様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年　６月１０日現在

機関番号：３２６５７
研究種目：基盤研究（C）
研究期間：2009 ～ 2011
課題番号：２１５００１４８
研究課題名（和文）　確定節文法の漸次学習方式と機械学習への応用

研究課題名（英文）　Incremental Learning of Definite Clause Grammars and Application to Machine Learning

研究代表者
中村　克彦　（NAKAMURA KATSUHIKO）
東京電機大学理工学部・教授
研究者番号：９００５７２４０

研究成果の概要（和文）：本研究の目的は，これまで進めてきた文脈自由文法（CFG）および確定節文法（DCG）に対する漸次学習方式の研究を基礎として，さらに広範囲の言語に対する効率の高い文法推論を確立し，機械学習へ応用することである．３年間の研究によって，規則集合探索方式の解析と改良，1 次元セルオートマトン(CA) の学習，ブール式充足可能性判定(SAT) ソルバによる文法推論，線形インデクス文法（LIG）の学習などについての新しい成果が得られた．

研究成果の概要（英文）：This research project aims to establish efficient incremental learning of a wider class of formal grammars and its applications to machine learning, based on the previous works on learning context-free grammars (CFGs) and definite clause grammars (DCGs). The results includes analysis and improvement of search strategy for rule sets, learning one-way cellular automata (OCAs) recognizing formal languages, learning grammars by using Boolean satisfiability problem (SAT) solver and learning linear-indexed grammars (LIGs).

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	500,000	150,000	650,000
2010 年度	500,000	150,000	650,000
2011 年度	500,000	150,000	650,000
総計	1,500,000	450,000	1,950,000

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：文法推論，機械学習，確定節文法，漸次学習，セルオートマトン，弱文脈自由文法，線形インデックス文法

１．研究開始当初の背景

形式文法を記号列の例から合成する文法推論は，機械学習の基礎としてまた認知科学における幼児の言語獲得のモデルとして国内外において盛んに研究されている．しかし，この分野の研究の多くは，制限の強い言語について計算量の観点から理論的な限界を明らかにすることが目的であり，実際に意味のある文法を学習する方式およびその応用についての研究は現在もあまりみられない．

この研究計画の基礎は，1999 年以来，続けてきた文脈自由文法(CFG: context-free

grammar）の漸次学習方式の研究である．この方式は Synapse システムに実装されて改良が続けられてきた．このシステムは，文脈自由文法の最小または最小に近い規則集合を合成できることを特長としており，研究開始当初には文脈自由言語を超えた言語を扱うことができる確定節文法（DCG: definite clause grammar）の学習が可能であるように拡張されていた．さらに，DCG によって構文的翻訳図式（syntax-directed translation scheme）の学習が実現され，その応用としてプログラム言語(拡張された算術式)の文法とコンパイラをプログラムの例と対応する中間言語コードの例から学習する試みに成功していた．

２．研究の目的

本研究の当初の目的は，これまで進めてきた CFG の文法推論方式を基礎として DCG に対する効率の高い帰納推論の方法を確立し，機械学習のいくつかの課題への応用を進めることであった．DCG は論理プログラミングで用いられる文法の表現形式であり，非終端記号を表す項への引数の追加によって文脈自由を超える文法も扱うことができる．

実際の 3 年間の研究では，DCG の学習方式とその応用に加えて，以下のような課題について研究が進められた．

(1) 規則集合探索方式の解析と改良

文法推論に膨大な計算量を必要とするという基本的問題を解決する方法の一つは，類似の言語の文法または言語の部分集合からの段階的漸次学習である．もうひとつの解決法は，最小ではなくなるべく小さな（準最小の）規則集合を求める探索方式の採用である．規則集合の探索方式の改良と実験による検討・解析は本研究全体の基礎となる課題である．

(2) 1 次元 1 方向性セルオートマトン（OCA）の学習

CA（cellular automata）は同一の順序機械を規則的に配置・接続してつくられるシステムであり，生物の自己複製や並列計算などのモデルとして用いられている．1 次元セルオートマトンの実時間言語認識能力については長い間研究されてきたが，多くの未解決の問題が残されている．CA は基本的に形式文法とは異なるモデルであるが，形式言語を認識する 1 次元 1 方向性 CA（OCA: one-way CA）の推移規則は Chomsky 標準形の CFG の規則と類似しており，形式言語を受理するための CA を CFG の合成と同様に扱うことができる．

