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　２７　年　　６　月　１５　日現在

機関番号：　３２７０６
研究種目：　基盤研究(C)
研究期間：　2012～2014
課題番号：　２４５６０４８２

研究課題名（和文）十分統計量と個別冗長度に着目したユニバーサルＶＦ符号の理論解析および設計法

研究課題名（英文）Theoretical Analysis and Design of Universal VF Code Based on Sufficient Statistic and Pointwise Redundancy

研究代表者

有村　光晴（Arimura, Mitsuharu）

湘南工科大学・工学部・講師

研究者番号：８０３１３４２７

交付決定額（研究期間全体）：（直接経費）　　3,800,000円

研究成果の概要（和文）：可変長ブロックの集合を固定長符号語の集合に写像するVF符号の理論的な性能評価を行った．まず，VF符号の部分クラスとして考えることのできる，固定長ブロックの集合を固定長符号語の集合に写像するFF符号について，符号化レートで圧縮性能を評価した場合と，符号化レートとその下限の差である冗長度で圧縮性能を評価した場合について，それぞれの最適な符号クラスの包含関係の条件について明らかにした．また，複数のブロックを切り出すマルチショット・タンストール符号について，平均符号化レートのエントロピーへの収束および符号化レートの概収束を証明し，１個のブロックを切り出すワンショット符号との違いを明らかにした．

研究成果の概要（英文）：Theoretical performance of variable-to-fixed length source codes is investigated. First, the performance of fixed-to-fixed length code, which is a subclass of variable-to-fixed length code, is investigated under two criteria, which are the asymptotic coding rate and the asymptotic redundancy. Under these two criteria, the optimal code class are defined and the conditions of inclusion relations of these two classes are formulated by the information-spectrum methods. Second, the asymptotic performance of a multi-shot Tunstall code is investigated for a stationary memoryless source. Geometrical mean of the leaf numbers of the Tunstall trees is newly defined. It is clarified that this quantity plays crucial roles at the evaluation of the asymptotic coding rate. Moreover, it is clarified that the situation is different between one-shot Tunstall code and multi-shot Tunstall code. The difference is characterized by the geometrical mean of the leaf numbers of Tunstall trees.

研究分野：　情報理論

キーワード：　無歪みデータ圧縮　情報理論　情報源符号化　情報スペクトル理論　VF符号　タンストール符号

１．研究開始当初の背景

(1) データ圧縮アルゴリズムはコンピュータおよびデジタル通信において様々なデータの圧縮に用いられているが，主に二つの立場からの研究が行われている．一つが，扱うデータを専門とする計算機科学者や画像処理研究者などによる，アルゴリズムの考案と実データに対する性能評価を中心とする研究であり，もう一つが，情報理論研究者による，扱うデータを確率過程もしくは任意の系列として扱うことによる圧縮方式の理論的な性能の評価を中心とする研究である．

(2) これらの研究はそれぞれ別々に行われており，ある新しいデータ圧縮アルゴリズムが考案されたとき，実験的にファイルを圧縮することによって，ある種類のデータに対して適している事が示されたとしても，データの種類そのものが理論的に定式化されていない限り，任意のデータに対してデータ圧縮アルゴリズムの性能が予測できる訳ではない．また，あるクラスの確率過程に対して漸近的な最適性が示されたとしても，それが有限長のあるファイルに対してどの程度適した評価であるかは分からない．

(3) 情報理論を用いてデータ圧縮アルゴリズムを評価する際に，ユニバーサル性という性質が存在する．ある圧縮アルゴリズムが，あるクラスの情報源クラス全体からの出力データ全てに対して，漸近的に最適な性能を示すとき，その圧縮アルゴリズムが，その情報源クラスに対してユニバーサルであるという．データ圧縮のユニバーサル性は，そのデータ圧縮アルゴリズムが適したデータ形式を情報理論的にモデル化するものであり，ある与えられたデータに対して高性能なデータ圧縮アルゴリズムを設計しようとする場合，最も重要な性質である．

(4) しかし，これまで情報理論的な研究においては，定常情報源，定常エルゴード情報源，マルコフ情報源，定常無記憶情報源といった，限られたクラスに対してユニバーサル性が評価されているのみで，より細かいデータクラスに対するユニバーサルデータ圧縮の理論的評価は行われていない．

(5) そこで，あるデータ圧縮アルゴリズムが与えられたとき，従来情報理論で扱われてきた情報源クラスだけではなく，テキストデータや画像データなど，用途の限られたデータ形式に対して，情報理論的なモデル化が可能になることが求められる．これにより，これまでよりもよりデータの性質に合致した圧縮性能の評価が理論的に可能になり，実用的な研究と理論的な研究の間のギャップが埋まると考えられる．

