様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　６　月　２２　日現在

機関番号：　２２３０１
研究種目：　若手研究(B)
研究期間：　2013～2015
課題番号：　２５７３００２３

研究課題名（和文）統計的学習問題に対する情報源符号化アルゴリズムの応用に関する研究

研究課題名（英文）A study on the application of a source coding algorithm to statistical learning problems

研究代表者

石田　崇（Ishida, Takashi）

高崎経済大学・経済学部・准教授

研究者番号：７０４０９６３９

交付決定額（研究期間全体）：（直接経費）　　1,700,000円

研究成果の概要（和文）：本研究では(1)情報源符号化問題における情報源モデルのエントロピー・レートの特性解析および(2)効率的な情報源符号化アルゴリズムを統計的学習問題へ応用し有効性を検証することを数値実験によって行った．(1)については従来よりも広いクラスのモデルを対象として，モデルの設定によってエントロピー・レートとその上界・下界の値の振る舞いにどのような影響を与えるかを実験した．(2)についてはあるクラスの決定木問題に対して符号化アルゴリズムを応用した手法の予測精度の解析を数値実験によって行った．また実データに対しても予測アルゴリズムを適用し手法の有効性を検証した．

研究成果の概要（英文）：In this study, we analyzed a property of entropy rate of a source models in source coding problems. An effect of model parameters to the behaviors of the value of entropy rate and the upper / lower bound of it were verified by some numerical experiments for the extended class of source model. In Addition, we analyzed an effectiveness of the prediction algorithm to which a source coding method is applied. The prediction precision of the algorithm applying to a certain class of a decision tree model was verified by some numerical experiments. Furthermore, the effectiveness of the algorithm to the actual data was also verified.

研究分野：　統計的学習理論

キーワード：　情報源符号化　統計的学習　機械学習

１．研究開始当初の背景

(1)情報源符号化問題

情報を効率良く伝送・記憶するための情報源符号化（データ圧縮）技術の重要性が改めて認識され，効率の良い符号化法の開発が求められている．情報源符号化問題では，

(a) 実データの確率構造を反映するより適切な情報源モデルクラスを設定する，

(b) 情報源モデルの理論的圧縮限界（エントロピー・レート）を求める，

(c) 情報源モデルに対して限界を達成する計算効率の良い符号化アルゴリズムを構成する，

ことが，実データに対して良い圧縮性能を達成する符号化アルゴリズム開発の鍵となる．

情報源モデルのエントロピー・レートは情報源モデルに対する理論的な圧縮限界を与えるという意味で工学的に非常に重要な指標である．情報源符号化の分野において情報源の確率モデルは当初から非常に重要な役割を果たしており，簡単な構造でありつつできるだけ実データの構造を反映させたモデルへと徐々に拡張されながら今日に至っている．

情報源モデルに対して理論的な圧縮限界であるエントロピー・レートを明らかにし，それを達成する効率的な符号化アルゴリズムを構成することが情報源符号化の一つの目的である．

(2)統計的学習問題

統計的学習理論は，人間が日常で行っている情報処理を統計的な手法を用いてコンピュータ上で実現することを目指すものである．長年にわたる学習アルゴリズムの研究に加えて，近年のコンピュータ処理性能の飛躍的な進歩によって，パターン認識や自然言語処理などの幅広い分野において実用化のレベルに到達する技術が開発されるようになってきた．これらの研究では，膨大なサンプルデータを利用し，確率モデルを仮定する統計的なアプローチが重要な役割を果たしている．特に近年ではベイズ統計学に基づく手法が急速に発展してきている．

２．研究の目的

情報源符号化アルゴリズムは，情報源の確率構造を逐次的に学習するメカニズムを内在していることが知られている．したがって，ある情報源クラスに対して理論的な最適性が保証された符号化アルゴリズムは，想定する確率モデルクラスに対して性能が保証された効率の良いオンライン学習アルゴリズムとも解釈でき，データ圧縮以外の一般的な学習問題にも応用が可能である．

