

機関番号：12612
 研究種目：基盤研究（C）
 研究期間：2008～2010年度
 課題番号：20500007
 研究課題名（和文） 形式言語に対する例からの学習を行う効率的アルゴリズムの開発・応用
 研究課題名（英文） Developments of efficient algorithms for learning from examples of formal languages and their applications
 研究代表者
 若月 光夫（WAKATSUKI MITSUO）
 電気通信大学・大学院情報理工学研究科・助教
 研究者番号：30251705

研究成果の概要（和文）： 決定性プッシュダウン変換器のスタック記号を1種類に限定した決定性限定ワンカウンタ変換器について、それが空スタック受理式及び実時間最終状態受理式の場合、その等価性判定が多項式時間でできることを証明した。また、決定性限定ワンカウンタオートマトンのある部分クラス等が、正例から多項式時間で極限同定可能なことを証明した。更に、正則言語の部分クラスに対する正例からの極限同定を利用した、ジュウシマツの歌文法の解析手法を改良し、自動化を図った。

研究成果の概要（英文）： We have developed polynomial time algorithms for checking the equivalence of deterministic restricted one-counter transducers, which are deterministic pushdown transducers having just one stack symbol, that accept by empty stack or by final state. We have proved that a subclass of deterministic pushdown automata called Szilard strict deterministic restricted one-counter automata and a subclass of finite state transducers (FST's for short) called strict prefix deterministic FST's are polynomial time identifiable in the limit from positive data. Furthermore, we have improved the method for analyzing songs of the Bengalese finch using an identification algorithm for the class of k-reversible languages, which is a subclass of regular languages.

交付決定額

(金額単位：円)

	直接経費	間接経費	合計
2008年度	1,200,000	360,000	1,560,000
2009年度	1,100,000	330,000	1,430,000
2010年度	1,200,000	360,000	1,560,000
年度			
年度			
総計	3,500,000	1,050,000	4,550,000

研究分野：総合領域

科研費の分科・細目：情報学・情報学基礎

キーワード：計算論的学習理論，正例からの学習，極限同定，等価性判定，アルゴリズム，決定性文脈自由言語，プッシュダウンオートマトン，プッシュダウン変換器

1. 研究開始当初の背景

人工知能の実現において機械学習は最も重要な研究分野の一つであり，計算論的学習理論はその可能性を数理的で厳密に解析を行うパラダイムである。そのような研究の対象として形式言語を選択し，本研究に先立つ

科学研究費基盤研究(C)「決定性文脈自由言語の部分クラスに対する学習アルゴリズムの開発と応用」においては，形式言語の部分クラスの中で実用上重要な決定性文脈自由言語を受理する決定性プッシュダウンオートマトン(DPDA)またはそれに対応した文法，

あるいは DPDA に出力機構を付与した決定性プッシュダウン変換器 (DPDT) に対し、その構造に妥当な制約を課した幾つかの部分クラスを対象として、以下に述べるような幾つかの研究成果を上げた。

(1) 学習アルゴリズム開発の基礎構築

① DPDA や DPDT が入力を読み込むことなく動作するとき、これらは ϵ -推移を持つと言い、そうでないとき、実時間であると言う。 ϵ -推移を持つある DPDT の部分クラスに対する等価性判定アルゴリズムを開発し、次の論文に発表した。

・清野和司, 富田悦次, 若月光夫: ϵ -推移を許したある決定性プッシュダウン変換器対の等価性判定, 電子情報通信学会論文誌 D, Vol. J90-D, No. 10, pp. 2675-2690 (2007).

これは ϵ -推移を許した DPDT 対に対して、その等価性判定可能な範囲を従来より大きく拡張した最新の成果であり、現時点ではこれを包含する結果は報告されていない。

② DPDT のスタック記号を 1 種類に限定した決定性限定ワンカウンタ変換器 (DROCT) と呼ぶ部分クラスについて、これが実時間でかつ空スタック受理式の場合、上記①に記載した論文のアルゴリズムをこの対象 DROCT 対に特化して単純化することによって、その等価性判定が効率的に行えること、すなわち多項式時間で行えることを次の発表で明らかにした。

