3F3-5

順序グラフパターン言語の多項式時間帰納推論

Ordered Graph Patterns Which Are Polynomial Time Inductively Inferable from Positive Data

日野 隆博 Takahiro Hino　鈴木 祐介 Yusuke Suzuki　内田 智之 Tomoyuki Uchida　宮原 哲浩 Tetsuhiro Miyahara

広島市立大学情報科学研究科
Graduate School of Information Sciences, Hiroshima City University

Ordered graphs, each of whose vertices has a unique order on edges incident to the vertex, can represent graph structured data such as Web pages, TeX sources, CAD, and map data. In order to design computational machine learning for such data, we introduce an ordered graph pattern g with ordered graph structures and structured variables and its ordered graph pattern language $GL(g)$ as the set of all ordered graphs obtained from g by replacing structured variables in g with arbitrary ordered graphs. In this paper, we show that the class $\mathcal{OGPL} = \{GL(g) \mid g$ is an ordered graph pattern$\}$ is polynomial time inductively inferable from positive data.

1. はじめに

近年のコンピュータやネットワーク技術の発展により Web 上には地図データや CAD のような隣接頂点の時計回り順序を持つ平面データや，Web ページや TeX ソース等の子の順序を持つ木構造データが増加している．Jiang と Bunke [Jiang 98] は，このようなデータを表現できる順序グラフを提案した．

順序グラフに共通する構造を表現するため，順序グラフに構造変数を加え，各頂点が辺と構造変数の順序を持つ新しいグラフパターンである順序グラフパターンを提案した [Hino 13]．順序グラフパターンの構造変数には，任意の順序グラフを代入できる．機械学習理論の分野では，帰納推論という学習手法が研究されている．帰納推論とは，与えられたデータからそれらを説明する一般的な規則を導き出す過程である．あるデータが入力されてから仮説を出力するまでの時間が，それまでの入力データサイズの多項式時間で抑えられるならば，多項式時間帰納推論可能であるという．順序グラフパターン言語のクラスが正データから多項式時間帰納推論可能であることを示した [Hino 14] ので本稿で報告する．

2. 順序グラフパターン

Λ と $\mathcal{X}$ は $\Lambda \cap \mathcal{X} = \emptyset$ を満たすアルファベットとする．$\Lambda \cup \mathcal{X}$ 上のグラフパターン g は 3 つ組 (V, E, H) で定義される．ここで V は頂点の集合，E は $V \times \Lambda \times V$ の要素の多重集合で，E の要素を辺とよぶ．H は $V \times \mathcal{X} \times V^+$ の部分集合で，任意の $h \in H$ に対して h 中の全ての頂点は異なるものとする．H の要素を変数とよぶ．頂点 $v \in V$ に対し，$l_g(v)$ を v を含む $E \cup H$ の要素の循環リストとする．$\ell_g(v)$ を v の順序付けリストとよぶ．以下の条件を満たす 4 つ組 $(V, E, H, \mathcal{L})$ を $\Lambda \cup \mathcal{X}$ 上の**順序グラフパターン**という．

(1) (V, E, H) は $\Lambda \cup \mathcal{X}$ 上のグラフパターンであり，グラフ (V, E) は連結である．

(2) 任意の頂点 $v \in V$ は複数の変数の要素とならない．

(3) $\mathcal{L} = \{\ell_g(v) \mid v \in V\}$ は V 中のすべての頂点の順序付けリストの集合である．$\mathcal{L}$ を順序付け集合とよぶ．

連絡先: 鈴木 祐介, 広島市立大学情報科学研究科, 〒 731-3194 広島市安佐南区大塚東 3-4-1, y-suzuki@hiroshima-cu.ac.jp

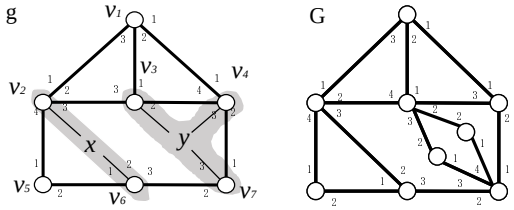

図 1: 順序グラフパターン g と順序グラフ G．g の各頂点 v のまわりの数字は順序付けリスト $\ell_g(v)$ の辺と変数の順序を表す．図の灰色で囲まれた部分は変数を表す．

