数理解析研究所講究録 1205 巻 2001 年 206-211

消去可能および消去不能変数を含む正則パターンの効率的な帰納学習

植村　仁 (Jin UEMURA)*　　佐藤　優子 (Masako SATO)**
大阪府立大学大学院理学系研究科
(Graduate School of Science, Osaka Prefecture University)*
大阪府立大学総合科学部
(College of Integrated Arts and Sciences, Osaka Prefecture University)**

序

パターンとは，定数及び変数からなる有限文字列であり，パターン言語は各変数へ定数列を代入して得られる文字列からなる．Sato et al.[9] は，変数への消去代入を許さない場合に正則パターン言語の和の族が包含に関する Compactness を有する条件を求めた．すなわち，高々 k 個の正則パターン言語の和の族が Compactness を有する必要十分条件は $\sharp\Sigma \geq 2k-1$ である．Arimura et al.[3] は，一般的な枠組みの下で Compactness を有する言語の族を正例から効率的に学習するアルゴリズムを与えた．本稿では，変数が消去可能な場合，すなわち，空列代入を許す場合の Compactness が成り立つ条件を求め，正則パターン言語の和の族が多項式の更新時間で学習可能であることを示す．また，本稿では，消去可能及び消去不能の 2 種類の変数を含むパターンを導入し，その一般化された正則パターン言語を正例から効率的に学習するアルゴリズムを与える．本稿での意味で一般化された正則パターン言語の族は，通常の正則パターン言語族と拡張正則パターン言語族 ([10]) の和を真に包含する．このような正則パターン言語の和の族は，正例から推論可能となるが，効率的な学習アルゴリズムは得られていない．しかし，一般化された正則パターン言語は，誤りを含む事例からの推論の枠組みとして提案されている「言語の近傍推論」([8]) を通常の正則パターン言語に適用しその効率的な学習アルゴリズムを構築する際，重要と考えられる．

1　消去可能及び消去不能変数を含むパターンとその言語

本稿で扱う**パターン**とは，$\Sigma \cup Y \cup Z$ 上の有限文字列である．ただし，Σ は定数記号からなるアルファベットで，Y, Z は，それぞれ，**消去可能** (erasing) 及び**消去不能** (nonerasing) と呼ばれる変数記号の加算集合で，これらは互いに素な集合とする．$X = Y \cup Z$ とし，X の元を変数とよぶ．変数を $x, x_1, x_2, \cdots$ で表し，Y, Z の変数をそれぞれ $y, y_1, y_2, \cdots$ 及び $z, z_1, z_2, \cdots$ で表す．

パターン p に対して，p の長さを $|p|$ で表し，全てのパターンからなる集合を $\mathcal{P}$ で表す．

代入とは $\mathcal{P}$ から $\mathcal{P}$ への準同型写像であり，(i) 定数をそれ自身に，(ii) 消去不能変数を定数又は消去不能変数を少なくとも 1 文字含むパターンに写すものとする．$x_1, \cdots, x_n$ 以外の変数を含まないパターン p への代入 θ を集合 $\{x_1 := p_1, \cdots, x_n := p_n\}$ で表し，$p\theta$ とかく．パターン p が パターン q の**例化** (または q が p の**汎化**) とは，$p = q\theta$ となる代入 θ が存在することであり，$p \preceq q$ と表す．パターン p, q が **等価**とは，$p \preceq q$ かつ $q \preceq p$ であることであり，$p \equiv q$ と表す．ただし $p \equiv q$ は必ずしも $p = q$ を意味しない．例えば，$y_1 \preceq y_1y_2$ かつ $y_1y_2 \preceq y_1$ ならば $y_1 \equiv y_1y_2$ であるが，$y_1 \neq y_1y_2$ である．しかし，本稿

では，消去可能(不能)変数間の名前の付け替えで等しくなるパターンは同一視する．パターン p が**標準形**であるとは，任意のパターン q に対して $p \equiv q$ ならば $|p| \leq |q|$ となることをいう．

パターン p に対して，$L(p) = \{w \in \Sigma^* \mid w \preceq p\}$ を p によって生成される言語という．明らかに，$p \preceq q$ ならば $L(p) \subseteq L(q)$ であり，また，$p \equiv q$ ならば $L(p) = L(q)$ である．Σ 上の言語 L が パターン言語であるとは，$L(p) = L$ となるパターン p が存在することである．