本研究は情報源符号化（データ圧縮）アルゴリズムを統計的学習問題へ応用することにより，理論的な性能が保証される効率の良い学習アルゴリズムを構築することが目的である．

本研究における具体的な課題は以下の通りである．

(1)情報源符号化問題において，想定している情報源モデルのエントロピー・レートの特性について解析を行うこと

(2)効率的な情報源符号化アルゴリズムを一般的な学習問題へ応用し有効性を検証すること

(3)以上を段階的に進め，情報源符号化アルゴリズムを統計的学習問題へ応用する端緒を開くこと

である．

３．研究の方法

(1)本研究で取り扱う言語構造を反映した情報源モデルは従来の枠組みよりも広いクラスのモデルを扱うため，その性質はいまだに明らかとなってない部分が多く残されている．現在までに，いくつかの制約を設定する事によってエントロピー・レートの解析や情報源符号化アルゴリズムの構成を行ってきた．しかし，対象となる実際のデータの確率構造を考えるとこれらの制約条件は厳しいものである．そこで制約条件をより緩和した広いモデルクラスの情報源に対して，陽な形でエントロピー・レートを導出することを目指し，数値実験によりその特性の分析を行う．

(2)情報源符号化アルゴリズムの統計的学習問題への応用として，あるクラスの決定木問題に対する適用とその特性の解析を数値実験によって行う．また実データに対しても適用しアルゴリズムの有効性を検証する．

４．研究成果

(1)情報源モデルのエントロピー・レートの解析：

本研究が対象とする情報源モデルは語頭条件を満たさない Word-valued source と呼ばれるクラスのモデルである（T.ishida 他，“Properties of a Word-valued source with a non-prefix-free Word set，”IEICE Trans. Fundamentals，vol.E89-A，no.12，pp.3710 - 3722，電子情報通信学会，Dec. 2006.）．このモデルクラスについては情報源のエントロピー・レートが陽な形では導出されておらず，エントロピー・レートの上界値と下界値が得られているのみであった．

語頭条件を満たさない Word-valued source とは 0-1 の情報源系列を単語単位で出力する情報源モデルである．ここで単語とは以下の図の例で示したような単語集合 W に属する各要素{0,1,00,01}のことである．

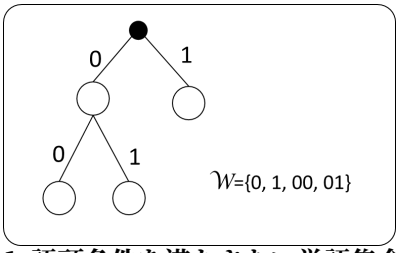

図 1:語頭条件を満たさない単語集合の例

単語集合上に与えられた確率分布 P(W)にしたがって，情報源系列が出力される．ただし，観測される系列は単語同士が連接された 0-1 の系列である．これは自然言語などの単語単位で確率構造を持つような系列の確率モデルとなっている．また，上記の単語集合では語頭条件（どの単語も他のどの単語の語頭とも一致しないこと）を満たしていないため，複数の単語系列が同一の 0-1 系列に写像されてしまう．例えば，000 という 0-1 系列は，0,0,0 という単語同士が連接したものなのか，0,00，もしくは 00,0 の連接なのか区別がつかない．このような問題により，観測される 0-1 系列の確率構造は複雑なものとなってしまう．これがエントロピー・レートを示すことのできない原因となっている．

本研究ではこのクラスの情報源の確率構造とエントロピー・レートの性質を明らかにするため，従来行っていた検証実験（石田ら，“語頭条件を満たさない単語集合をもつ Word-valued source の性質について，” 電子情報通信学会技術研究報告，IT2003-5，pp.23 - 28，電子情報通信学会，2003 年 5 月）に継続して，新しい条件やモデルの設定において数値実験を行った．

