

A Study of Bayes Prediction Algorithm for an Expanded Model Class

AUCHI Hiromu

1 研究背景・目的

時系列データに対し、ある時点までのデータが与えられたもとで次時点以降のデータを逐次予測する問題は多くの適用例を持ち、様々な分野で広く研究されている。予測に基づく情報源符号化法は、このような逐次予測問題の一つとして捉えることができるが、近年、情報源が従う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。これは、過去のデータが与えられたもとで次時点のデータの生起確率をベイズ基準のもとで算出し、算出された予測確率(符号化確率)を用いて符号化を行う手法であり、一般の時系列データへの予測問題に対する有効性が示されている [1]。

従来のベイズ符号化アルゴリズムは、情報源が従う確率分布のクラスとして、マルコフ情報源の一種である木情報源を仮定する。また、Willemsら [2] は、複数のサブセットにより構成される情報源クラスを仮定した場合の Context Tree Weighting 法(以下 CTW 法)を提案している。ここで CTW 法は、与えられたデータから最適な階層構造の重みを学習し、符号化を行う手法である。しかし、情報源符号化の問題では前述のモデルクラスでも十分であるが、他の一般的な時系列データでは、必ずしも次時点のデータが過去のデータの出現順序を条件として生起するとは限らない。そのような事象の例として、プロ野球選手の打撃成績の時系列が挙げられる。プロ野球選手が次時点にヒットを打つ確率は、現在の調子の良さ、すなわち過去の「ヒット」、「凡退」の厳密な順序ではなく、直近一定期間内の「ヒット」の回数により決定されやすいと考えられる。本研究では、前述の例のような、直近一定期間内のシンボルの出現回数を条件として、次時点のシンボルの生起確率が決定されるモデルクラスの存在に着目する。このようなモデルクラスに対し、マルコフ情報源を仮定する従来のベイズ符号化アルゴリズムでは、過度に複雑なモデルを仮定することになり、オーバーフィッティングを起こすと考えられる。また、実際の時系列データの真のモデルクラスが直近一定期間内の累積出現回数を条件とするモデルクラスか、出現順序を条件とするマルコフ情報源であるか、またその両者であるかは未知であり、事前にモデルクラスを仮定することは難しい。

このような問題に対し、著者らは最適なモデルクラスを選択するパラメータを導入し、モデルクラスが未知の場合への適用を可能にしたベイズ予測アルゴリズムを提案している [3]。しかし、実際の時系列データでは、真のモデルクラスに出現順序を条件とする場合と、累積出現回数を条件とする場合が混在している可能性が考えられる。そのような複雑なモデルクラスに従う時系列データに対しては、最適なモデルクラスを選択するパラメータのみでは対応できない。

上に示した著者らの問題点の解決のため、本研究では、真のモデルクラスに出現順序を条件とする場合と、累積出現回数を条件とする場合が混在しているような複雑なモデルクラスをもつデータに対し、より真の構造に近いモデルを学習可能なベイズ予測アルゴリズムを提案する。提案手法の有効性について、人工データと実際の時系列

データを用いた実験により評価を行う。

2 ベイズ符号化アルゴリズム

ベイズ符号は、情報源の確率構造が完全に与えられていることを前提としないユニバーサル符号化の一種であり、ベイズ基準の下で冗長度(平均符号長とエントロピーの差)を最小とする符号である。ベイズ符号化法では、情報源から出現するシンボルの符号化確率を算出し、二次的に木情報源モデルが生成される。本節ではマルコフ情報源の一種である木情報源モデルと、松嶋らにより提案された効率的なベイズ符号化アルゴリズム [1] 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 A を $A = \{a_1, a_2, \dots, a_{|A|}\}$ 、長さ n の情報源系列を $x^n = x_1 x_2 \dots x_n$ と定義する。ただし $x_t \in A$ であり ($t = 1, 2, \dots, n$)、これを t 時点でのシンボルと呼ぶ。マルコフ情報源では、 t 時点から D だけ過去までの系列 $x_{t-D}, x_{t-(D-1)}, \dots, x_{t-1}$ によって一意に決まる状態 s_{t-1} のもとで、次のシンボルの生起確率が決められている。この D 次のマルコフ情報源は、深さ D の $|A|^D$ 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。木情報源は、マルコフ情報源の一種であり、葉ノードが複数の深さに深さに存在することを許容している。いま、最大深さ \bar{D} の考える木情報源モデル集合を \mathcal{M} 、ある木情報源モデル $m \in \mathcal{M}$ の状態集合を $S(m)$ とする。ここで m はある木構造を持つ木情報源モデルであり、 \mathcal{M} において唯一の構造を持つ。