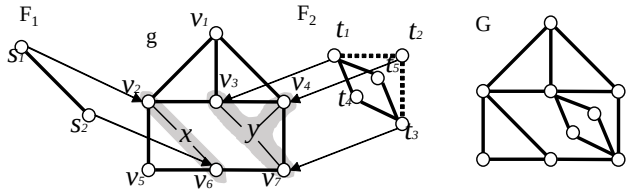

図 2: 順序グラフパターンに対する代入の例．破線で描かれた辺は代入すると消える特別な辺である．

$\Lambda \cup \mathcal{X}$ が文脈より明らかな場合，$\Lambda \cup \mathcal{X}$ を省略する．順序グラフとは変数をもたない順序グラフパターンをいう．スタート辺とよばれる特別な辺が指定されている順序グラフパターンを**スタート辺付き**順序グラフパターンとよぶ．全てのスタート辺付き順序グラフパターンの集合と全てのスタート辺付き順序グラフの集合をそれぞれ $\mathcal{OGP}$ と $\mathcal{OG}$ とする．本論文ではスタート辺付き順序グラフパターンとスタート辺付き順序グラフのみを取り扱うため，以降はスタート辺付きという表記を省略する．順序グラフと順序グラフパターンの例を図 1 に示す．

順序グラフパターン g と順序グラフ G に対し，g の変数を適切な順序グラフで置き換え，順序付けリストを更新することで G が得られるなら，g と G は**マッチ**するという．例えば図 2 では，順序グラフパターン g の変数 x を順序グラフ F_1 で，変数 y を順序グラフ F_2 でそれぞれ置き換えることで順序グラフ G が得られるので g と G はマッチする．順序グラフパターン $g \in \mathcal{OGP}$ に対し，g の**順序グラフパターン言語**を $GL(g)$ で表し，集合 $\{G \in \mathcal{OG} \mid g$ と G がマッチする $\}$ と定義する．順

MINL-$\mathcal{OGPL}$ (S)
入力: 順序グラフの空でない有限集合 $S \subseteq \mathcal{OG}$
出力: S に対して極小な言語を持つ順序グラフパターン g
1: スタート辺とただ１つ変数からなる
順序グラフパターン g を作成する;
2: 幅優先探索を用い，S 中の順序グラフに共通する構造を
見つけ g に追加する;
3: g の各頂点の次数を調べ，変数の要素であるか決定する;
4: g の各変数の連結性を調べ，変数を可能な限り分割する;
5: g を出力する

図 3: MINL-$\mathcal{OGPL}$.

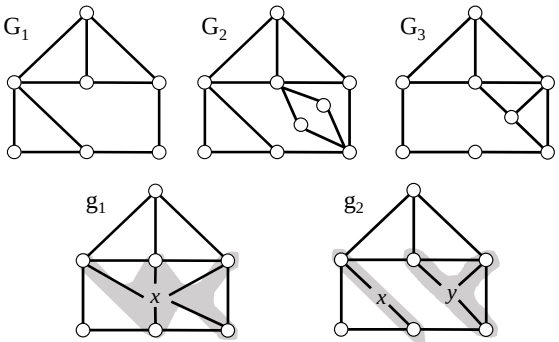

図 4: MINL-$\mathcal{OGPL}$ の出力例．順序グラフの集合 $\{G_1, G_2, G_3\}$ に対して，MINL-$\mathcal{OGPL}$ は g_2 を出力する．

序グラフの空でない有限集合 S に対し，$S \subseteq GL(g)$ であり，$S \subseteq GL(g') \subsetneq GL(g)$ となるような $g' \in \mathcal{OGP}$ が存在しないとき，$g \in \mathcal{OGP}$ は S に対して極小な言語を持つ順序グラフパターンであるという．本論文で学習対象とする順序グラフパターン言語のクラスを $\mathcal{OGPL} = \{GL(g) | g \in \mathcal{OGP}\}$ と定義する．

3. 順序グラフパターン言語の正データからの多項式時間帰納推論可能性

言語クラス $\mathcal{C}$ に対し，$\mathcal{C}$ の所属性問題と極小言語問題が多項式時間で解け，$\mathcal{C}$ が有限の厚みを持つならば，$\mathcal{C}$ は正データから多項式時間帰納推論可能である [Angluin 80, Shinohara 82]．この枠組みに基づいて $\mathcal{OGPL}$ が正データから多項式時間帰納推論可能であることを示す．任意の順序グラフの空でない有限集合 $S \subseteq \mathcal{OG}$ に対し，$|\{L \in \mathcal{OGPL} \mid S \subseteq L\}|$ が有限であるならば，$\mathcal{OGPL}$ は有限の厚みを持つという．