(3) SAT ソルバによる文法推論

ブール式の充足可能性判定問題（SAT: Boolean satisfiability problem）とは与えられたブール式に対してこれを真にする命題変数への代入を求める問題である．各種の探索問題を制約充足問題としてブール式の形式で記述し，SAT ソルバを用いてこれらの制約を満足する解を求めることができる．最近は非常に高速な SAT ソルバがつくられており，さまざまな問題解決に使われている．この方式を文法推論に応用することによって SAT ソルバの進歩を文法の合成に役立てることができる．

(4) 線形インデクス文法（LIG）の学習

LIG は Aho および Duske らによって提案された拡張プッシュダウン・オートマトンにもとづいた文法であり，文脈自由言語(CFL)を超えた弱文脈依存言語(mildly context-sensitive language）と呼ばれるクラスの言語を表わすことができる．このクラスのために互いに等価ないくつかの文法モデルがあるが，LIG は簡潔であり，CFG に比べて自然言語の文法に適していること，多項式時間で構文解析が可能であることなど好ましい特性をもっている．なお，LIG は拡張 DCG に含まれるため, LIG の学習は DCG の学習とも密接な関係をもっている．

３．研究の方法

本研究は理論的な検討とこれを実証する文法推論システムを作成して文法の学習の実験を行うことによって進められる．実際にはこれまで改良を続けてきた Synapse システムを拡張して，OCA および LIG の規則集合を漸次学習する機能を組み込みことが基本となる．ただし，SAT ソルバによる文法推論はこれとは異なり独自のシステムを開発する必要がある．

Synapse は，正例を上向き構文解析した結果（不完全な導出木）から導出木を完成させるためのブリッジ法と呼ばれる規則合成方式と，与えられた例を満足する最小または準最小の規則集合を探索する機能から構成されている．

(1) ブリッジ法による規則生成

漸次学習システムは，学習する言語の正例のおよび負例の順序集合またはその言語の記号列を識別・生成するプログラム，および初期規則の集合(オプション)を入力して，この言語のための規則集合を探索する．ブリッジ法による規則合成手順は次のような非決定的な手続きによって表わされる．

① 正例の記号列に対して，それまでに得ら

れた規則集合を用いて上向きの構文解析を行う．正例は入力によって与えられるか，または与えられたプログラムによって生成する．構文解析の結果が成功ならばこの記号列に対する規則合成は終了する．

② 構文解析が失敗したとき，構文解析の結果である不完全な導出木に対して，欠けた部分を探索し，導出木を完成するような規則を生成する．

③ 合成された規則集合が言語以外の記号列を導出しないことをテストする．これには，与えられた負例をテストするか，または規則集合から指定された長さまで記号列を順に導出し，その記号列を与えられたプログラムによって調べる．どの負例を導出しないなら終了．このプロセスが失敗したときは前の分岐点にバックトラックする．

(2) 規則集合の探索

漸次学習システムは次のいずれかの探索方式によって，すべての正例を導出し，負例を導出しない規則集合を出力する．

① 全域（最小規則）探索(global search)：与えられたすべての正例を導出し，すべての負例を導出しない最小の規則集合を規則数についての反復深化によって求める．

② 直列探索(serial search)：各正例について順にこの記号列を導出し，すべての負例を導出しない最小の規則を反復深化によって求め，規則集合に漸次追加する．これは探索木の一つの経路についてのみ規則集合を求めている．

③ 最良優先探索(best-first search)： 探索木の各節点を評価し，もっとも最終的な規則集合に近いとみられる節点から探索木を拡張する．この方式は，探索を進める候補の節点をある個数に限定するため，ビーム探索とも呼ばれる．

(3) 漸次学習の実験結果

本研究の文法推論方式および応用はすべて Prolog で実装されている．サンプル数やプロセッサによらない探索の量の指標として，目的の規則集合が得られるまでに生成された全規則数 GR を用いている．

