様式 C-7-2

自己評価報告書

平成２３年　４月１８日現在

機関番号：１１３０１
研究種目：基盤研究（B）
研究期間：2008～2011
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

研究分野：情報科学
科研費の分科・細目：情報学・知能情報学
キーワード：機械学習，人工知能，データ圧縮，知識発見，アルゴリズム，情報基礎

１．研究計画の概要

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．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．関連する種々の要素技術を含む，下記の４つの研究項目に力点をおく．

(1) コルモゴロフ複雑性の活用　本申請課題において，知識発見の定式化に関して基盤となるのは，アルゴリズム情報理論であり，そこで用いられるコルモゴロフ複雑性と非圧縮性は，それ自身で重要な理論的ツールとして，計算量理論やアルゴリズム理論，形式言語理論などでの活用例が紹介されている．

(2) データ圧縮技法とその応用　これまでの研究において，我々は，圧縮されたデータを陽に展開することなく，パターン照合したり，繰り返し構造や最長共通文字列を検出するといった，圧縮文字列処理に適した一連のアルゴリズムやデータ構造の開発を行ってきた．本研究においても，それをさらに押し進め，新たな展開を目指す．

(3) 時間限定の条件下での万能人工知能の再定式化　M. Hutter による万能人工知能の定式化は，計算機に無制限の力を許すという仮定がもとになっており，それが故に種々の数学的性質がエレガントに証明されている反面，直接的な実用に結びつかないという欠点があると考えられる．本研究は，計算時間に種々の条件（多項式時間，多項式領域，対数領域など）を課したときの数学的展開を行うことでそのギャップを埋めていく．

(4) 現実の種々の問題への適用と実働化　既存の万能人工知能の定式化は，機械学習，帰納推論や統計的意志決定，ゲーム理論などを統一的に記述しようという意欲的なものであり，それはある程度は成功していると言えるが，反面，上述の通り現実的な応用例とは大きなギャップがある．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て，その実用性を検証していく．

２．研究の進捗状況

(1) コルモゴロフ複雑性の理論に基づいて，類似性を測る指標として有用な正規化情報距離を，画像の類似性に応用した．この際，非可逆圧縮が本質的な働きをするという知見が得られ，これを裏付けるために既存の理論の拡張を行った．

(2) データ圧縮に関しては，データに内在する繰り返し構造をうまく抽出することが重要であるが，文字列の繰り返し構造を抽象化した概念である連(run)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．また，文法を用いて指数的に圧縮された文字列を，陽に展開することなく，回文構造やスクエア構造の検出をしたり，最長共通部分文字列の計算を行ったりする多項式時間アルゴリズムを開発した．また，基本形式体系(EFS)に非終端記号を導入することによる記述力の変化や既存の形式言語理論との対応関係を解明した．

(3) 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．さらに，計算論的な枠組みにおける教示の理論に，例数の制限を導入したときに生じる性質として，矛盾を含む説明が最適教示法になる場合があること，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などを示す定理を証明した．

(4) 現実の問題への応用として，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．

３．現在までの達成度

②おおむね順調に進展している．

特に，データ圧縮や知識発見にとって本質的な，繰り返し構造に関する基礎研究において，大きな発見を成し遂げた．長さn の文字列の中に含まれる繰り返し構造（「連」と呼ぶ）が最大数の下限に関して，それまでに知られていた $0.927n$ を大幅に上回る $0.945n$の文字列を発見し，また平均値の厳密な解析にも成功した．実ロボットに対する知識獲得やエージェント技術，計算学習理論もそれぞれ進展している．また応用例としても，コンピュータによる大貧民や将棋の大会，ロボカップサッカー等でそれぞれ優秀な成績を修めている．

４．今後の研究の推進方策

この3年間の研究において得られた知見に基づいて，研究計画最終点度前年度の応募として申請した基盤研究(B)「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」(H23～H26）として研究を継続する．上述の通り，これまでおおむね順調に研究が展開できているため，大きな計画の変更の必要性や問題点はないと考えている．

５． 代表的な研究成果

（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. 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, Article 4, 2010, pp.1-19, 査読あり
2. K. Kusano, W. Matsubara, A. Ishino & A. Shinohara, *“Average Value of Sum of Exponents of Runs in a String”*, Inter. Journal of Foundations of Computer Science Vol.20, Issue 6, pp.1135-1146, 2009, 査読あり
3. R. Nakamura, S. Inenaga, H. Bannai, T. Funamoto, M. Takeda, & A. Shinohara, *“Linear-Time Off-Line Text Compression by Longest-First Substitution”*, Algorithms, Vol.2, Issue 4, pp.1429-1448, 2009, 査読あり
4. 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 Trans. on Information and Systems. E92-D(9), pp.1752-1761, 2009, 査読あり
5. W. Matsubara, S. Inenaga, A. Ishino, A. Shinohara, T. Nakamura, & K. Hashimoto. *“Efficient algorithms to compute compressed longest common substrings and compressed palindromes”*, Theoretical Computer Science, Vol.410, Issues 8-10, pp.900-913, 2009, 査読あり

〔学会発表〕（計 35 件）

1. W. Matsubara, A. Ishino & A. Shinohara, *“Inferring strings from runs”*, PSC2010, 2010 年 9 月 1 日，プラハ、チェコ.
2. K. Kusano & A. Shinohara, *“Average Number of Runs and Squares in Necklace”*, PSC2010, 2010 年 9 月 1 日，プラハ、チェコ.
3. K. Matsuta, H. Kobayashi, & A. Shinohara, *“Multi-Target Adaptive A*”*, AAMAS2010, 2010 年 5 月 13 日，トロント、カナダ
4. H. Kobayashi & A. Shinohara, *“Complexity of Teaching by a Restricted Number of Examples”*, COLT2009, 2009 年 6 月 21 日，モントリオール、カナダ
5. W. Matsubara, K. Kusano, H. Bannai, & A. Shinohara, *“A Series of Run-rich Strings”*, LATA2009, 2009 年 4 月 7 日，タラゴナ、スペイン