図 1 に木情報源モデル $m \in \mathcal{M}$ の例を示す。図 1 の例では、 $S(m) = \{s_{00}, s_{10}, s_{01}, s_{11}\}$ である。

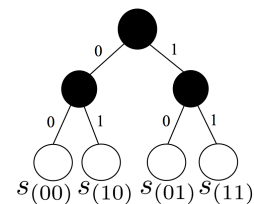


図 1: $D = 2$ の木情報源モデル m の例 ($A = \{0, 1\}$)

2.2 効率的なベイズ符号化アルゴリズム

モデル m の事前確率を $P(m)$ 、 m のパラメータを $\theta(m)$ とすると、ベイズ符号の符号化確率は次の式 (1) で表される。

$$\begin{aligned}
 &P_C(x_t|x^{t-1}) \\
 &= \sum_{m \in \mathcal{M}} \int_{\theta(m)} P(x_t|x^{t-1}, \theta(m), m) \\
 &\quad \times P(\theta(m)|x^{t-1}, m) P(m|x^{t-1}) d\theta(m)
 \end{aligned} \tag{1}$$

いま木情報源モデルの最大深さ \bar{D} は与えられているが、真の木情報源モデル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最大深さ \bar{D} の完全 $|A|$ 分木の全部分木で表せる。モデル数 $Md(\bar{D})$ は

$$Md(\bar{D}) = \begin{cases} 1 & \bar{D} = 1 \\ Md(\bar{D} - 1)^{|A|} + 1 & \text{otherwise} \end{cases} \quad (2)$$

で表される。式 (2) で表されるように、 \bar{D} が大きくなるにつれモデル数は膨大となる。このため、考え得る全てのモデルを混合した混合木を考え、ノードに重みパラメータを導入することにより、効率的に符号化確率を計算可能なアルゴリズムが松嶋らにより提案されている [1]。

以下にノーテーションを定義する。 $t-1$ 時点で得られている過去の系列によって決まる葉ノードを $s_{t-1}^{\bar{D}}$ とする。ここで、 s_{t-1} を $s_{t-1}^{\bar{D}}$ と根ノード s_{t-1}^0 を結ぶパス上の任意のノードとすると、各記号の出現確率ベクトルを $\theta(s_{t-1})$ 、ノード s_{t-1} の重みパラメータを $g(s_{t-1}|x^{t-1})$ と表せる。符号化確率 $AP_D(x_t|x^{t-1})$ は式 (1) で算出することが出来るが、重みパラメータの導入により以下のように効率的に計算することが出来る。

$$AP_D(x_t|x^{t-1}) = P_C(x_t|s_{t-1} = s_{t-1}^0) \quad (3)$$

ただし、 $P_C(x_t|s_{t-1})$ を

$$P_C(x_t|s_{t-1}) = \begin{cases} P(x_t|x^{t-1}, s_{t-1}) & s_{t-1} \text{ が葉ノード} \\ (*) & \text{otherwise} \end{cases} \quad (4)$$

$$(*) = (1 - g(s_{t-1}|x^{t-1}))P(x_t|x^{t-1}, s_{t-1}) + g(s_{t-1}|x^{t-1})P_C(x_t|s'_{t-1}) \quad (5)$$

$$g(s_{t-1}|x^t) = \frac{g(s_{t-1}|x^{t-1})P_C(x_t|s'_{t-1})}{P_C(x_t|s_{t-1})} \quad (6)$$

とし、 s'_{t-1} は s_{t-1} のパス上の子ノードとする。

上式 (3) から式 (6) のように、混合木のただ1つのパスについて、再帰計算と重みパラメータの更新を同時に行うことにより、効率的に符号化確率が計算可能となる。

3 シンボルの出現順序と累積出現回数の両者を条件とするモデルクラスに対するベイズ予測アルゴリズム [3]

2節では、マルコフ情報源の一種である木情報源モデルを仮定したベイズ符号化アルゴリズムについて示した。これに対し、一般の予測問題への適用場面を考慮し、次時点のデータが、直近一定期間内のシンボルの累積出現回数に従い生起するモデルクラスを仮定した場合のベイズ予測アルゴリズムが提案されている。

