様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　１８　日現在

機関番号：　３２７０６
研究種目：　基盤研究(C)（一般）
研究期間：　2015～2017
課題番号：　１５Ｋ０６０８８

研究課題名（和文）逐次仮説検定および十分統計量との関連に着目したＶＦデータ圧縮法の解析および設計

研究課題名（英文）Analyses and design of variable-to-fixed length data compression algorithms focused on the relationship with the sequentially hypothesis testing and the sufficient statistic

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）　　3,300,000円

研究成果の概要（和文）：本研究では逐次仮説検定および十分統計量との関連に着目したVFデータ圧縮法の解析および設計法の構築を目指した．特に，十分統計量を理論的に拡張し，漸近十分統計量なるものを定義し，これとユニバーサル符号との関連を理論的に調べた．
その結果，Lempel-Ziv符号のバリエーションを含む，ブロックを切り出して符号化する種類のアルゴリズムから漸近十分統計量として分割木を取り出すことができた．さらに，この分割木を用いた二段階符号を構築し，このアルゴリズムがユニバーサルである，すなわち，様々な情報源クラスに対して漸近的に情報源のエントロピーレートを達成することを証明した．

研究成果の概要（英文）：In this research we aim for the construction of the methods of analyses and design of variable-to-fixed length lossless data compression algorithms using some methods of mathematical statistics. Specifically, some relationships between the sufficient statistic and the universality of lossless source codes are investigated.
At first, the sufficient statistic is extended to the asymptotically sufficient statistic, and some asymptotically sufficient statistics are extracted from the existing lossless data compression algorithms. Then two-step algorithms are constructed. In their algorithms, asymptotically sufficient statistics are encoded in the first step. In the second step, the index of the given data sequence in the set of sequences which generates the same statistic as the given data sequence.
The asymptotic optimalities of the two step codes are proved using the zero-rateness and the asymptotic sufficiency of the statistics encoded in the first step of the two-step codes.

研究分野：　情報理論

キーワード：　データ圧縮　十分統計量　ユニバーサル符号

１．研究開始当初の背景

数理統計学の分野では逐次仮説検定や十分統計量と呼ばれる概念が存在する．他方，情報理論の分野ではデータ圧縮のユニバーサル性という性質が存在する．これらは，パラメトリックな情報源すなわち確率過程の集合を与えた時，パラメータに依存しないような統計量や符号化アルゴリズムが存在するという意味で，関連があると考えられる．しかし，このテーマに基づく研究の開始当初の時点では，情報理論的な理論解析は十分と言えるほど存在していなかった．本課題研究を開始する時点で，我々によって，固定長のデータ集合を可変長の符号語集合に変換する FV 符号において，十分統計量を拡張した漸近十分統計量と，ユニバーサルデータ圧縮の関連を示す結果が得られていた．しかし，可変長のデータ集合を固定長の符号語集合に変換する VF 符号について，十分統計量との関連を理論的に解析されてはいなかった．また，VF 符号は，可変長データの確率がほぼ一様になるという意味で，逐次仮説検定と関連があると考えられる．

また，これまでユニバーサルデータ圧縮アルゴリズムが各種提案され，その圧縮性能が評価されてきたが，評価の手法や尺度がばらばらであり，統一的な方法論および尺度に基づいた性能評価が行われていない．これについても，数理統計学のいくつかの概念を用いたデータ圧縮アルゴリズムという，広いクラスの情報源符号を考えることにより，種々の符号に対して統一的な性能評価を行うことができると考えられる．

２．研究の目的

本研究の目的は，逐次仮説検定および十分統計量という数理統計学の概念と，ユニバーサルデータ圧縮アルゴリズムという情報理論の概念の関連づけを行うことである．これにより，情報理論において様々なアルゴリズムの様々な情報源クラスに対して求められてきたユニバーサル性および冗長度の評価を統一的に行うことが可能となる．また，統計量を用いた二段階符号という新しい枠組みの符号を提案することにより，統計学における十分統計量がデータ圧縮のユニバーサル性と理論的にどのように関連づけられているかを明らかにする．これにより，これまで検討されていなかった新しいデータ圧縮アルゴリズムのクラスを提案することで，これまで以上に高性能な圧縮性能を達成することを目的とする．特に，対象とする符号のクラスとして VF 符号を考える．このクラスの情報源符号は，出力される符号語が固定長となることから，この後に暗号化や誤り訂正符号を続けて行う場合にも都合が良いというメリットが存在する．そこで，VF 符号と逐次仮説検定および十分統計量の理論的関連を，符号化アルゴリズムを通して構築することを目的とする．

