生じていると考えられる．

5 まとめと今後の課題

本研究では，シンボルの一定期間内の累積出現順序を重要視するモデルクラスに対するベイズ予測アルゴリズム提案し，人工データを用いた実験により有効性を示した．

今後の課題として提案手法の実データへの適用や, 計算量の削減などが挙げられる．

参考文献

[1] 須子統太, 野村亮, 松嶋敏泰, 平澤茂一, “決定木モデルにおける予測アルゴリズム”, 電子情報通信学会研究報告,COMP, コンピュテーション, Vol.103, pp.93-98, July 2003.

[2] 岩間大輝, “ベイズ符号化法によって推定された木情報源の類似度を用いた自動文書分類” , *Journal of Japan Industrial Management Association* 64(3), 438-446, 2013-10

[3] T.Matsushima and S.Hirasawa, “Reducing the Space Complexity of a Bayes Coding Algorithm using an Expanded Context Tree”,*IEEE Transactions on Information Theory*, Vol 24,No.5,pp530-536,Sep,1978.