定理 1 クラス $\mathcal{OGPL}$ は有限の厚みを持つ．

クラス $\mathcal{OGPL}$ の所属性問題
入力: 順序グラフ $G \in \mathcal{OG}$ と順序グラフパターン $g \in \mathcal{OGP}$
出力: $G \in GL(g)$ であるか否か

補題 1 [Hino 13] クラス $\mathcal{OGPL}$ の所属性問題は多項式時間で解ける．

クラス $\mathcal{OGPL}$ の極小言語問題 (MINL)
入力: 順序グラフの空でない有限集合 $S \subseteq \mathcal{OG}$
出力: S に対して極小な言語を持つ順序グラフパターン g

クラス $\mathcal{OGPL}$ の極小言語問題を解くアルゴリズム MINL-$\mathcal{OGPL}$ の概略を図 3 に示す．MINL-$\mathcal{OGPL}$ は，まず初めにスタート辺とただ１つの変数からなる順序グラフパターンを作

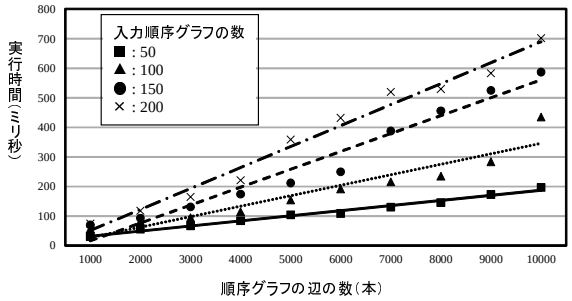

図 5: MINL-$\mathcal{OGPL}$ の実行時間．

る．次に各順序グラフのスタート辺から順序付けリストを元に幅優先探索を行い，共通する辺を確定させていくことで，変数に相当する部分を狭めていく．共通する辺を確定する作業が終了すると，変数との接続の有無や，変数の分割を行うことで極小な言語を持つ順序グラフパターンを生成する．アルゴリズム MINL-$\mathcal{OGPL}$ の出力例を図 4 に示す．順序グラフの集合 $\{G_1, G_2, G_3\}$ に対して，極小な言語を持つ順序グラフパターンは g_2 であり，g_1 は極小な言語を持つ順序グラフパターンではない．

定理 2 クラス $\mathcal{OGPL}$ の極小言語問題は多項式時間で解ける．

S 中に含まれる辺の本数が最大の順序グラフの辺の本数を E_{max} とする．MINL-$\mathcal{OGPL}$ の計算量は $O(|S| \times E_{max})$ である．

MINL-$\mathcal{OGPL}$ の効率性を評価するため，MINL-$\mathcal{OGPL}$ を，主記憶メモリが 12.0GB，OS が Microsoft Windows7 64bit SP1, 3.47/3.47GHz の Xeon プロセッサを持つ計算機上に言語 Java で実装し，評価実験を行った．結果を図 5 に示す．実験結果より，入力順序グラフの集合の要素数と順序グラフの辺の数に比例して実行時間が増加していることが確認できる．

以上の定理 1,2，補題 1 より，以下の定理が示せる．

定理 3 スタート辺付き順序グラフパターン言語のクラス $\mathcal{OGPL}$ は正データから多項式時間帰納推論可能である．

4. おわりに

本研究では，順序グラフパターン言語のクラス $\mathcal{OGPL}$ が正データから多項式時間帰納推論可能であることを示した．また提案したアルゴリズムを計算機上に実装し評価実験を行った．今後の課題として，スタート辺のない順序グラフパターンへの応用や，地図データ等の実データを用いた実験が考えられる．

参考文献

[Angluin 80] D. Angluin : Inductive Inference of Formal Languages from Positive Data. Information and Control, 45(2): pp.117-135, 1980.
[Jiang 98] X. Jiang and H. Bunke : On the Coding of Ordered Graphs. *Computing*, 61(1):23–38, 1998.
[Hino 13] T. Hino et al. : Polynomial Time Pattern Matching Algorithm for Ordered Graph Patterns, Proc. of ILP 2012, LNCS 7842,pp.86-101,2013.
[Hino 14] T. Hino et al. : Ordered Graph Patterns Which Are Polynomial Time Inductively Inferable from Positive Data, 7th IADIS International Conference on Information Systems 2014,pp.263-270,2014.
[Shinohara 82] T. Shinohara. : Polynomial Time Inference of Extended Regular Pattern Languages. RIMS Symposium on Software Science and Engineering, pp.115-127, 1982.