・清野和司, 富田悦次, 若月光夫: 実時間空スタック受理式決定性限定ワンカウンタ変換器の多項式時間等価性判定アルゴリズム, 電子情報通信学会コンピュータ研究会, 2007 年 10 月 16 日, 東北大学工学部電子情報システム・応物系 1 号館。

(2) 学習アルゴリズムの開発

① 構造が非常に単純な正則言語の部分クラスを準同型写像で変換して得られるような、ある種の正則言語の部分クラスに対して、正例からの極限同定を統一的に扱える手法を開発し、次の論文に発表した。

・M. Wakatsuki, E. Tomita, G. Yamada: A unified algorithm for extending classes of languages identifiable in the limit from positive data, Lecture Notes in Artificial Intelligence, Vol. 4201, pp. 161-174 (2006).

② 拡張有限オートマトンに出力機構を付与した変換器のある部分クラスに対して、正例からの多項式時間極限同定アルゴリズムを開発し、次の国際会議で発表した。

・M. Wakatsuki, E. Tomita: Polynomial time identification of finite state transducers in some class, AICT 2007, 2007 年 9 月 20 日, Beijing University of Posts and Telecommunications, Beijing,

China.

以上のような研究成果から、これらの手法を更に上位の言語クラスに対する学習アルゴリズムの開発に適用するための手がかりを得た。

2. 研究の目的

前記のようにこれまでに蓄積されてきた実績を最大限活用して、DPDA あるいは DPDT の中の、より広範な部分クラスに対して、計算論的な手法により MAT 学習等の質問を用いた学習や極限同定による学習を行うアルゴリズムを開発する。またそれと並行して、学習アルゴリズム開発の基礎となる等価性・包含性判定アルゴリズムを開発し、時間計算量解析を行って効率化を図る。更に、開発した学習アルゴリズムを適用することによって、ゲノム情報処理等の実際の問題の解決に応用し、本研究手法の有効性を明確にして、前基盤研究(C)の成果を大きく発展させることを本研究の目的とする。

3. 研究の方法

(1) 学習アルゴリズム開発の基礎構築

1. (1)①で述べた DPDT のある部分クラスに対する等価性判定アルゴリズムを、対象とする DPDT 対に特化して単純化することによって、多項式時間で等価性判定を行うアルゴリズムを開発する。

(2) 学習アルゴリズムの開発

① DPDA のスタック記号を 1 種類に限定した決定性限定ワンカウンタオートマトン (DROCA) が、実時間で、かつ各入力記号に対応する推移規則が高々 1 つである場合、Szilard strict DROCA と呼ぶ。この Szilard strict DROCA に対する、正例からの多項式時間極限同定アルゴリズムを開発する。

② 1. (2)①で述べた、準同型写像で変換して得られる言語クラスに対する正例からの極限同定の統一的手法を改良するとともに、正例から多項式時間極限同定可能である新たな言語クラスを明らかにする。

③ 1. (2)②で述べた、有限状態変換器のある部分クラスに対する正例からの極限同定アルゴリズムを拡張し、より上位のクラスに適用できるように改良する。

(3) 学習アルゴリズムの応用

ジュウシマツのさえざり (単に鳥の歌と呼ぶ) の解析は、バウトと呼ばれる一息ごとの連続音声データを単位として扱われる。バウトは幾つかのソングユニットから構成されており、各ソングユニットの時系列データは正則言語の真部分クラスである k -可逆言語として表現できる。我々は、Angluin が提案した k -可逆言語の正例からの極限同定アルゴリズムを組み込んだ、鳥の歌の解析ツール「EUREKA」を開発してきている。バウトから

ソングユニットへの分割は現在、研究者の手作業に頼っているため、この分割処理を自動的に行えるよう、解析システムの改良を図る。

4. 研究成果

(1) 学習アルゴリズム開発の基礎構築

① 入力を読み込むことなく出力が限りなく長く続くような不自然な場合以外は ϵ -推移を許した空スタック受理式 DROCT について、1. (1) ②で述べた実時間空スタック受理式 DROCT 同士の等価性判定アルゴリズムを拡張することによって、その等価性判定が多項式時間で行えることを証明した。

② 上記①と受理方式の異なる最終状態受理式の実時間 DROCT に対し、対象 DPDT の持つ性質を利用することによって、その等価性判定が多項式時間で行えることを証明した。

受理記号列の集合部分が本質的に正則を超えるような変換器に対する等価性判定が多項式的に行えるとの結果は、上記①及び②の結果の他には、次の論文で示されているだけである。

• C. Bastien, J. Czyzowicz, W. Fraczak, W. Rytter, Equivalence of simple functions, Theoretical Computer Science, Vol. 376, pp. 42-51 (2007).

