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　 ５ 月 ２２ 日現在

機関番号：１１３０１
研究種目：基盤研究(B)
研究期間：2011～2014
課題番号：２３３０００５１

研究課題名（和文）データ圧縮に基づく知識発見の理論と応用に関する研究

研究課題名（英文）Knowledge discovery based on data compression: a study in theory and practice

研究代表者
篠原　歩（SHINOHARA, AYUMI）

東北大学・情報科学研究科・教授

研究者番号：００２２６１５１

交付決定額（研究期間全体）：（直接経費）　　10,300,000円

研究成果の概要（和文）：知識発見の原理の究明と実働化を目指して，データ圧縮技術の関連に着目しながら，理論と応用の両面から研究を展開した．圧縮したままのデータ処理，文字列照合，文字列の繰り返し構造に関する組み合わせ論，マルチエージェントシステム，例から概念を同定する問題の計算量，強化学習，ゲームの解析，実問題への応用などに関する一連の成果を得た．

研究成果の概要（英文）：We study various topics concerning with data compression and knowledge discovery, from both theoretical and practical points of view. We made several contributions on compressed string processing, string matching, combinatorial properties on the repetitive structures in strings, multi-agent system, computational learning theory, reinforcement learning, game analysis, and practical applications.

研究分野：情報科学

キーワード：データ圧縮　機械学習　人工知能　アルゴリズム

１．研究開始当初の背景

通信技術と計算機の飛躍的な発展に伴い，数百ギガバイトから数テラバイトにも及ぶデータが実験・観測によって収集・蓄積され，また，自動計測やデータの自動収集技術等の発展により，各種の企業等においても，同様に巨大な情報が蓄積されている．このような巨大な情報を対象にして，科学的な仮説や知識、意志決定に必要な基準等を発見する効率的な手法の開発が強く求められている．蓄積された膨大なデータの中から，ユーザにとって真に必要なものを効率よく抽出する技術の開発が渇望されている．また，さまざまな機能を持ったロボットが登場し，真に自律的な動作を行える知能システムへの期待がますます高まっている．こうした状況を背景に，大量のデータから有用な知識を発見する技術の開発には，実用的な成果が期待されている．一方，データ圧縮は，通信やデータ保存のコストが極めて高価であった計算機科学の黎明期から，高速大容量の通信とデータの電子化が日常的になった今日に至るまで，最も重要な情報技術の一つとして深くかつ幅広く研究されている．

我々は，計算量理論に基づいた機械学習の理論を初期のテーマとして研究活動に着手し，その発展型としての機械発見の研究にも継続的に取り組んできた．また，データ圧縮技術ならびにそれを利用した様々なデータ処理の効率化に取り組んで来た．さらに，我々は自律移動ロボットによる競技に継続的に参戦してきたが，この開発の過程において，知的に振る舞うべく実ロボットに期待される知識処理の膨大さと，現実のハードウェアに実装して処理できる作業内容との大きなギャップを痛感していた．競技に勝つための手先のチューニングに囚われすぎず，自律ロボットが真に自律的な動作を行えるようにするために必要な原理を根本から再考したいという動機付けがあった．

２．研究の目的

本研究は，知識発見の原理の究明と実働化を目指して，特にデータ圧縮技術との関連に着目しながら，理論と応用の両面から研究を展開することを目的とする．特に，知識発見を，データ縮約のプロセスであるとみなすことによって，非可逆的なデータ圧縮としてとらえ，定式化することによって，その原理を明らかにし，またこれまでに蓄積してきたデータ圧縮技法を効果的に適用することによってその実用性を検証していくことを主たる目標とする．　そのために，データ圧縮や学習に関連した種々の問題に対して，理論的な解析と効率の良い解法を与えることを目指す．そして現実の問題にも適用し，その効果を検証する．

３．研究の方法

(1) データ圧縮技法とその応用：圧縮されたデータを陽に展開することなく，パターン照合や処理を行う，圧縮文字列処理の研究を推し進めた．

(2) 文字列の多重集合であるマルチトラック文字列を対象として，トラック間の置換を許した順列パターン照合を新たに考察した．これは複数のセンサーデータ系列に内在するパターンの検出をモデル化したものである．置換によってトラック数 N に関して N!通りの組み合わせ爆発が起こるため，素朴な方法では応用例で要求される大きさのNに対して全く歯が立たない．この照合を高速に行うためのアルゴリズムとデータ構造の開発を押し進めた．

(3) 文字列に含まれる連の数の解析：圧縮しやすい文字列には，繰り返し構造が多く含まれているが，そもそも文字列には，どれだけ多くの繰り返しが含まれうるのか．この根本的な疑問に端を発して，文字列の連の数を数学的に解明しようという研究が近年盛んになっている．連(run)とは，文字列に含まれる 2 回以上の繰り返しで，それ以上，左にも右にも延長できないものをいう．長さ n の文字列に含まれる連の最大数 $\rho(n)$について，その下界と，連の数の平均数についての解析を行った．

