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　１４　日現在

機関番号：　１３２０１
研究種目：　挑戦的萌芽研究
研究期間：　2012～2014
課題番号：　２４６５００７３

研究課題名（和文）不可逆圧縮過程における保存量の研究

研究課題名（英文）Invariant properties in an irreversible transformation

研究代表者
村山　立人（Murayama, Tatsuto）

富山大学・大学院理工学研究部（工学）・講師

研究者番号：８０３６０６５０

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：情報の符号化における不可逆圧縮過程を前提にして、そこに保存量が定義できないかどうかを数値的・数学的に検証した。特に、情報理論でよく利用される行列モデルを分析したところ、復号の方法に強い制約条件を課さない限りは、一般的に保存量を構成するのは困難であることが判明した。一方、いわゆる粗視化の方法では、厳密に保存量が存在するような符号化の方法を提案することは容易だが、工学的な立場からの有効性は非常に低かった。最終的には、これに関連した計測と制御の問題について非自明な発見をできたので、その主要成果を国際学会で公表した。

研究成果の概要（英文）：We have conducted preliminary research on possibilities of finding some invariant properties in an irreversible transformation, which is defined by a simple matrix model in information theory. At this point, there seems no general way to state an invariance theorem without imposing a set of strong assumptions for the decoding model itself. In a coarse graining approach, we obtained some exact results but, eventually, they were found out to be quite trivial from the point of view of practical engineering. However, we were able to report related non-trivial findings in an international conference.

研究分野：　情報統計力学

キーワード：　情報理論　統計力学　不可逆過程　計測　保存量

１．研究開始当初の背景

(1) 現代の社会では、より少ない記憶領域で経済的に情報を保存するために可逆圧縮過程が広く利用されている。しかし、データ系列を完全に再現しなければならないこの可逆過程では、エントロピー限界を超える水準で情報を集約することはできない。

(2) 情報理論では、完全なデータ系列の再現が困難な圧縮水準でも、距離公理を満足する自然な歪み測度を忠実度規範とした圧縮過程が議論できる。例えば JPEG、MPEG、MP3 などは、人間の視覚や聴覚の感覚特性、あるいは感覚情報を統合する脳の認識特性を経験的に考慮した忠実度規範を持っていると解釈できる。

(3) 一方、深宇宙探査やゲノム情報処理などのデータマイニング諸分野では、データスケールが人間の扱える範囲を超えており、すでにコンピュータでの処理が常識になっている。このように、データの入力先が人間の感覚器からコンピュータに置き換わった分野では、人間の感覚・認識特性との親和性は相対的に重要ではなくなり、コンピュータでの統計的処理に適した情報集約の方法を再検討する必要がある。

(4) これに加えて、深宇宙探査やゲノム情報処理などのデータマイニング諸分野では、近年の急速な技術革新によって、大容量の計測データがオンデマンドで取得できる恵まれた環境になりつつある。このように、最新の機器によってデータ取集のコストが急速に低下していく状況では、むしろ取得した計測データの維持・管理コストが研究活動の経済性を決定するボトルネック要因となる。そのため、データ解析との親和性が極めて高く、かつ圧縮効率も非常に優れた符号化の技術が提唱できれば、それが潜在的に持つ市場規模は大きいと考えられる。

(5) さらに、数学的に不可逆圧縮過程における保存量を考えるという立場の研究としては、その「保存量」として何を考えるのかという自由度が大きい。本研究では各種の基本的な統計量、特に確率分布の推定という枠組での十分統計量の保存を分析の対象に限定しているが、学術的意味ではその他のバリエーションも数多く定義できる。例えば、計測データの非自明な対称性をマイニングして保存するためには群論的アプローチは極めて有効だと思われるし、保存量が何かに関わらずその個数自体も重要な視点を与える。このように本研究が成功した場合には、それを手掛かりにして、さらに新しい卓越した成果が得られると期待できる。