この論文は、我々が対象とした DPDT の部分クラスとは比較不能な関係にある、単純決定性プッシュダウン変換器同士の等価性判定が多項式的に行えることを示しているが、thickness (各スタック記号をポップアップするのに必要とする入力記号列の最短の長さ) をパラメータの一つとして用いており、対象 DPDT の記述長の指数オーダーになり得る点で、我々の結果と対照的である。

上記の等価性判定に関する成果は、対象 DPDT に対する質問による学習に利用できる。

(2) 学習アルゴリズムの開発

① 3. (2) ①で述べた Szilard strict DROCA に対する正例からの極限同定アルゴリズムを開発し、入力記号の種類を定数とみなせば、多項式時間極限同定可能なことを明らかにした。

② 3. (2) ②で述べた準同型写像による変換によって拡張された言語クラスに対して、正例から極限同定を行う統一的手法を改良し、biprefix コードによって拡張された言語クラス等が、正例から多項式時間極限同定可能なことを示した。

③ 一般化順序機械を含む有限状態変換器 (FST) のうち、strict prefix deterministic FST と呼ぶ部分クラスに対して、正例からの極限同定が多項式時間で行えることを証明した。

(3) 学習アルゴリズムの応用

EUREKA における鳥の歌構造の解析には、N グラムと呼ばれる確率的な音声認識モデル

と、k-可逆言語の極限同定アルゴリズムを併用している。N グラムモデルと k-可逆言語との関係を明らかにし、この解析手法の改良を行った。また、バウトからソングユニットを自動的に分割する手法を開発した。

(4) その他

上記に関連して、下記 5. 主な発表論文等に示したように、最大クリーク抽出アルゴリズムやバイオインフォマティクス、量子セルオートマトンを用いた画像圧縮についての有効な研究成果を得た。

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 7 件)

1. Khan Md. M. S., T. Nishino, K. Sasahara, M. Takahashi, K. Okanoya: A feasible approach for automatic detection and recognition of the Bengalese finch songnotes and their sequences, Journal of Intelligent Learning Systems and Applications, Vol. 2, No. 4, pp. 221-228 (2010). 査読有
2. M. Wakatsuki, E. Tomita: Polynomial time identification of strict prefix deterministic finite state transducers, Lecture Notes in Artificial Intelligence, Vol. 6339, pp. 313-316 (2010). 査読有
3. E. Tomita, Y. Sutani, T. Higashi, S. Takahashi, M. Wakatsuki: A simple and faster branch-and-bound algorithm for finding a maximum clique, Lecture Notes in Computer Science, Vol. 5942, pp. 191-203 (2010). 査読有
4. T. Matsunaga, C. Yonemori, E. Tomita, M. Muramatsu: Clique-based data mining for related genes in a biomedical database, BMC Bioinformatics, Vol. 10, No. 205 (2009). 査読有
5. Y. Kakishita, K. Sasahara, T. Nishino, M. Takahashi, K. Okanoya: Ethological data mining: an automata-based approach to extract behavioral units and rules, Data Mining and Knowledge Discovery, Vol. 18, pp. 446-471 (2009). 査読有
6. M. Wakatsuki, E. Tomita: Polynomial time identification of strict deterministic restricted one-counter automata in some class from positive data, IEICE Transactions on Information and Systems, Vol. E91-D, No. 6, pp. 1704-1718 (2008). 査読有
7. 清野和司, 富田悦次, 若月光夫: 実時間

空スタック受理式決定性限定ワンカウンタ変換器の多項式時間等価性判定, 電子情報通信学会論文誌 D, Vol. J91-D, No. 5, pp. 1188-1201 (2008). 査読有

[学会発表] (計 12 件)