表 1 は漸次学習によって合成された文法の規則数，生成された全規則数（GR），計算時間(秒)を示している．この結果は AMD Athlon(tm) 64 X2 Dual Core 2.21 GHz プロセッサおよび SWI-Prolog を用いて得られた．この表において，記号“ - ”は言語がその文法のクラスでは表せないことを，また，“？”は表せるか否かが不明であり，まだ学習できていないことを意味している．OCA の規則数 $14+8^*$ は 8 個の基本的な初期規則をあらかじめ与えて 14 個の規則が合成されたことを意味している．

学習された各言語は次の通りである．

(a) カッコ言語：同じ数の a と b からなり，どの左部分系列中の a の数も b の数を超えない記号列の集合，

(b) $\{a,\ b\}$ 上の回文の集合．

(c) 同じ個数の a と bからなる記号列の集合．

(d) a の数が b の数の 2 倍であるような記号列の集合．

(e) ww の形をもつ$\{a,\ b\}$ 上の記号列の集合．

(f) 集合 $\{a^n\, b^n\, c^n \mid n \geq 1\}$

(g) 集合 $\{a^n\, b^n\, c^m \mid 1 \leq m \leq n\}$

(h) 集合 $\{a^i\, b^j\, c^i\, d^j \mid i,\ j \geq 1\}$

表1 基本的な言語に対する CFG, OCA, LIG の規則集合の規則数(#R)，全合成規則数(GR)，時間(秒)

言語	CFG			OCA			LIG		
	$\#R$	GR	time	$\#R$	GR	time	$\#R$	GR	time
(a)	4	6	0.06	13	122	0.17	2	6	< 0.01
(b)	10	23	0.19	$14+8^*$	267	4.4	8	81	0.06
(c)	7	12	0.22	$12+8^*$	159	6.4	4	12	0.03
(d)	10	21	0.84	?	?	?	7	3538	2.34
(e)	-	-	-	?	?	?	6	5559	2.78
(f)	-	-	-	11	241	0.50	4	488	0.05
(g)	-	-	-	?	?	?	5	1174	1.30
(h)	-	-	-	?	?	?	6	1855	0.70

表2 言語(a), (d), (f) に対する CFG, OCA, LIG の規則集合

(a)	CFG	$s \to s\,s,\ s \to a\,p,\ p \to s\,b,\ s \to a\,b.$
	OCA	$p \to p\,p,\ a \to a\,p,\ p \to p\,a,$ $b \to p\,b,\ p \to b\,p,\ b \to p\,s,\ a \to s\,p,$ $p \to b\,a,\ b \to s\,b,\ b \to b\,b,$ $a \to a\,s,\ a \to a\,s,\ p \to a\,b.$
	LIG	$s\,[a] \to b\,s,\ s \to a\,s\,[a].$
(d)	CFG	$p \to b\,r,\ r \to s\,a,\ p \to p\,s,$ $s \to s\,s,\ p \to a\,q, q \to s\,b,\ p \to b\,a,$ $s \to p\,a,\ s \to a\,p,\ p \to a\,b.$
	LIG	$s \to a\,p\,[a],\ s \to b\,s[b],\ s\,[a] \to b\,s,$ $s\,[b] \to a\,p,\ p \to b\,p[b],\ p \to a\,s,$ $p\,[a] \to b\,p.$
(f)	OCA	$c \to p\,b,\ b \to p\,p,\ p \to c\,c,$ $b \to b\,p,\ p \to c\,p,\ b \to a\,c,$ $c \to b\,b,\ a \to a\,a,\ s \to a\,p,\ p \to b\,c,$ $a \to a\,b.$
	LIG	$q\,[a] \to b\,q,\ s\,[a] \to b\,q,$ $p \to s\,c,\ s \to a\,p\,[a].$

このうち，言語 (a) - (d) は文脈自由であるが，(e) - (h) は非文脈自由言語である．表 2 は合成された言語 (a)，(d)，(f) のため規則

集合を示している．言語(h) は分子生物学における RNA の 2 次構造にみられる 基本的な対が交差する pseudo-knot と呼ばれる系列に関連している．