２．研究の目的

(1) 本研究では，十分統計量と個別冗長度に着目することで，可変長データの集合を固定長符号語の集合に写像する VF 符号が，良く圧縮できるようなデータのクラスを理論的に定式化することを目的とする．

特に，これまで提案されてきた圧縮性能に関する様々な指標を総合的に用いることで，データ圧縮アルゴリズムの多面的な高性能化が可能となる．これにより，現在までに提案されているものよりも圧縮性能が高い VF 符号を提案することを目指す他，圧縮後のデータが固定長であるという利便性を生かした VF 符号の応用例を新規に提案することを目指す．

(2) 方法論としては特に，情報スペクトル的方法を用いることによって，従来の定常性や無記憶性，マルコフ性，整合性など，確率過程において良く用いられる仮定を全て取り払った，極めて一般的な情報源クラスに対する解析を行う．

(3) 次に，統計学における十分統計量の概念を導入し，これとデータ圧縮のユニバーサル性の関連を理論的に明らかにすることで，あるデータ圧縮アルゴリズムがユニバーサルであるような情報源クラスを，これまで以上に細かく解析する．

(4) 最後に，個別符号化レートと自己情報量の差である個別冗長度を評価の尺度に用いることで，符号化レートだけを評価する場合よりも細かい評価を行う．前述の情報スペクトル的方法においては主に，符号化レートと情報源の確率的な性質から決まる情報スペクトル的量とを比較し，前者が後者に収束するという形で定理が述べられており，個別冗長度を用いた研究例はほとんど行われていないため，新規の結果が得られることが期待される．

(5) これらの方法論を用いることで，VF 符号が最適に圧縮できるようなデータの種類を，情報理論的に明らかにすることを目的とする．

３．研究の方法

(1) 研究は大きく以下の３つに分けて行う．まず，情報スペクトル的方法を用いた，VF 符号および FF 符号の最適性に関する解析を行う．次に，VF 符号化アルゴリズムの，有限アルファベットに対する理論的な性能評価を行う．最後に，部分列数え上げ法の理論的な解析と，その VF 符号への応用に関する検討を行う．以下で順に述べる．

(2) VF 符号は FF 符号をその特別な場合とし

て含む. FF 符号についての解析は情報スペクトル理論の枠組みで進んでいるが, VF 符号の解析はこれまで行われてきていない．そこでまず，情報スペクトル理論の枠組みの中で，FF 符号と VF 符号の比較を行う．さらに，FF 符号によって漸近的に最適な符号のクラス，VF 符号によって漸近的に最適な符号のクラスの２つを定義し，これらの符号クラスの包含関係を調べる．また，FF 符号によって漸近的に最適な符号のクラスに含まれる全ての符号において，その符号化レートが最適値である上エントロピースペクトル上限に収束する条件を明らかにする．

(3) VF 符号の一つであるタンストール符号において，系列から複数のブロックを切り出して符号化したアルゴリズムについてその符号化レートを評価する．これまでのタンストール符号の解析においては，系列から一つのブロックを切り出したアルゴリズムについての解析しか行われていない．これにより，実際用いられているアルゴリズムに近いモデルに対する理論的解析を目指す．特に，この解析においては，複数のブロックを切り出す際に用いるタンストール木の葉の数を固定せず任意のばらばらの値に設定した状態で解析を進めることで，これまで提案されていない新しい高性能な VF 符号化アルゴリズムの発見も目指す.

(4) 最近提案された部分列数え上げ圧縮法の漸近的な圧縮性能について情報理論的に解析を行う．この符号は個別系列に対して，全ての長さのブロックの頻度を数え上げて符号化するものであり, VF 符号としての拡張も期待されるアルゴリズムである．しかし，情報理論的な解析はほとんど行われておらず，この状態では VF 符号としての定式化も十分にできるとは言えない．そこでまずは，FV 情報源符号としての理論的な性質を明らかにすることを目指す.

４．研究成果

(1) 可変長ブロックを固定長の符号語に写像する VF 符号は，その特別な場合として，固定長ブロックを固定長の符号語に写像する FF 符号を含む. FF 符号の性能評価の際に，これまでは符号化レートが用いられてきたが，本研究では，FV 符号の評価に多く用いられている冗長度を FF 符号の評価に導入し，符号化レートを評価の尺度に用いた場合とで状況が変わることを示した．特に，定常性やエルゴード性を仮定しない一般情報源を対象として FF 符号の符号化レートと冗長度のそれぞれで最適な符号のクラスを定義し，その包含関係が，情報スペクトルの幅に関する上界および下界の不等式において等号が成立するかどうかによって決まることを明らかにした(論文③，学会発表⑦⑨⑫).