1. 中西裕陽, 富田悦次, 若月光夫, 西野哲朗: 最大クリーク問題の多項式時間的可解性について, 2010 年度冬の LA シンポジウム, 2011 年 2 月 3 日, 京都大学数理解析研究所.
2. 若月光夫, 清野和司, 富田悦次, 西野哲朗: 実時間最終状態受理式決定性限定 1 カウンタ変換器の多項式時間等価性判定アルゴリズム, 2010 年度冬の LA シンポジウム, 2011 年 2 月 1 日, 京都大学数理解析研究所.
3. 大畑和樹, 西野哲朗, 若月光夫: 量子セルオートマトンに基づく画像圧縮のための画像変換アルゴリズム, 2010 年度冬の LA シンポジウム, 2011 年 2 月 1 日, 京都大学数理解析研究所.
4. Khan Md. M. S., T. Nishino, K. Sasahara, M. Takahashi, K. Okanoya: Information-theoretic analysis for understanding the behavior of song learning by the Bengalese finch, 情報処理学会数理解析研究会, 2010 年 12 月 16 日, 九州大学伊都キャンパス総合学習プラザ.
5. 鈴木徹, 若月光夫, 西野哲朗: 鳥の歌文法解析の自動化, 2010 年度夏の LA シンポジウム, 2010 年 7 月 20 日, 九段浜温泉 ひみのはな.
6. 中西裕陽, 富田悦次, 若月光夫: 最大クリーク抽出の単純な最大時間計算量評価と多項式時間的可解性, 2009 年度冬の LA シンポジウム, 2010 年 2 月 3 日, 京都大学理学研究科北部構内数学教室.
7. 若月光夫, 清野和司, 富田悦次, 西野哲朗: 空スタック受理式決定性限定ワンカウンタ変換器の多項式時間等価性判定アルゴリズム, 2009 年度冬の LA シンポジウム, 2010 年 2 月 2 日, 京都大学理学研究科北部構内数学教室.
8. 富田悦次, 須谷洋一, 東貴紀, 高橋真也, 若月光夫: 単純でより高速な最大クリーク抽出アルゴリズム, 情報処理学会アルゴリズム研究会, 2010 年 1 月 26 日, 九州大学西新プラザ.
9. 常田宏和, 若月光夫, 西野哲朗: 鳥の歌構造解析における k 可逆オートマトンと N グラムモデルの関係について, 2009 年度夏の LA シンポジウム, 2009 年 7 月 23 日, かんぼの宿 松島.
10. 若月光夫, 富田悦次: 準同型写像によって拡張されたある言語クラスに対する

正例からの極限同定, 情報処理学会数理解析研究会, 2009 年 3 月 5 日, 沖縄科学技術研究基盤機構 OIST Seaside House.

11. 大久保誠也, 西野哲朗, 若月光夫: オートマトンの学習困難性を安全性の基盤とする暗号システムについて, 2008 年度冬の LA シンポジウム, 2009 年 2 月 4 日, 京都大学数理解析研究所.
12. 若月光夫, 高橋真也, 富田悦次: 最大クリーク抽出アルゴリズムの共有メモリ型並列計算機上での並列化, 情報処理学会数理解析研究会, 2008 年 9 月 18 日, 電気通信大学.

[図書] (計 3 件)

1. E. Tomita, T. Akutsu, T. Matsunaga: “Efficient algorithms for finding maximum and maximal cliques: Effective tools for bioinformatics” in “Biomedical Engineering, Trends in Electronics, Communications and Software”, A. N. Laskovski (Ed.), InTech, pp. 625-640 (2011).
2. 富田悦次: “形式言語”, 広中平祐編 (分担執筆)「第 2 版 現代数理解析辞典」, III, 2-1, 丸善, pp. 391-396 (2009).
3. 富田悦次, 横森貴: “オートマトン・言語理論” (第 21 刷・改定増刷), 森北出版, 205 ページ (2009).

[その他]

ホームページ等

<http://kjk.office.uec.ac.jp/Profiles/0001/0000405/profile.html>

<http://www.nishino-lab.jp/laboratory/wakatsuki/index.html>

6. 研究組織

(1) 研究代表者

若月 光夫 (WAKATSUKI MITSUO)

電気通信大学・大学院情報理工学研究所・助教

研究者番号: 30251705

(2) 研究分担者

富田 悦次 (TOMITA ETSUJI)

電気通信大学・名誉教授

研究者番号: 40016598

西野 哲朗 (NISHINO TETSURO) (2009-2010 年度)

電気通信大学・大学院情報理工学研究所・教授

研究者番号: 10198484