LIG の規則は $p\,[\sigma] \to a\,q[\tau]$ および $p\,[\sigma] \to q[\tau]\,a$ の形式をもっている．ただし，a は終端記号，$p,\ q$ は状態(非終端記号)である， σ と τ はスタック記号であり，空の場合がある．

図 1 は合成された OCA によって得られた言語 $\{a^n b^n | n \geq 1\}$ と $\{a^n b^n c^n | n \geq 1\}$ を認識する OCA の状相の推移を表わす空間－時間推移図である．状態 2，3，4 がそれぞれ入力記号 $a,\ b,\ c$ を表わしている．言語 $\{a^n b^n | n \geq 1\}$ のための規則は $2 \to 2\,2,\ 1 \to 2\,3,\ 2 \to 3\,3,\ 4 \to 1\,2,\ 1 \to 2\,4,\ 2 \to 2\,1$ であり，1 が受理状態（開始記号）となっている．

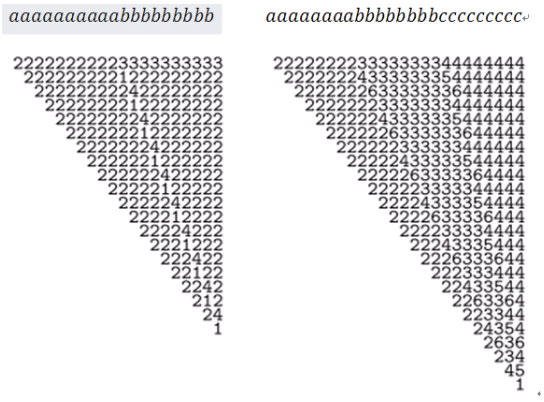

図1 言語 $\{a^n b^n | n \geq 1\}$ および $\{a^n b^n c^n | n \geq 1\}$ を受理する OCA の空間時間推移図

前述の CFG およびランダムに生成した CFG に対して，前述の 3 種類の探索方式を比較した結果，次のような結果が得られた．

① 全域探索は最小に近い規則集合が得られるが，計算時間は長い．

② 直列探索は計算時間は一般に短いが，規則の数は全域探索に比べて最大 2，3 倍となることがある．

③ 最良優先探索は全域探索と直列探索の中間的な結果を示した．

(4) SAT ソルバによる文法推論

この方式では，文法推論の問題を制約充足問題としてブール式の形式で記述し，SAT ソルバを用いてこれらの制約を満足するブール変数への代入を求めることによって，CFG および OCA を合成する．

CFG の学習の場合，与えられた正負の記号列例からブール制約（ブール式の集合）を以下の手順に従って構成する．ただし，CFG は Chomsky 標準形であることが仮定されている．

① 正負の例に出現する任意の部分文字列 $a_1 \cdots a_n$ と非終端記号 p に対して次のようなブール式を制約に加える．

$$T^p_{a_1 \cdots a_n} \leftrightarrow \bigvee_{i=1}^{n-1} \bigvee_{q,r \in N} (R^p_{qr} \wedge T^q_{a_1 \cdots a_i} \wedge T^r_{a_{i+1} \cdots a_n}).$$

ここで，$T^p_{a_1 \cdots a_n}$ は 非終端記号 p が文字列 $a_1 \cdots a_n$ を導出することを表す命題変数である． また，R^p_{qr} は規則 $p \to q\,r$ が規則集合に存在することを表す．

② 規則集合を最小化するために，規則集合の大きさ （真に割り当てられる，変数 R^p_{qr} の数）を表す制約を加える．

③ 任意の正例 w に対して 1 つの変数からなるブール式（単位節）T^s_w，負例 w に対しては $\neg\, T^s_w$ を制約に加える． ここで，s は開始記号を表す．

これらの制約によって，すべての正例を導出され，負例を導出されないことが保証される．規則集合は変数 R^p_{qr} の割り当てによって決定される．これらの制約を満足するような変数への代入（モデル）を SAT ソルバによって求めることによって正負の例に矛盾しない規則集合が得られる．OCA の学習に対しても同様のブール式を求めることができる．

SAT ソルバを用いた文法学習の実験の結果，規則合成に要する計算時間は規則数および非終端記号の数に対して指数的に増大する結果が得られた． さらに，ランダムに生成した CFG を対象に学習実験を行った結果，規則数よりも非終端記号の数のほうがより計算時間との相関が強いことがわかった．規則数は最大 14，非終端記号の数は最大 7 であった．