この際，情報スペクトルの左端の位置が漸近的にある一点に収束する場合に，冗長度による最適な符号のクラスと符号化レートによる最適な符号のクラスが完全に一致することが分かった．これに対応して，情報スペクトルの右端の位置が漸近的にある一点に収束する場合について，操作的に意味のある条件を導いた(論文②，学会発表⑨).

これらの結果は VF 符号の部分クラスである FF 符号に関する結果であるため，VF 符号に対しても同様の結果が成立するかどうか確認することが今後必要である．また，FF 符号に関する結果と VF 符号に関する結果を比較することで，新たな評価尺度で VF 符号の優位性が示される可能性があると考えられる.

(2) 系列から可変長のブロックを１個だけ切り出して固定長で符号化する one-shot タンストール符号について，以前の研究で，従来のタンストール符号の評価方法とは異なる平均符号語長の評価を行った. VF 符号の平均符号語長は従来，固定長の符号語長を切り出した可変ブロック長の期待値で割ることによって定義されていた．この平均符号語長の定義として，各系列に対してブロック長で符号語長を割ることで求まる個別符号化レートの期待値を用い，このような定義でもタンストール木を成長させることで平均符号語長が定常無記憶情報源のエントロピーに収束することを証明した.

今回の研究では，この解析方法を，系列から可変長のブロックを複数切り出して固定長で符号化する multi-shot タンストール符号に適用した．その結果，one-shot タンストール符号の場合と同様に，符号化が進むにつれて平均符号語長が定常無記憶情報源のエントロピーに収束することを証明した(学会発表④)．また，この際に，複数のタンストール木に対して木のデカルト連結および木の葉の数の幾何平均という操作を導入し，この二つが鍵となることを示した.

また，one-shot および multi-shot タンストール符号について，木の葉の数およびその幾何平均を増加させたときに符号化レートが定常無記憶情報源のエントロピーレートに概収束することを示した(学会発表①).

ただし，木の葉の数を固定したとき，タンストール符号は切り出されるブロック長の期待値が最大になるという意味での最適性を有するため，上記のように個別符号化レートの期待値に対して最適であることは保証されない．そこで，ある固定された葉の数の様々な木の中で，個別符号化レートの期待値が最適となるような VF 符号について考察を行った(学会発表②③)．この結果，タンストール符号においては確率が最大の葉を拡張していくことで最適な木が生成されるのに対し，確率 P(w)と木の深さ d(w)から求まる P(w)/{d(w)(d(w)+1)}という値を木の葉に対し

て計算し，この値が最大となる葉を拡張していくことで，平均個別符号化レートが最適な木を生成できることが明らかになった．さらに，この木はタンストール木に比べて，符号化レートの最悪値およびオーバーフロー確率の意味でも性能が高くなることを理論および実験の両面から定性的に示した．

このアルゴリズムについては，より詳細な性能評価が，理論・実験の両面から必要であるが，最悪冗長度および冗長度のオーバーフロー確率がタンストール符号に比べて抑えられることが，実際にデータを圧縮する際にも期待されるため，これまでよりも高性能な VF 符号が開発できることが期待される．

(3) 最近提案されているデータ圧縮アルゴリズムの一つとして CSE (Compression by Substring Enumeration，部分列数え上げ)法が存在する．このアルゴリズムの情報理論的評価を進めた(論文①，学会発表⑤⑧⑪)．これにより，有限次数のマルコフ情報源に対する最悪冗長度が明らかになった．

この圧縮アルゴリズムは，全ての長さの部分列の頻度を数え上げるため，任意の次数のマルコフ情報源だけでなく，定常エルゴード情報源などに対して最適であることが期待されるだけでなく，タイプを用いたデータ圧縮を実用的に一般化したものであるため，今後 VF 符号版の CSE 法を開発するにあたっても，この特徴が優位に働くことが期待される．

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３件）

① Ken-ichi Iwata, Mitsuharu Arimura and Yuki Shima, “Evaluation of Maximum Redundancy of Data Compression via Substring Enumeration for k-th order Markov Sources,” IEICE Trans. Fundamentals, Vol E97-A, No. 8, pp. 1754-1760, Aug., 2014. DOI: 10.1587/transfun.E97.A.1754