(4) 例から概念を同定する機械学習について，決定性有限オートマトンを同定する問題の複雑さの解析に取り組んだ．また，正の例と負の例の個数が著しく異なるクラス不均衡データからの学習に対する効率のよい学習アルゴリズムの開発に取り組んだ．

(5) 強化学習の枠組みにおいて，エージェントと環境の間の通信の遅延が学習の効率に与える影響を調べた．また，ある学習領域で得られた知見を類似した他の学習領域に適用する転移学習を理論的に考察した．

(6) ゲームの解析：完全情報ゲームに対する必勝法を理論的に解析すると共に，計算機を併用して検証する手法の開発に取り組んだ．また，パズルを解く問題の計算量と，それを効率よく解くための方法についての考察を行った．

(7) 実問題への適用：大量のデータからのパターン発見問題，自律ロボットの制御，ET ロボコンへの応用，ゲームアルゴリズムなど，これまでに具体的に取り扱ってきた経験と知見を活かしながら，その実用性を検証した．

４．研究成果

(1) 直線的プログラムと呼ばれる文法を用いて指数的に圧縮された文字列に対して，入力長に対する多項式時間で動作する種々の圧縮文字列処理アルゴリズムを開発した．これは，文法を陽に展開することなく，すべての連と平方，ギャップ付きの回文を検出することができるものである．これらの問題は，入力が通常の文字列であればいずれも効率よく解ける問題であるが，指数的に圧縮した文字列に対しては，それを展開するだけで指数時間かかってしまうため，文字列の反復性に着目した巧妙な処理によってこれを克服した．

(2) 文字列の多重集合としてのマルチトラック文字列について，まずオートマトンを用いた効率のよいアルゴリズムを提案し，その計算量を解析した．また，データベースに蓄えたマルチトラックデータに対する検索に有用な索引構造として，マルチトラック接尾辞木とマルチトラックポジションヒープを提案し，その構築アルゴリズムを示した．さらに，通常の文字列に対して文字の置換を許したパラメータ化照合に対してもポジションヒープが定義でき，それを効率よく構築できることを示した．またマルチトラックポジションヒープを縮約したものを提案した．これは長さｎの時系列データＮ本からなるマルチトラックに対して O(n)領域に収まるデータ構造であり，O(nN)領域を必要としていた既存の索引構造よりもはるかに省メモリである．これらのデータ構造についての理論解析とともに，実データに対する実験を行い．その効果を検証した．また近似照合を行うためのデータ構造も新たに提案し，計算機実験によってその挙動を確かめた．

(3) 文字列に内在する繰り返し構造である連について，連を多く含む文字列を生成する準同形写像の探索を行った．その結果，連の個数に関しては，これまでに知られていた最良の下界と厳密に一致する下界を与える文字列を生成する，より簡潔な準同形写像を見つけ出すことができた．さらに連の指数和の最大数については，既知の物よりも真によい下界を与える準同形写像を見つけ出すことができた．一方，環状に両端の繋がった構造を持つ文字列に対しても，連の平均数を厳密に表す閉じた数式を導出することに成功した．

(4) 決定性有限オートマトン(DFA)を正の例と負の例から同定する問題について，接頭辞集合に対するDFAの最小無矛盾問題の計算量を解析した．DFA の最小無矛盾問題は，入力例に矛盾しない状態数最小の DFA を見つける問題であり，計算学習理論において学習可能性との深い繋がりが指摘されている．既存研究において，一般の入力に対する困難性は知られていたが，本研究では，対象を接頭辞集合，すなわち入力のすべての文字列がある文字列の接頭辞になっている場合に限定してもやはりこの問題が NP 困難であり，近似精度を保証した多項式時間アルゴリズムを構築することも困難であることを証明した．
また，不均衡データからの学習について，経験的カーネルを用いることで既存手法に比べて精度を同程度に保ちながらより高速に処理できる方法を提案し，その効果を計算機実験によって検証した．

(5) 強化学習における遅延の影響について，遅延の大きさが既知の場合と未知の場合それぞれについて効果的に働く学習アルゴリズムを提案し，その効果を検証した．また，転移学習について，サンプル量の削減が原理的には可能であることを証明し，それを裏付ける実験結果を得た．

(6) 完全情報ゲームである一般化三並べの変種として，通常は交互に石を１つずつ盤面に置いていくが，これを一手にｐ個ずつ石を置くという拡張を提案し，さまざまな条件下における必勝法の有無などに関する網羅的な解析を行った．また盤面をトーラス状にしたときのゲームの勝敗も解析した．これらについて数学的な解析と計算機による実証を並行して行った．
また，覆面算というパズルについての考察も行った．覆面算を一般化した問題はNP 困難であることを示し，一方で覆面算を網羅的に生成して解析するためのツールとして有限オートマトンを用い，その挙動を調べた．

(7) 現実の問題への応用に関しては，自律型ロボットのプログラミングに関して，実データの収集や強化学習を自動で行うための補助ツールとして我々が構築していたものを改良し，長時間に渡って人手を介することなく処理が行えることを実証した．