３．研究の方法

以下のような方法で研究を行なった．まず，既に存在する符号から統計量を取り出した．これは，データの関数として与えられるので，具体的な符号やデータそのものも含まれるが，実際には符号化に使われる中間情報である，データのタイプ(経験頻度分布)や Lempel-Ziv78 符号における分割木および分割数，符号の符号語長などを考える．

次に，これらの統計量を用いた二段階情報源符号を構築する．これは，以下のような手続きである．まず一段階目で，取り出された統計量が符号化される．次に二段階目で，統計量が等しくなるような情報源系列の集合の中で，符号化される系列を指定するための情報が符号化される．

この二段階符号について，統計量の十分性を理論的に拡張した漸近十分性と，統計量のゼロレート性を用いることで符号の漸近最適性および冗長度を理論的に求める．特に，冗長度の解析の際に，これまでの研究においては，個々の符号に依存した組み合わせ論的な方法が用いられてきた．それに対して，本研究で用いる解析では，統計量の十分性を用いることで簡単な数式で冗長度を評価し，漸近最適性を示すことができる．

これらの解析により，ある情報源クラスに対してユニバーサル符号が存在すれば，必ずゼロレートの漸近十分な統計量が存在すること，また逆に，ゼロレートの漸近十分な統計量が存在すれば，それを用いてユニバーサルな二段階符号を構築できることが理論的に示せる．よって，ある情報源クラスに対して，ゼロレートの漸近十分統計量の存在とユニバーサル符号の存在の同値性が示せることになる．

また，逐次仮説検定について，これと VF 情報源符号との理論的関連を示す．これは，VF 符号を逐次仮説検定の一つのアルゴリズムとして見ること，および逐次仮説検定の最適性と VF 符号の最適性が同値であることを理論的に示す．これによって，VF 符号を逐次仮説検定の一つのアルゴリズムとして見ることができること，また，逐次仮説検定問題を応用して VF 符号を設計できることが分かる．

以上のような方法によって，数理統計学における逐次仮説検定や十分統計量などの概念と，情報理論におけるデータ圧縮アルゴリズムのユニバーサル性の理論的関連を示す．これにより，これまでユニバーサルデータ圧縮アルゴリズムを構成する際に考えられていなかった，新しい種類のアルゴリズムのクラスを構築することができ，これまでよりも高い圧縮性能を持つアルゴリズムを提案できる可能性が存在する．また，データ圧縮アルゴリズムのユニバーサル性を，数理統計学の概念を用いて統一的に証明することによって，データ圧縮におけるユニバーサル性の

理解をより深めることを目指す.

４．研究成果

VF 符号として分割数が固定された Lempel-Ziv78 符号を考え, この符号と十分統計量との関係を定理として提示した. これは, 以前に研究代表者らによって行われた, FV 符号と十分統計量の関係を VF 符号に適用したものと見ることができる. 符号から取り出す漸近十分統計量としては符号化する系列の分割が終了した際の分割木を用いた. これにより, VF 符号から抽出される統計量を用いる VV 二段階符号を新しく提案した. VV 符号は VF 符号に比べて符号割り当ての自由度が高くなるため, 圧縮性能を向上できる可能性が存在することを示すことができた.

さらに, この研究の元となっている Tunstall 符号の圧縮性能を平均符号語長の収束および符号語長の概収束符号化定理を示した. 特に, これまでは Tunstall 分割木を用いて１回だけ切り出しを行なった場合の平均符号語長しか求められていなかったが, 本研究では任意の分割木を複数個用いて, それぞれの分割木で複数のブロックを系列から切り出して符号化するという, より通常用いられるアルゴリズムについて性能解析を行い, 複数の分割木それぞれの葉の数の幾何平均が符号化レートのエントロピーへの収束に効いていることが明らかになった. これにより, VF 符号で複数のブロックを切り出して符号化する際に, LZ78 符号のように木を成長させるだけでなく, 様々な木の成長方法が有り得ることが明らかになった.

文脈自由文法のチョムスキー標準形を用いるデータ圧縮アルゴリズムのクラスを新しく定義し, このクラスに含まれるアルゴリズムの圧縮性能について検討した. その結果, LZ78 アルゴリズムおよび MPM アルゴリズムは, 圧縮法として最適ではない部分が存在することが明らかになり, その部分を修正することで圧縮性能を向上させることができた.