② Mitsuharu Arimura, Hiroki Koga and Ken-ichi Iwata, “A Characterization of Optimal FF Coding Using a New Optimistically Optimal Code,” IEICE Trans. Fundamentals, Vol. E96-A, No. 12, pp. 2443-2446, Dec., 2013. DOI: 10.1587/transfun.E96.A.2443

③ Mitsuharu Arimura, Hiroki Koga and Ken-ichi Iwata, “Redundancy-Optimal FF Codes for a General Source and its Relationships to the Rate-Optimal FF Codes,” IEICE Trans. Fundamentals, Vol E96-A, No. 12, pp. 2332-2342, Dec., 2013. DOI: 10.1587/transfun.E96.A.2332

〔学会発表〕（計１２件）

① 有村光晴, “Multishot Tunstall 符号の定常無記憶情報源に対する概収束符号化定理,” 第 37 回情報理論とその応用シンポジウム(SITA2014)予稿集, pp. 112-117, 富山県黒部市, Dec. 9-12, 2014.

② 有村光晴, “平均個別符号化レートが最適な VF 符号,” 第 37 回情報理論とその応用シンポジウム(SITA2014)予稿集, pp. 118-123, 富山県黒部市, Dec. 9-12, 2014.

③ 有村光晴, “平均個別符号化レートが最適な VF 符号と Tunstall 符号の比較,” 第 37 回情報理論とその応用シンポジウム(SITA2014)予稿集, pp. 124-129, 富山県黒部市, Dec. 9-12, 2014.

④ Mitsuharu Arimura, “On the Coding Rate of a Multishot Tunstall Code for Stationary Memoryless Sources,” Proc. 2014 International Symposium on Information Theory and its Applications (ISITA2014), pp. 284-288, Melbourne, Australia, Oct. 26-29, 2014.

⑤ 岩田賢一, 有村光晴, “多値アルファベット版部分列数え上げデータ圧縮法に対する最悪冗長度,” 電子情報通信学会技術研究報告, No. IT2013-55, pp. 7-12, 名古屋大学, March 10-11, 2014.

⑥ 有村光晴, 岩田賢一, “可算無限アルファベットの情報源に対して VF 符号が存在する条件,” 電子情報通信学会技術研究報告, No. IT2013-9, pp. 41-46, 福井県あわら市, May 24, 2013.

⑦ Mitsuharu Arimura, Hiroki Koga and Ken-ichi Iwata, “Relationships between the Classes of the Redundancy-Optimal and the Rate-Optimal FF Codes for a General Source,” Proc. 8th Asian-European Workshop on Information Theory: Fundamental Concepts in Information Theory (AEW8), pp. 4-8, Kamakura, Kanagawa, Japan, May 17-19, 2013.

⑧ 岩田賢一, 有村光晴, 嶋優希, “Lossless Data Compression via Substring Enumeration のマルコフ情報源に対する最悪冗長度,” 電子情報通信学会技術研究報告, No. IT2012-76, pp. 95-100, 関西学院大学, March 7-8, 2013.

⑨ 有村光晴, 古賀弘樹, 岩田賢一, “冗長度と符号化レートの両方で最適な FF 符号,” 電子情報通信学会技術研究報告, No. IT2012-61, pp. 71-76, 電気通信大学, Jan.

21, 2013.

⑩ Mitsuharu Arimura and Ken-ichi Iwata, “On Variable-to-Fixed Length Coding of A General Source with Infinite Alphabet,” Proc. 2012 International Symposium on Information Theory and its Applications (ISITA2012), pp. 485-488, Honolulu, Hawaii, USA, Oct. 28-31, 2012.

⑪ Ken-ichi Iwata, Mitsuharu Arimura and Yuki Shima, “On the Maximum Redundancy of CSE for I.I.D. Sources,” Proc. 2012 International Symposium on Information Theory and its Applications (ISITA2012), pp. 489-492, Honolulu, Hawaii, USA, Oct. 28-31, 2012.

⑫ **有村光晴**, **古賀弘樹**, **岩田賢一**, “FF **符号における冗長度と符号化レートの関係について**,” **電子情報通信学会技術研究報告**, No. IT2012-2, pp. 7-12, **福岡県飯塚市**, May 25, 2012.

〔図書〕（計　1 件）

① 山本博資, 古賀弘樹, 有村光晴, 岩本貢 訳, 「情報理論 基礎と広がり」, 共立出版, 2012, 580 ページ. (原著：Thomas M. Cover and Joy A. Thomas, Elements of Information Theory, 2nd Ed., John Wiley & Sons, Inc., 2006.)

６．研究組織

(1)研究代表者

有村　光晴（ARIMURA Mitsuharu）

湘南工科大学・工学部・講師

研究者番号：８０３１３４２７