以下では、このようなモデルクラスを仮定した場合のベイズ予測アルゴリズムについて概略を述べる。

3.1 累積出現回数のみを条件とするベイズ予測アルゴリズム

2節で示した木情報源モデルは、過去のシンボルの出現順序を条件とするモデルであった。例えば、従来の木情報源モデルでは系列 $x^3 = 101$ と $x^3 = 011$ が与えられたもとで次のデータが出現する確率は区別される。しかし、ここでは一定期間内の累積回数に従い次のデータの出現確率が決まるモデルクラスを対象とするので、 $x^3 = 101$ と $x^3 = 011$ が与えられたもとで次のデータが出現する確率が同じとなるようなモデルクラスについて考える。

対象とするモデルクラスは、以下の図2のような情報源モデルで表現することができる。

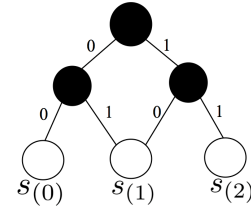


図 2: 深さ $D = 2$, $A = \{0, 1\}$ の情報源モデルの例

図2に示した情報源モデルを混合した混合モデルは、シンボルの累積出現回数を条件とするノードを持つモデルであるため、直近一定期間内のシンボルの累積出現回数を考慮した予測値を算出することができる。図2のモデルを m とし、その状態集合を $S(m)$ とすると、 $S(m) = \{s_0, s_1, s_2\}$ と表せる。また最大深さ $\bar{D} = 2$ までのモデル集合を以下の図3に示す。

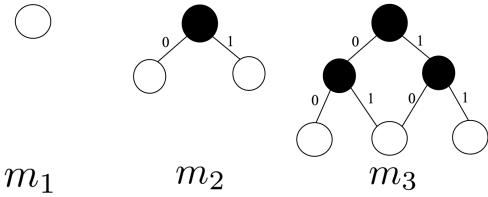


図 3: 最大深さ $\bar{D} = 2$, $A = \{0, 1\}$ の情報源モデル集合

図3に示したモデル集合では、各モデルの葉ノードは全て同じ深さに存在するモデルとなっている。そのため混合モデルを考えたとき、重みパラメータはある深さに対したただ一つ持てばよい。ここで、系列 x^{t-1} で定まる深さ D のノードを $s^D(x^{t-1})$ 、深さ D に対してただ一つ存在する重みパラメータを $g(D|x^{t-1})$ とする。

また、この情報源モデルを用いて、予測値は従来手法と同様、ある一つのパスを用いた再帰計算により算出される。

3.2 シンボルの出現順序と累積出現回数の両者を条件とするモデルクラスに対するベイズ予測アルゴリズム [3]

2節では出現順序を条件とする予測アルゴリズムについて、3.1節では累積出現回数のみを条件とするベイズ予測アルゴリズムについて示した。しかし、実際の時系列データへの適用を考えたとき、一般に真のモデルクラスは未知であるため、従来の出現順序を条件とするベイズ予測アルゴリズムと、累積出現回数を条件とするベイズ予測アルゴリズムのどちらの予測モデルを選択すべきかを事前に決めることはできない。そこで、モデル集合の選択確率を表すパラメータを導入した、実際のデータが従うモデルクラスが未知の場合へ適用可能なベイズ予測アルゴリズムが提案されている [3]。

3.2.1 モデル集合選択パラメータ

真のモデルクラスが未知の場合への適用のために、従来の出現順序を条件とする予測モデルと累積出現回数のみを条件とする予測モデルに対し、どちらのモデル集合を選択すべきかを表すパラメータ $r(T|x^t)$ を導入することにより組み合わせる ($0 \leq r(T|x^t) \leq 1$)。従来の混合木モデルを T_1 、累積出現回数のみを条件とする混合モデルを T_2 としたとき、予測確率は式 (7) で表される。

$$P(x_t|x^{t-1}) = r(T_1|x^{t-1})P(x_t|x^{t-1}, T_1) + r(T_2|x^{t-1})P(x_t|x^{t-1}, T_2) \quad (7)$$

ただし,

$$r(T_2|x^{t-1}) = 1 - r(T_1|x^{t-1}) \quad (8)$$

を満たす.

3.2.2 予測確率算出アルゴリズム

以下に予測確率算出アルゴリズムを示す

Step1) 時点 $t-1$ までの系列 x^{t-1} に対し従来モデルからの予測確率 $P(x_t|x^{t-1}, T_1)$ を式 (3)~式 (6) を用いて算出する.