部分列数え上げ法によるデータ圧縮アルゴリズムの圧縮性能について理論的解析を行い, k 次マルコフ情報源に対する最悪冗長度の漸近式を求めた.

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 4 件）

① Mitsuharu Arimura, “A Variable-to-Fixed Length Lossless Source Code Attaining Better Performance than Tunstall Code in Several Criterions,” IEICE Transactions on Fundamentals, 査読有, Vol.E101-A, No.1, pp.249-258, Jan., 2018.
DOI: 10.1587/transfun.E101.A.249

② Ken-ichi Iwata and Mitsuharu Arimura, “Lossless Data Compression via Substring Enumeration for k-th Order Markov Sources with a Finite Alphabet,” IEICE Transactions on Fundamentals, 査読有, Vol.E99-A, No.12, pp.2130-2135, Dec., 2016.
DOI: 10.1587/transfun.E99.A.2130

③ Mitsuharu Arimura, “Average Coding Rate of a Multi-Shot Tunstall Code with an Arbitrary Parsing Sequence,” IEICE Transactions on Fundamentals, 査読有, Vol.E99-A, No.12, pp.2281-2285, Dec., 2016.
DOI: 10.1587/transfun.E99.A.2281

④ Mitsuharu Arimura, “Almost Sure Convergence Coding Theorems of One-shot and Multi-shot Tunstall Codes for Stationary Memoryless Sources,” IEICE Transactions on Fundamentals, 査読有, Vol.E98-A, No.12, pp.2393-2406, Dec., 2015.
DOI: 10.1587/transfun.E98.A.2393

〔学会発表〕（計 8 件）

① 有村光晴, 長岡浩司, “情報源近似と漸近十分統計量を用いた強ユニバーサル FV 符号の最悪冗長度の評価,” 電子情報通信学会技術研究報告, 査読無 No.IT-2017-106, pp.19-24, 東京理科大学 葛飾キャンパス, March 8-9, 2018.

② Mitsuharu Arimura, “A Variable-to-Fixed Length Lossless Source Code Optimizing a Different Criterion of Average Coding Rate from Tunstall Code,” Proc. 2016 International Symposium on Information Theory and its Applications (ISITA2016), 査読有, pp.667-671, Monterey, CA, USA, Oct.30-Nov.2, 2016.

③ Mitsuharu Arimura, “A Grammar-Based Compression Using a Variation of Chomsky Normal Form of Context Free Grammar,” Proc. 2016 International Symposium on Information Theory and its Applications (ISITA2016), 査読有, pp.246-250, Monterey, CA, USA, Oct.30-Nov.2, 2016.

④ Mitsuharu Arimura and Hiroshi Nagaoka, “A Two-Step Universal VV Code Using an Asymptotically Sufficient Statistic Extracted from a VF Code,” Proc. 2016 International Symposium on Information Theory and its Applications (ISITA2016), pp.6-10, Monterey, CA, USA, Oct.30-Nov.2, 2016.

⑤ 有村光晴, “文脈自由文法のチョムスキー標準形を用いた文法圧縮アルゴリズム,” 電子情報通信学会技術研究報告, 査読無, No.IT-2016-12, pp.69-74, 北海道小樽市 小樽経済センター, May 19-20,

2016.

⑥ 有村光晴，長岡浩司，“VF 符号から抽出される漸近十分統計量を用いた二段階 VV ユニバーサル符号の構築，” 電子情報通信学会技術研究報告，査読無，No.IT-2016-13，pp.75-80，北海道小樽市 小樽経済センター，May 19-20，2016.

⑦ 有村光晴，“任意の分割木系列に対する multi-shot Tunstall 符号の平均符号化レート，” 第 38 回情報理論とその応用シンポジウム(SITA2015)予稿集，査読無，pp.445-450，岡山県倉敷市，Nov.24-27，2015.

⑧ 有村光晴，“十分統計量と VF 符号のユニバーサル性の関係，” 第 38 回情報理論とその応用シンポジウム(SITA2015)予稿集，査読無，pp.208-213，岡山県倉敷市，Nov 24-27，2015.

〔図書〕（計　0 件）

〔産業財産権〕

○出願状況（計　0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

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

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：

(4)研究協力者
（　　）