Synapse を基礎とする漸次学習の結果と比較すると，CFG の学習に要した時間はほぼ同等であった．OCA の学習では括弧言語および，言語 (f) $\{a^n b^n c^n \mid \mathrm{n} \geq 1\}$ の学習に成功してほぼ同様の規則集合が得られた．ブリッジ法にもとづく文法推論方式では漸次学習によって計算量の問題を解決しているが，SAT による文法推論の方式では規則数および非終端記号の数による計算量の限界に現時点では対処できていない．一方，規則集合が最小であることは保障されている．

４．研究成果

(1) CFG および DCG を漸次学習するための Synapse システムについて，学習の高速化と機能の充実化のために規則集合探索方式を改良し，全域探索，直列探索，最良優先探索

3 種類の探索戦略を比較した.
(2) 形式言語を並列認識する 1 次元 1 方向性のセルオートマトン (OCA) の規則集合の漸次学習方式について，理論的な解析をすすめ，非文脈自由言語を含むいくつかの基本的言語を受理する OCA の規則集合を合成することができた.
(3) SAT ソルバによる CFG および OCA の学習システムを作成し，ブリッジ法にもとづく文法学習と比較した.
(4) 基本的な弱文脈依存言語の文法を合成できるシステムを作成した．これまで Synapse で用いられてきた文脈自由文法の規則合成方式（ブリッジ方式）を応用できることが明らかになった.
(5) 本研究課題に関連する研究として非同期並列的に適用される規則集合がその規則集合自身を合成できるような自己複製 (self-replication) のモデルを作成してシミュレーションを行い，新しい人工生命のモデルであることを確認した(学会発表②).
残された研究課題として，次のようなものがあげられる.
(1) LIG を代表とする弱文脈依存文法の効率の高い構文解析法およびこれにもとづく漸次学習の方式を明らかにし，この方式を自然言語文法の学習および LIG にもとづく構文的トランスデューサの学習へ応用すること.
(2) 輪郭図形の構文的パタン学習への応用.
輪郭図形をその構成要素の循環系列で表すことによって，輪郭形状のパタン・クラスを循環言語（ループ形の文字列の集合）として表すことができる．循環言語に対する DCG の学習が可能となれば，形状の分類や検索などのための新しい構文的パタン認識・学習が実現できる.

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　1 件）

① Keita Imada and Katsuhiko Nakamura, Search for Semi-Minimal Rule Sets in Incremental Learning of Context-Free and Definite Clause Grammars, IEICE TRANSACTIONS on Information and Systems, 査読有, E93-D, http://search.ieice.org/bin/summary.php?id=e93-d_5_11972010, pp.1197 -1204.

〔学会発表〕（計　4 件）

① Katsuhiko Nakamura and Keita Imada, Toward Incremental Learning of Mildly Context-Sensitive Grammars, ICMLA 2011, 査読有, Honolulu, Hawaii, USA, Dec. 20, 2011, IEEE, DOI 10.1109/ICMLA.2011.146, pp. 223 -228.
② Katsuhiko Nakamura and Keita Imada, Incremental Learning of Cellular Automata for Parallel Recognition of Formal Languages, DS 2010, 査読有, Canberra, Australia, Oct. 7, 2010, LNAI 6332, pp.117-131.
③ Katsuhiko Nakamura, Asynchronous Parallel Self-Replication Based on Logic Molecular Model, Artificial Life XII Conference, 査読有, Odense, Denmark, Aug. 22, 2010, http://mitpress.mit.edu/books/chapters/0262290758chap74.pdf.
④ Keita Imada and Katsuhiko Nakamura, Learning Context Free Grammars by Using SAT Solvers, ICMLA 2009, 査読有, Florida, USA, Dec. 14, 2009, IEEE, DOI 10.1109/ICMLA.2009.28, pp.267-272.

６．研究組織
(1)研究代表者

中村　克彦（NAKAMURA KATSUHIKO）
東京電機大学・理工学部・教授
研究者番号：90057240

(2)研究分担者　なし

(3)連携研究者　なし