ここでは，一例として単語の系列と 0-1 系列の写像の特性とエントロピー・レートの関連性について検証した結果を示す．なお実際にはこの情報源クラスではエントロピー・レートの存在性自体についても明らかにはされていないが，(-1/n)log P(x^n)によって定義されるエントロピー密度レートを十分に長い 0-1 系列について数値実験で計算した平均値の収束値を情報源のエントロピー・レートの推定値として扱っている．

単語の系列として以下のような 8 つのモデルを想定したときの，単語の並べ替えによる単語系列のパターン数とそれを連接して生成される 0-1 系列のパターン数は以下のように計算される．

表 1：単語系列のパターン数と 0-1 系列のパターン数の対応

model		m	n	単語系列のパターン数	0-1系列のパターン数
1	0,1,00,01	4	6	24	14
2	0,0,1,00,01	5	7	60	20
3	0,1,1,00,01	5	7	60	33
4	0,1,00,00,01	5	8	60	27
5	0,1,00,01,01	5	8	60	37
6	0,0,1,1,00,01	6	8	180	55
7	0,0,1,1,00,00,01	7	10	630	119
8	0,0,1,1,00,01,01	7	10	630	185
9	0,0,1,1,00,00,01,01	8	12	2520	462

同じ数の単語（単語数：m）の組み合わせであってもそこから生成される 0-1 系列のパターン数はモデルごとに異なってくることが分かり，単語の中に含まれる 0 と 1 の数や並び順が写像に影響するものと考えられる．

次に，モデル 1～8 の単語の比率と同じ確率分布 P(W)をもつ情報源モデルに対して，エントロピー・レート（の推定値）H(X)とその上界・下界を計算した結果を以下に示す．

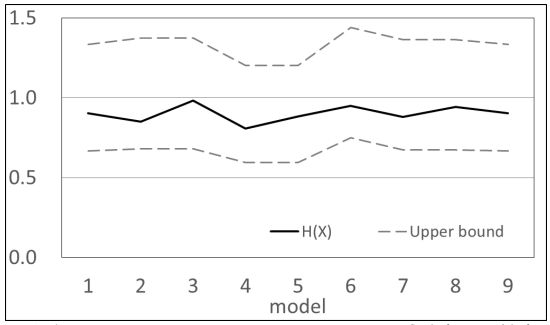

図 2：エントロピー・レートと上界，下界

例えば，モデル 2 とモデル 3 を比較したときに，上界値と下界値は同じ値をとっているが，H(X)の値は異なっていることが分かる．すなわち，単語集合の組み合わせによってそこに含まれる 0 と 1 の個数や並びのパターン，0-1 系列に写像される単語の系列の本数，割合が H(X)の値に影響していることが確認できる．

以上のような結果を通して，本課題では情報源のエントロピー・レートを陽に導出するには至らなかったものの，0-1 系列の確率構造やエントロピー・レートの値，単語集合中に含まれる 0 と 1 の個数と写像の関連性などについて一定の示唆を得ることができた．

(2) 情報源符号化アルゴリズムの統計的学習問題への応用：

目的変数がポアソン分布にしたがう決定木を用いた予測についての従来研究（峯苫ら，“目的変数がポアソン分布に従う決定木モデルにおけるベイズ最適な予測アルゴリズム，” 平成 24 年度日本経営工学会秋季研究大会予稿集，pp.24 - 25，日本経営工学会，2012 年 11 月.）について，実データへの応用を目指して，数値実験によって予測アルゴリズムの特性を解析した．

この予測アルゴリズムは冗長度の点において理論的な圧縮性能が保障された情報源符号化法の一つであるベイズ符号法（T. Matsushima 他，“A class of distortionless codes designed by Bayes decision theory，” IEEE Transactions on Information Theory，vol.37，no.5，pp.12889 - 1293）の計算効率の良いアルゴリズム（T.Matsushima 他，“A Bayes coding algorithm using context tree，” Proceedings of 1994 IEEE International Symposium on Information Theory，Jun. 1994），（T.Matsushima 他，“A Bayes coding algorithm for FSM sources，”