Step2) 同様に, x^{t-1} に対し累積出現回数を条件とするモデルからの予測確率 $P(x_t|x^{t-1}, T_2)$ を算出する.

Step3) 予測確率を式 (9) により算出し, モデル集合選択パラメータ $r(T_1|x^{t-1})$ を式 (10) で更新する.

$$AP_D(x_t|x^{t-1}) = r(T_1|x^{t-1})P(x_t|x^{t-1}, T_1) + r(T_2|x^{t-1})P(x_t|x^{t-1}, T_2) \quad (9)$$

$$r(T_1|x^t) = \frac{P(x_t|x^{t-1}, T_1)r(T_1|x^{t-1})}{AP_D(x_t|x^{t-1})} \quad (10)$$

これらの Step1-3 を全学習系列について繰り返す.

4 提案手法

3節で示した手法は, 出現順序を条件とする予測モデルか, 累積出現回数のみを条件とする予測モデルのどちらを選択するかをモデル選択パラメータの導入により学習していた. しかし, 実際の時系列データでは, 真のモデルの構造に出現順序を条件とするノードと, 累積出現回数を条件とするノードが混在する情報源からデータが生起している可能性が考えられる. そのようなモデルに従うデータに対しては, 3節の手法では対応ができない. 図4に, 3節の手法が対応できないモデルの例を示す.

図4に示したモデルクラスへのベイズ符号化アルゴリズム化アルゴリズムの適用を行うため, 本提案手法では, 混合モデルの各ノードに, 出現順序と累積出現回数どちらを条件とすべきかを表現するパラメータを導入することで, より真の構造に近い予測モデルを構築し, 予測精度の向上を図る.

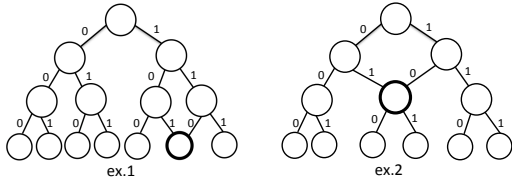


図4: 3節の手法が対応できないモデル例

4.1 ノード選択パラメータ

提案モデルの深さ D の各ノード s^D に対し, 混合木 T_1 のノード $s_{T_1}^D$ を選択する確率 $r(s_{T_1}^D|x^t)$ と, 混合モデル T_2 のノード $s_{T_2}^D$ を選択する確率 $r(s_{T_2}^D|x^t)$ を導入する ($0 \leq r(s_{T_1}^D|x^t) \leq 1$). このとき, 提案モデルのノード s の確率構造 $P(x_t|x^{t-1}, s)$ は次の式 (11) で表される.

$$P(x_t|x^{t-1}, s^D) = r(s_{T_1}^D|x^{t-1})P_C(x_t|x^{t-1}, s_{T_1}^D) + r(s_{T_2}^D|x^{t-1})P_C(x_t|x^{t-1}, s_{T_2}^D) \quad (11)$$

ただし,

$$r(s_{T_2}^D|x^{t-1}) = 1 - r(s_{T_1}^D|x^{t-1}) \quad (12)$$

また, 更新式は次の式 (13) で表される.

$$r(s_{T_1}^D|x^t) = \frac{r(s_{T_1}^D|x^{t-1})P_C(x_t|x^{t-1}, s_{T_1}^D)}{P(x_t|x^{t-1}, s^D)} \quad (13)$$

4.2 提案アルゴリズム

以下に提案手法のアルゴリズムを示す.

Step1-1) 混合木 T_1 の葉ノード $s_{T_1}^D$ に対して, 式 (14) を, 混合モデル T_2 の葉ノード $s_{T_2}^D$ に対して式 (15) を計算する.

$$P_C(x_t|s_{T_1}^D) = P(x_t|x^{t-1}, s_{T_1}^D) \quad (14)$$

$$P_C(x_t|s_{T_2}^D) = P(x_t|x^{t-1}, s_{T_2}^D) \quad (15)$$

Step1-2) 提案モデルの葉ノード s^D に対して Step1-1) で求めた $P_C(x_t|s_{T_1}^D)$ と $P_C(x_t|s_{T_2}^D)$ を用いて, 式 (16) を計算し, ノード選択パラメータを式 (17) で更新する.