２．研究の目的

(1) 本研究の目的は、深宇宙探査やゲノム情報処理などのデータマイニング諸分野で、特定の事象の検出を主目的に日々大量に蓄積されている計測データの効率的な情報集約のための方法論の確立である。特に、本研究では、計測データが離散時間の確率過程、あるいは確率変数の列として解釈できるとき、そこで定義される基本的な統計量を保存させる不可逆圧縮過程による情報集約の形式を提唱する。これによって、もとの計測データを特徴づける統計的性質を最大限に継承させながら、同時に圧縮過程で時系列が受ける情報理論的意味での損失を最小限に抑える理想的な不可逆データ圧縮技術を提供することができる。

(2) 完全なデータ系列の再現性を保証する可逆圧縮過程は、そのデータ系列を特徴づける任意の統計量を保存させることができる圧縮過程である。この見方をすると、エントロピー限界を超えて情報を縮約する不可逆圧縮過程とは、特定の限定された統計量だけを保存させることを可能にする圧縮過程だと解釈できる。このように、計測データの情報縮約を目的に、データ系列を特徴づける統計量を保存させる不可逆圧縮過程に着目した研究は見当たらない。さらに、不可逆過程において保存量を考えるという物理学的思考を情報理論分野に持ち込んでいるという意味でも、学術的に興味深い内容になっていると考えられる。

(3) 本研究におけるデータ系列は、特定の確率分布に従って繰り返し生成される確率変数の集合だと仮定する。そして、このデータ系列の背後にある確率分布を特徴づけるパラメータを確率変数の実現値から推定する逆問題を、一般的なデータマイニングのモデルケースとして考える。本研究では、パラメータの十分統計量を保存させる不可逆圧縮過程をデータ系列に適用し、古典的な歪み測度の最小化との両立性を情報統計力学・タイプ理論等を援用して分析する。

(4) 不可逆データ圧縮過程の実践的研究では、数学的に定義された歪み測度による忠実度規範を陽には考えず、むしろ人間の感覚・認識特性を主観的に反映した情報縮約の方法の発見が主流になっている。本研究では、情報理論の原点に立ち戻り、数学的に定義できる歪み測度に基づく忠実度規範を不可逆圧縮過程に導入する。これによって、計測データを特徴づける特定の統計量が保存する圧縮過程を設計するときに、データ系列がこの不可逆過程によって失う情報量の大きさを定量的に把握し、その最小化との両立性を数学的に議論できるようにしていく。

(5) 本研究は、古典的な情報縮約と統計的推定を現代的なデータマイニングという視点で融合させる。結果として、レート・歪み関

数を拡張した圧縮限界の表現が得られると予想できるが、これは情報理論の新しい適用先が開拓できたという学術的意義があり、これが本研究の究極的な目的である。

３．研究の方法

(1) 本研究では、実践的なデータマイニングの諸分野で、データ解析の対象になる統計量を保存させる不可逆データ圧縮技術による情報集約を提唱する。ただし、実際に計測されるデータの種類や計算の対象となる諸統計量は代表的なものに限っても非常に多岐にわたる。そこで、本研究では計測データの生成モデルを仮定し、データとは特定の確率分布に従って繰り返し生成される確率変数の集合だと解釈する。そして、このデータ系列の背後にある確率分布を特徴づけるパラメータを確率変数の実現値から推定する逆問題を、一般的なデータマイニングのモデルケースとして分析する。具体的には、パラメータの十分統計量を保存させる不可逆圧縮過程をデータ系列に適用し、古典的な歪み測度の最小化との両立性を情報統計力学・タイプ理論等を援用して分析する。