Proceedings of 1995 IEEE International Symposium on Information Theory, Sep. 1995）が応用されたものである．

ここでの問題設定は説明変数としてK次元の質的変数ベクトル X が与えられたもとで，計数データの目的変数Yを予測するものである．決定木によって説明変数の層別パターンを表現し，ベイズ符号化アルゴリズムの混合決定木上で目的変数のベイズ予測を行う．

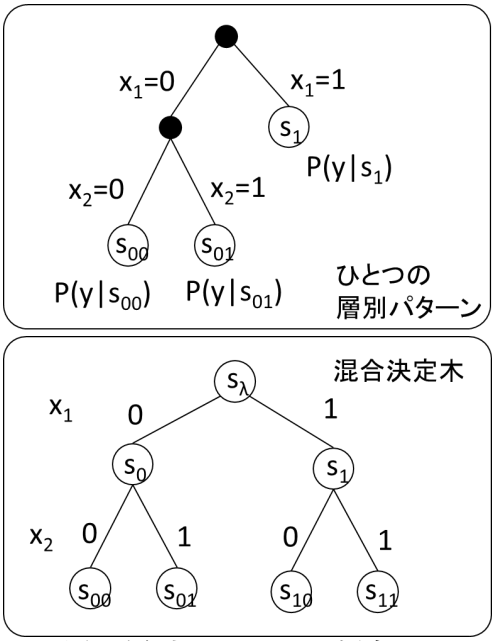

図 3:層別のパターンと混合モデル

本研究で数値実験によって解析したアルゴリズムの特性は次の通りである．

(a) 交互作用が予測に与える影響

(b) 予測アルゴリズムで層別の順番（決定木を開く順番）が固定されていることの予測精度への影響

これらの実験において，予測性能を比較する手法として①必ず全ての変数で層別するモデル1つのみを用いた決定木アルゴリズム，②ポアソン回帰モデル（AIC によるモデル選択，交互作用項なし），③（AIC によるモデル選択，交互作用項一部あり），④（全変数使用，交互作用項なし），⑤（全変数使用，交互作用項あり），⑥CART モデル，⑦Random Forests Regression モデルを用いた．

(a)交互作用効果の検証については，ポアソン分布の平均値λについて

$\ln \lambda = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \gamma \beta_{12} x_1 x_2$

として与えることでγによって交互作用の影響を数値化できるようにした．

以下にパラメータ設定と実験結果の一例を示す．

表 2a:ポアソン分布のパラメータ設定(例)

β	設定値
β0	1.6
β1	-0.9
β2	0.0
β12	1.6

表 2b:ポアソン分布のパラメータ設定(例)

ノード	A：$\lambda=e^A$	λ
s00	β0	4.95
s01	β0＋β2	4.95
s10	β0＋β1	2.01
s11	β0＋β1＋β2＋β12	9.97

パラメータγを 0 から 1 へ変化させた時の予測アルゴリズムの誤り率のグラフの一例を以下に示す．

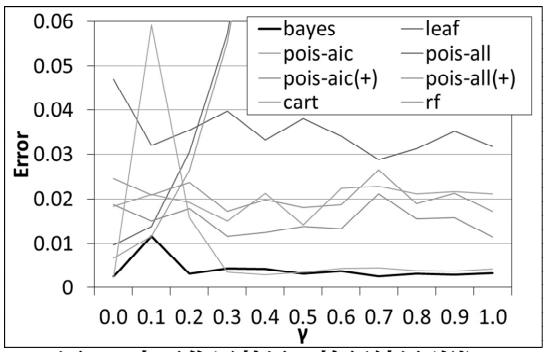

図 4: 交互作用効果の検証結果(例)

