様式 C-19

科学研究費補助金研究成果報告書

平成２４年　５月１７日現在

機関番号：１１３０１
研究種目：基盤研究（B）
研究期間：2008～2010
課題番号：20300052
研究課題名（和文）　データ圧縮に基づく知識発見の研究

研究課題名（英文）　A study on knowledge discovery based on data compression

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

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した．知識発見を非可逆的なデータ圧縮としてとらえて新たな類似性の指標を導入し，その性質を調べた．また，圧縮したままのデータ処理，文字列の連の数に関する組み合わせ論，形式言語理論，マルチエージェントシステム，計算学習理論，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 studied various topics concerning with data compression and knowledge discovery, from both theoretical and practical points of view. We proposed a new similarity measure based on lossy compression, and analyzed its properties. We made several contributions on compressed string processing, combinatorial properties on the number of runs in strings, formal language theory, multi-agent system, computational learning theory, game analysis, and practical applications.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	2,400,000	720,000	3,120,000
2009 年度	2,100,000	630,000	2,730,000
2010 年度	2,100,000	630,000	2,730,000
年度			
年度			
総　計	6,600,000	1,980,000	8,580,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：知識発見とデータマイニング，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

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，通信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．データマイニングを主キーワードとして国際会議や研究会も数多く開催され，多くの研究者が精力的に取り組んでいる．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サッカーを題材としたロボカップ４足ロボットリーグにチームを結成して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．試合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，ユーザの関心に依存したフィルタリングと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． さらに，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．

３．研究の方法

(1) コルモゴロフ複雑性に基づく類似性指標：データ圧縮の基本的なアイディアは，データに内在する繰り返しや冗長性を検出し，それを簡潔に表現し直すことで表現長を短くするというものである．すなわち，繰り返しや冗長性が多いほど高い圧縮が行える．逆にいうと，高い圧縮が行えるデータには，繰り返しや共通部分など，類似したものが多く含まれていることになる．ここに着目し，データ圧縮のしやすさをデータ間の類似性を測る尺度として活用する方法が知られている．この方法には，コルモゴロフ複雑性の理論による裏付けがあり，また種々の応用例も示されている．既存研究においては，これらはすべて可逆圧縮アルゴリズムによるものであったが，本研究においては非可逆圧縮にも適用できるかどうかを検証し，実際の画像データなどに適用してその効果を探った．

(2) 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．

(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，$\rho(n)<n$ という予想がなされているが，証明は与えられていない．この $\rho(n)$ の下界と，連の数の平均数についての解析を行った．

(4) 形式言語理論について，特に基本形式体系(EFS)を用いた言語に関する性質を調べた．EFS は，計算学習理論においても，学習可能性をクラス分けするために好都合な性質を持っていることが知られている．

(5) マルチエージェントシステム：複数のエージェントから構成されるシステムによる問題解決について，特に通信規約の学習問題と，探索問題についての検証を行った．

(6) 計算学習理論：学習の容易性・困難性を計算量理論に基づいてアプローチする計算学習理論において，特に教師側の立場からの視点として，教示の理論の研究を推進した．特に，教示に用いる例の個数に制限があった場合に挙動を解析した．

(7) 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，不完全情報ゲームに対して効果的に働くアルゴリズムの研究を行った．

(8) 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，ロボカップや 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) コルモゴロフ複雑性に基づく類似性の指標を，画像の類似性指標として用いた場合の効果と問題点を検証した．圧縮方法によって効果が大きく違うことが確かめられた．その結果，特にフラクタル圧縮がこの方式と相性がよいことがわかった．また，既存の理論を画像や音声データで主に用いられる非可逆圧縮にも適用できるように拡張した．

(2) 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，文法を用に展開することなく，①回文構造の検出，②スクエア構造の検出，③最長共通部分文字列の計算，④非平方性の検証，を行う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線形時間で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．

(3) 文字列に内在する繰り返し構造である連について，文字列に包含されうる連の平均数を解析し，それを厳密に表す閉じた数式を導出することに成功した．また，探索的な手法によって，連を多く含む文字列を計算機実験によって発見し，その観察に基づいた数学的な解析によって，連の最大数の下限をこれまでに知られていたものから大幅に更新することができた．さらに，探索的な手法とビット演算を駆使した効率のよい実装により，連を多く含む文字列を計算機実験によって発見する手法を示した．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．またこれまでは主として文字種が 2 の場合について，連の個数の解析がなされてきたが，我々は文字種が 3 以上になったときにそれがどう影響を受けるかについての考察も行った．また，連を多く含む文字列を計算機で効率よく求めるという課題に関する計算量理論的な問題として，連に関する情報から元の文字列を推測する逆問題の NP 困難性を証明した．さらにこの問題は，バイナリ文字列に対しては効率良く解けることを示した．

(4) 基本形式体系(EFS)で表現される言語の部分クラスにおいて，自己組織化と呼ばれる連接に類似した操作の閉包性を示した．また，EFS に非終端記号を導入することによって，記述力がどう変化するかについての考察をおこない，また既存の形式言語との対応関係を証明した．