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　６件）

1. Tomohiro I, Wataru Matsubara, Shunsuke Inenaga, Hideo Bannai, Masayuki Takeda, Kazuyuki Narisawa, Ayumi Shinohara, “Detecting Regularities on Grammar-Compressed Strings”, Information and Computation, 査読有り, Vol. 240, 2015, 74-89.
   DOI: 10.1016/j.ic.2014.09.009
2. ディプタラマ, 成澤和志, 篠原歩, “一般化三並べの拡張：一石 p 手”, 情報処理学会論文誌, 査読有, Vol. 55, No. 11, 2014, 2344-2352.
   http://id.nii.ac.jp/1001/00106953/
3. 本田耕一, 八鍬友貴, 成澤和志, 篠原歩, “一般化三並べの拡張：目標動物の組み合わせ”, 情報処理学会論文誌, 査読有, Vol. 55, No. 11, 2014, 2336-2343.
   http://id.nii.ac.jp/1001/00106952/
4. Kazuhiko Kusano and Ayumi Shinohara, “Average Number of Occurrences of Repetitions in a Necklace”, Discrete Applied Mathematics, 査読有, Vol. 163, 2014, 334-342.
   DOI: 10.1016/j.dam.2013.05.019
5. Kosuke Bannai, Kazuyuki Narisawa, Ayumi Shinohara, “Similarity Measure using Lossy Compression and its Application to Image Retrieval”, The GSTF International Journal on Computing, 査読有, Vol. 1, No. 3, 2011, 45-50.

〔学会発表〕(計 35 件)

1. Yusuke Sato, Kazuyuki Narisawa, Ayumi Shinohara, “A Simple Classification Method for Class Imbalanced Data using the Kernel Mean”, The 6th International Conference on Knowledge Discovery and Information Retrieval, 2014 年 10 月 22 日, ローマ (イタリア)
2. Takashi Katsura, Yuhei Otomo, Kazuyuki Narisawa, Ayumi Shinohara, “Position Heaps for Permuted Pattern Matching on Multi-Track String”, The 41th International Conference on Current Trends in Theory and Practice of Computer Science, 2015 年 1 月 24 日, ペプポド・シュネツコ (チェコ)
3. Kazuya Yaguchi, Naoki Kobayashi, Ayumi Shinohara, “Efficient Algorithm and Coding for Higher-Order Compression”, Data Compression Conference 2014, 2014 年 3 月 27 日, ソルトレイクシティー (アメリカ)
4. Kouta Oguni, Kazuyuki Narisawa, Ayumi Shinohara, “Reducing Sample Complexity in Reinforcement Learning by Transferring Transition and Reward Probabilities”, The 6th International Conference on Agents and Artificial Intelligence, 2014 年 3 月 6 日, アンジェ (フランス)
5. Tomoki Komatsu, Ryosuke Okuta, Kazuyuki Narisawa, Ayumi Shinohara, “Bounded Occurrence Edit Distance: A New Metric for String Similarity Joins with Edit Distance Constraints”, The 40th International Conference on Current Trends in Theory and Practice of Computer Science, 2014 年 1 月 27 日, ノビー・スモコベック (スロバキア)
6. Kazuhiko Kusano, Kazuyuki Narisawa, Ayumi Shinohara, “On Morphisms Generating Run-Rich Strings”, The Prague Stringology Conference 2013, 2013 年 9 月 2 日, プラハ工科大学 (チェコ)
7. Kaori Ueno, Shinichi Shimozono, Kazuyuki Narisawa, Ayumi Shinohara, “On the hardness of approximating the minimum consistent DFA from prefix samples”, ICALP2013 Satellite workshop on Learning Theory and Complexity, 2013 年 7 月 7 日, ラトビア大学 (ラトビア)
8. Takashi Katsura, Kazuyuki Narisawa, Ayumi Shinohara, Hideo Bannai, Shunsuke Inenaga, “Permuted Pattern Matching on Multi-track Strings”, The 39th International Conference on Current Trends in Theory and Practice of Computer Science, 2013 年 1 月 28 日, シュピンドレルール・ムリーン (チェコ)
9. Kazuhiko Kusano, Kazuyuki Narisawa, Ayumi Shinohara, “Computing Maximum Number of Runs in Strings”, The 19th International Symposium String Processing and Information Retrieval, 2012 年 10 月 21 日, カルタヘナ (コロンビア)
10. Junya Saito, Kazuyuki Narisawa, Ayumi Shinohara, “Prediction for Control Delay on Reinforcement Learning”, Special Session on Machine Learning, The 4th Inter. Conf. on Agents and Artificial Intelligence, 2012 年 2 月 6 日, ヴィラモウラ (ポルトガル)

６．研究組織

(1)研究代表者

篠原　歩 (SHINOHARA, AYUMI)
東北大学・大学院情報科学研究科・教授
研究者番号：00226151

(2)研究分担者

成澤　和志 (NARISAWA, KAZUYUKI)
東北大学・大学院情報科学研究科・助教
研究者番号：40583323