これらの結果により，交互作用の影響が強くなっても混合モデルによるベイズ予測手法の精度がよいことが確認できた．

(b)層別の順番が固定されている問題に関しては，真のモデルの層別の順序が既知の場合と未知の場合とで予測精度がどのように変化するか検証した．実験結果の一例を以下に示す．

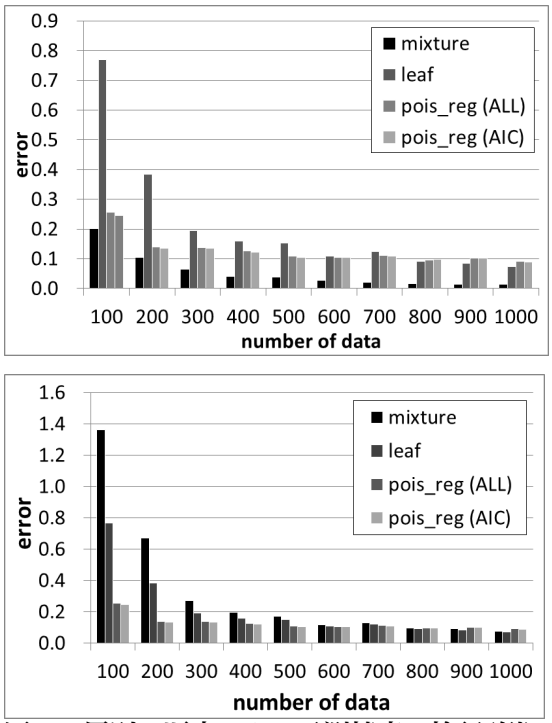

図 5：層別の順序による予測精度の検証(例)
（順序が既知(上図)・未知(下図)の場合）

上記の結果から真のモデルの層別の順序が既知か未知かは予測の精度に大きく影響することから，あらかじめ何らかの手法で順序を推定する必要があることがわかる．

さらに，RAND 研究所による社会実験データ（P.Deb 他，“The Structure of Demand for Health Care: Latent Class versus Two-Part Models，” Journal of Health Economics，vol.21，no.4，pp.601 – 625，2002.）に対して予測アルゴリズムを適用して予測精度の検証を行った．実験結果の一例を以下に示す．

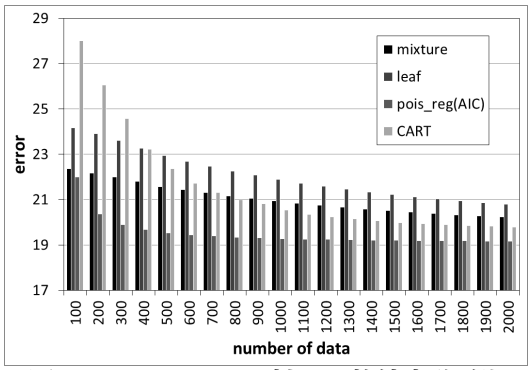

図 6：RAND データに対する数値実験(例)

この実験では，ベイズ符号によるアルゴリズムの予測精度が必ずしもよくなるとは限らず，AIC によるモデル選択を用いたポアソン回帰モデルや CART の方がよい性能を示す場合が見られた．実データへの応用については想定しているモデルと対象データがもっている特性とのかい離が問題であると考えられるため，その差を埋めるための検討を引き続き行う必要がある．

今後の展望として，本課題から得られた上記の知見をもとに数値実験を継続して，対象とするモデルクラスの拡張や予測アルゴリズムの改良を行っていく必要がある．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計０件）
〔学会発表〕（計０件）
〔図書〕（計０件）
〔産業財産権〕
○出願状況（計０件）
○取得状況（計０件）
〔その他〕
ホームページ等：なし

６．研究組織
(1)研究代表者
石田　崇（ISHIDA，Takashi）
高崎経済大学・経済学部・准教授
研究者番号：　７０４０９６３９