(2) 計測データの収集は時間的・空間的に非常に広範囲にわたって断続的に行われることも多く、そのため当初の目的とする統計量の計算のために有効なデータが事後的あるいは逐次的に追加されていくことが予想される。本研究が提唱する不可逆圧縮過程では、データ系列はその圧縮水準に対応した情報損失を被ることになるが、データ処理の計算目的になる統計量はいつでも保存される。そのため、追加データによる統計量の再計算は数学的に全く同じ手順のアルゴリズムに帰着する。このように、事後的なデータ処理に対する親和性も考慮した圧縮方法の検討を中心課題にして、情報理論的な分析を進める。

(3) 初年度は、十分統計量を保存させる代償としてレート・歪み関数が被るペナルティーを定量的に分析する。つまり、通常の圧縮限界に比較して、歪み測度の期待値がどの程度大きくなるのかを最も簡単なモデルであるベルヌイ分布のパラメータ推定問題等で評価する。コンピュータによる数値的評価と数学的手法による理論的評価を組み合わせる。次年度以降は、研究の進捗状況によってアプローチを検討する。

４．研究成果

(1)数値的な知見としては、情報理論では代表的な系列モデルであるベルヌイ系列及びマルコフ系列について、十分統計量を保存させるような変換を発見法的に構成し、これが不可逆圧縮の過程として利用価値があるかどうかを検証した。その結果、ベルヌイ系列については、十分統計量を保存させるような変換自体は自明な方法でもかなりの数が構成できるが、マルコフ系列については、その構成自体が非常に困難であった。つまり、変換則をランダムに与えても、圧倒的な確率でこれは要請を満たさないので、数値的な検証すら難しい状況であった。

(2)ベルヌイ系列で十分統計量を保存させるような変換則のうち、不可逆圧縮過程として意味があると考える方法の多くはいわゆる「粗視化」と呼ばれるメカニズムで説明ができる符号化に分類できた。これは、ランダムに変換則を与えた当然の帰結とも言えるが、非自明な構成で十分統計量を保存させることが極めて困難であることが強く示唆される結果である。

(3)上記の結果を出したプログラムで、センシングと符号化の有名問題を再考したところ、非常に興味深い現象を発見した。つまり、ネットワーク全体が利用できる通信帯域に上限があり、非常に多数の計測デバイスで同一の対象物からの信号を同時に推定するとき、計測ノイズの関数として、最適な圧縮率が一次転移するのである。一次転移とは、関数をグラフに書いた時に、どこかで非連続にジャンプしていることを意味する。このような現象の報告は初めてであり、大偏差統計の理論による解析的な説明と合わせて国際会議で口頭発表を行った。

(4)本研究では、データを圧縮するとき、データ自体ではなくそれを特徴づける確率分布の統計量（特に十分統計量）に注目をしている。このような立場は、古典的な意味での統計学では常識だが、これを明示的に不可逆圧縮過程と組み合わせて議論している研究者は少ない。しかし、このアプローチは、時系列の信号処理のパラダイムでは非常に有効に機能すると考えられるので、この研究の内容を発展させる方向で新しい研究テーマを提案している。現在、採択されている研究課題としては、新学術領域研究（研究領域提案型）の課題番号 26120516 がこれに相当する。こちらに関連した研究成果については、当該領域の公式ページなどを参照されたい。

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）

〔学会発表〕（計１件）

T. Murayama, K. Okino, M. Tajima, P. Davis, ``Aggregation Principle for Independent Noisy Observations: A Scaling-law Perspective,'' 2nd Korea-Japan Joint Workshop on Complex Communication Sciences
(KJCCS'13, Oct 19, Okinawa)

〔その他〕

ホームページ等
(1)ResearcherID
http://www.researcherid.com/rid/E-7575-2012

(2)My Citations
https://scholar.google.com/citations?hl=en&user=Ys0focEAAAAJ

(3)新学術領域研究「スパースモデリングの深化と高次元データ駆動科学の創成」
http://sparse-modeling.jp/

６．研究組織
(1)研究代表者
村山　立人（MURAYAMA, Tatsuto）
富山大学・大学院理工学研究部（工学）・講師
研究者番号：　８０３６０６５０