(5) マルチエージェントシステムにおける通信規約学習に関して，学習に必要なメッセージのサイズに関する理論的な証明と計算機実験を行った．また，代表的な探索アルゴリズムである A*を状況の変化に効率よく追随させる拡張として知られている適応 A*について，その目標が複数ある場合に対してこれを一般化し，その理論保証を与えると共に，計算機実験によってその効果を確認した．

(6) 計算論的な枠組みにおける教示の理論において，教示に用いる例の個数を制限すると，教示の可能性が大きく変化することがわかった．すなわち，矛盾を含む例を含んだ教示を行う方が，無矛盾な例のみを用いた教示よりも教示の効率がよくなる場合があることを，具体的な概念クラスを用いて証明した．また，例数を制限したときには，通常とは異なる教示戦略が必要となることを示す定理を証明した．

(7) 完全情報ゲームである一般化 3 並べの変種として，目標とする型の複数の OR をとるゲームを提案し，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．さらに，目標とする型と共に「禁止型」を新たな条件として加えたゲームを提案し，目標型と禁止型のさまざまな組み合わせに対する必勝法の有無に関する網羅的な解析と計算機による実証を行った．一方，不完全情報ゲームである大富豪・大貧民ゲームに対して，モンテカルロ法に基づいた局面評価法を提案し，この手法が極めて有効であることを競技会において実証することができた．

(8) 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，シミュレーションによる仮想環境と，実際にロボットが動く実環境とを融合した

拡張仮想現実環境を構築した．このことにより，ロボットの学習において，人手の介在する手間を大幅に削減するとともに，またロボットの可動部分の消耗を減らし，かつ学習効率を上げることができるようになった．また，ETロボコンの競技機体に対して，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を作成し，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確認した．

５．主な発表論文等

〔雑誌論文〕（計５件）

1. W. Matsubara, S. Inenaga, A. Shinohara, “An Efficient Algorithm to Test Square-Freeness of Strings Compressed by Balanced Straight Line Programs”, Chicago Journal of Theoretical Computer Science, 査読有, Article 7, 2010, 1-19.
2. K. Kusano, W. Matsubara, A. Ishino, A. Shinohara, “Average Value of Sum of Exponents of Runs in a String, International Journal of Foundations of Computer Science”, 査読有, Vol.20, No.6, 2009, 1135-1146
3. H. Kobayashi, T. Osaki, T. Okuyama, J. Gramm, A. Ishino, A. Shinohara, “Development of an interactive augmented environment and its application to autonomous learning for quadruped robots”, IEICE Transactions on Information and Systems, 査読有, E92-D(9), 2009, 1752-1761
4. W. Matsubara, S. Inenaga, A. Ishino, A. Shinohara, T. Nakamura, K. Hashimoto, “Efficient algorithms to compute compressed longest common substrings and compressed palindrome”, Theoretical Computer Science, 査読有, 410, 2009, 900-913
5. 小林 隼人, 畑埜 晃平, 石野 明, 篠原 歩, 間引き：ロボットのスキル発見における評価の削減手法, 人工知能学会論文誌, 査読有, Vol. 24, No. 1, 2009, 191-202

〔学会発表〕（計35件）

1. W. Matsubara, A. Ishino, A. Shinohara, “Inferring strings from runs”, The Prague Stringology Conference 2010, 2010年9月10日, プラハ（チェコ）
2. K. Kusano and A. Shinohara, “Average Number of Runs and Squares in Necklace”, The Prague Stringology Conference 2010, 2010年9月10日, プラハ（チェコ）
3. K. Matsuta, H. Kobayashi, A. Shinohara, “Multi-Target Adaptive A*”, The 9th International Conference on Autonomous Agents and Multi-Agent Systems, 2010年5月13日, トロント（カナダ）
4. T. Kasai, H. Kobayashi, A. Shinohara, “The Size of Message Set Needed for the Optimal Communication Policy”, The 7th European Workshop on Multi-Agent Systems, 2009年12月18日, アイアナパ（キプロス）
5. H. Kobayashi, A. Shinohara, “Complexity of Teaching by a Restricted Number of Examples”, The 22nd Annual Conference on Learning Theory, 2009年8月31日, モントリオール（カナダ）
6. T. Kasai, H. Kobayashi, A. Shinohara, “Improvement of the Performance by using Received Message on Learning of Communication Codes”, 8th International Conference on Autonomous Agents and Multi-Agent Systems, 2009年5月15日, ブダペスト（ハンガリー）
7. W. Matsubara, K. Kusano, H. Bannai, A. Shinohara, “A Series of Run-rich strings”, The 3rd International Conference on Language and Automata Theory and Applications, 2009年4月7日, タラゴナ（スペイン）
8. W. Matsubara, S. Inenaga, A. Shinohara, “Testing Square-Freeness of Strings Compressed by Balanced Straight Line Program”, 15th Computing: The Australasian Theory Symposium, 2009年1月20日, オークランド（ニュージーランド）
9. K. Kusano, W. Matsubara, A. Ishino, A. Shinohara, “Average Value of Sum of Exponents of Runs in Strings”, The Prague Stringology Conference 2008, 2008年9月3日, プラハ（チェコ）
10. W. Matsubara, K. Kusano, H. Bannai, A. Ishino, A. Shinohara, “New Lower Bounds for the Maximum Number of Runs in a String”, The Prague Stringology Conference 2008, 2008年9月2日, プラハ（チェコ）

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