$$P_C(x_t|s^D) = P(x_t|x^{t-1}, s^D) = r(s_{T_1}^D|x^{t-1})P_C(x_t|s_{T_1}^D) + r(s_{T_2}^D|x^{t-1})P_C(x_t|s_{T_2}^D) \quad (16)$$

$$r(s_{T_1}^D|x^{t-1}) = \frac{r(s_{T_1}^D|x^{t-1})P_C(x_t|s_{T_1}^D)}{P_C(x_t|s^D)} \quad (17)$$

Step2-1) $s_{T_1}^D, s_{T_2}^D$ それぞれの親ノード $s_{T_1}^{D-1}, s_{T_2}^{D-1}$ について式 (18), (19) を計算し, 重みパラメータを式 (20), (21) で更新する.

$$P_C(x_t|s_{T_1}^{D-1}) = (1-g(s_{T_1}^{D-1}|x^{t-1}))P(x_t|x^{t-1}, s_{T_1}^{D-1}) + g(s_{T_1}^{D-1}|x^{t-1})P_C(x_t|s_{T_1}^D) \quad (18)$$

$$P_C(x_t|s_{T_2}^{D-1}) = (1-g(s_{T_2}^{D-1}|x^{t-1}))P(x_t|x^{t-1}, s_{T_2}^{D-1}) + g(s_{T_2}^{D-1}|x^{t-1})P_C(x_t|s_{T_2}^D) \quad (19)$$

$$g(s_{T_1}^{D-1}|x^t) = \frac{g(s_{T_1}^{D-1}|x^{t-1})P_C(x_t|s_{T_1}^D)}{P_C(x_t|s_{T_1}^{D-1})} \quad (20)$$

$$g(s_{T_2}^{D-1}|x^t) = \frac{g(s_{T_2}^{D-1}|x^{t-1})P_C(x_t|s_{T_2}^D)}{P_C(x_t|s_{T_2}^{D-1})} \quad (21)$$

Step2-2) 提案モデルのノード s^D の親ノード s^{D-1} に対して, 式 (14), (18), (19) を用いて式 (22) を計算し, ノード s^{D-1} の重みを式 (23) で更新する. ノード選択パラメータは式 (17) で更新.

$$P_C(x_t|s^{D-1}) = (1-g(s^{D-1}|x^{t-1}))P(x_t|x^{t-1}, s^{D-1}) + g(s^{D-1}|x^{t-1})P_C(x_t|s^D) \quad (22)$$

$$g(s^{D-1}|x^t) = \frac{g(s^{D-1}|x^{t-1})P_C(x_t|s^D)}{P_C(x_t|s^{D-1})} \quad (23)$$

Step3) s^{D-1} が根ノードのとき予測値を式 (24) で算出. そうでなければ $s^D = s^{D-1}$ として Step2-1) へ.

$$AP_D(x_t|x^{t-1}) = P_C(x_t|s^0) \quad (24)$$

5 人工データを用いた実験

5.1 実験条件

真のモデルに、出現順序と累積出現回数を条件とするノードが混在するような人工データを発生させる。情報源アルファベットは $\mathcal{A} = \{0, 1\}$ 、真のモデルの深さは $D^* = 5$ とする。学習系列長を 10 から 100 までは 10 刻み、100 から 300 までは 50 刻み、300 から 500 までは 100 刻みとし、テストデータ 100 件に対して逐次予測を行なった場合の予測対数損失の平均値を算出する。それらのデータを各学習系列長に対して 1000 セット用意し、その平均を実験結果とする。予測対数損失は以下の式 (25) で表される。実験に用いる予測モデルの深さは $\bar{D} = 10$ とした。

$$L = - \sum_{t=1}^{N_T} \log P(y_t | x^n) \quad (25)$$

ここで、 N_T をテストデータ数、 y_t を t 時点でのテストデータ系列のシンボル、 n を学習系列長とする。

5.2 実験結果

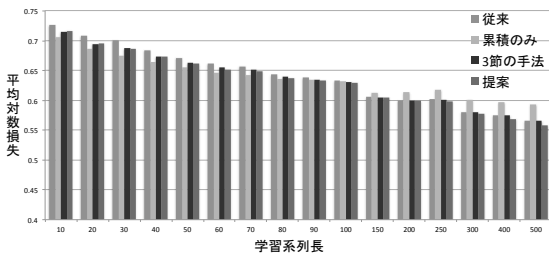


図 5: 実験結果 (人工データ)

実験結果より出現順序と累積出現回数を条件とするノードが混在するようなモデルクラスに対して、学習系列長が大きいとき、提案手法が最も低い対数損失値を示した。この結果から、本研究で対象とするモデルクラスに対する提案手法の有効性が示された。しかし、学習系列長が小さい場合は、提案手法は高い対数損失値を示している。これは、提案手法が最も多くのパラメータを持ち、ある程度多くの学習データを必要とするためと考えられる。

6 実データを用いた実験

6.1 実験条件

実際の時系列データとして、プロ野球選手の打撃成績データを用いる。情報源アルファベットとして、 $\mathcal{A} = \{0, 1\} = \{\text{非出塁}, \text{出塁}\}$ とする。学習系列長を 10 から 100 までは 10 刻み、100 から 300 までは 50 刻みとし、テストデータ 100 件に対して逐次予測を行なった場合の予測対数損失の平均値を算出する。これらを、総打席数が 400 以上の全 68 人に対して行い、その平均値を実験結果とする。実験に用いる予測モデルの深さは $\bar{D} = 5$ とした。

6.2 実験結果

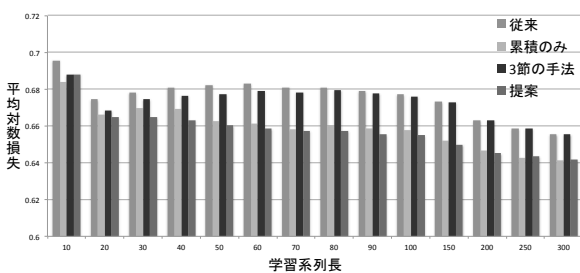


図 6: 実験結果 (実データ)

実験結果より、多くの学習系列長について、提案手法が最も低い予測対数損失値を示した。これは、ノードごとに最適なモデルクラスを選択するパラメータを導入したことで、より真の構造に近いモデルを学習できたからであると考えられる。

7 考察

人工データによるシミュレーション実験の結果より、学習系列長を増やしたとき、提案手法が最も低い予測対数損失値を示した。このことから、本研究が対象とする複雑なモデルクラスに対する提案手法の有効性が示された。しかし学習系列長が小さいときは、比較手法が優れた予測対数損失値を示している。これは、提案手法が最も多くのパラメータの学習を必要とすることに起因しており、適切なパラメータの学習にはある程度の学習系列長が必要であることが示唆される。

他方、実データを用いた実験により、実際の時系列データに対する提案手法の有効性が示された。これは、ノードごとに最適なモデルクラスを選択するパラメータの導入により、複雑な構造を持つと考えられるデータに対しその構造を捉えたモデルを構築できたためと考えられる。それに対し、モデル選択パラメータを導入した 3 節の手法は良い予測が行えていない。その理由として、今回対象とした時系列データに出現順序を条件として次データの確率が決まる場合と累積出現回数を条件として決まる場合が混在していることが考えられる。そのようなデータに対し、3 節の手法はどちらのモデルクラスが最適であるかのみを考慮しているために対応できなかったと考えられる。また、学習系列長が大きいときに累積のみを条件とする予測モデルが最も低い予測対数損失値を示す理由としては、データの非定常性の影響も考えられる。このようなデータに対し、累積のみを条件とする予測モデルは最も学習パラメータの数が少なく、短い学習系列長で学習が完了するため、確率構造が変わったデータの影響が小さい。すなわち、非定常性の影響を受けにくかったと考えられる。

8 まとめと今後の課題

本研究では、ノードごとに最適なモデルクラスを選択するパラメータを導入することにより、複雑な真の構造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た。また、人工データと実際の時系列データを用いた実験によりその有効性を示した。今後の課題として、時間、空間計算量を削減したアルゴリズムの提案、提案モデルのパラメータを解析することによる実データの分析、提案モデルの漸近解析などが挙げられる。

参考文献

- [1] T. Matsushima, S. Hirasawa, "Reducing the Space Complexity of a Bayes Coding Algorithm using an Expanded Context Tree," *Proc. ISIT2009*, pp. 719–723, 2009.
- [2] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "Context Weighting for General Finite-Context Sources," *IEEE Trans. Inform. Theory*, Vol. 42, No.5, pp. 1514–1520, Sep 1996.
- [3] 阿内宏武, 三川健太, 雲居玄道, 後藤正幸, "シンボルの累積出現回数を条件とするモデルクラスのベイズ予測アルゴリズムについて," 第 37 回情報理論とその応用シンポジウム予稿集, 2014.
- [4] 「プロ野球ヌルデータ置き場」<http://lcom.sakura.ne.jp/NulData/index.html> 2015 年 8 月 1 日アクセス.