1.1 正則パターン言語

パターン p が**正則**とは，どの変数も p に高々 1 回しか出現しないことをいう．正則パターンの族及びその言語族をそれぞれ $\mathcal{RP}, \mathcal{RPL}$ で表す．また，消去可能変数(消去不能変数)だけを含む正則パターンの族及び言語族をそれぞれ，$\mathcal{RP}_e(\mathcal{RP}_{ne})$ 及び $\mathcal{RPL}_e(\mathcal{RPL}_{ne})$ と表す．正則パターン yaz が生成する言語 $L(yaz)$ は，$\mathcal{RP}_e \cup \mathcal{RP}_{ne}$ に含まれるパターンでは生成されないので，$\mathcal{RPL}_e \cup \mathcal{RPL}_{ne} \subset \mathcal{RPL}$ である．

パターン p を $p = w_0\alpha_1 w_1 \alpha_2 \cdots w_{n-1}\alpha_n w_n$ とおく．ただし，$w_0, w_n \in \Sigma^*, w_i \in \Sigma^+, \alpha_i \in X^+$ $(i = 1, \cdots, n-1)$．p が正則パターンでかつ標準形ならば，任意の i に対して $\alpha_i \in Y \cup Z^+$ が成り立つ．従って，正則パターン p, q が標準形でかつ同値ならば，変数名の付け替えを除いて，$p = q$ である．このことは標準形である正則パターンの族は，関係 $\preceq$ に関して半順序集合となることを意味する．故に，任意の正則パターン p に対して，$L(q) = L(p)$ となる正則パターンの標準形 q は一意に定まる．

$\mathcal{L}$ を言語族とし，$L \in \mathcal{L}$ とする．空でない有限集合 $S \subseteq \Sigma^*$ が $\mathcal{L}$ における L の**特徴集合**であるとは，任意の $L' \in \mathcal{L}$ に対して，$S \subseteq L'$ ならば $L \subseteq L'$ となることをいう．

正則パターン p に対して，$S_1(p) = L(p) \cap \Sigma^{|p|}$，すなわち，$p$ の各変数に定数を代入して得られる文字列の集合とする．$S(p) = S_1(p) \cup S_1(c(p))$ とおく．ただし，$c(p)$ は p に含まれるすべての消去変数に空列を，消去不能変数にはそれ自身を代入して得られる正則パターンとする．

次の補題は，Sato et al. [9] による消去不能な正則パターンの場合と同様にして，証明される．

補題 1.1. $\sharp\Sigma \geq 3$, $p_1 z p_2, q$ を正則パターン，$a_1, a_2, a_3 \in \Sigma$ を異なる定数とする．このとき，$p_1 a_i p_2 \preceq q$ $(i = 1, 2, 3)$ ならば，$p_1 z p_2 \preceq q$ である．

補題 1.2. $\sharp\Sigma \geq 3$, $p_1 y p_2, q$ を正則パターンとする．$p_1 z p_2 \preceq q$ かつ $p_1 p_2 \preceq q$ ならば，$p_1 y p_2 \preceq q$ である．

補題 1.1 ，補題 1.2 により，正則パターン言語の正例からの帰納推論で重要な役割を果たす次の等価性定理が得られる．

定理 1.3. $\sharp\Sigma \geq 3$, p, q を正則パターンとする．以下の命題は等価である．

$$\text{(i) } S(p) \subseteq L(q), \qquad \text{(ii) } L(p) \subseteq L(q), \qquad \text{(iii) } p \preceq q$$

上記の定理より，次の結果が得られる．

系 1.4. $\sharp\Sigma \geq 3$, p を正則パターンとする．このとき，集合 $S(p)$ は $\mathcal{RPL}$ における言語 $L(p)$ の特徴集合である．

1.2 正則パターンの効率的な学習

p を正則パターンとする．p が標準形であれば，明らかに $|p| \geq 2|c(p)|+1$ となる．また，文字列 $w \in \Sigma^*$ に対して，$w \in L(p)$ ならば，$|c(p)| \geq |w|$ である．従って，文字列 w を含む正則パターン言語の数は高々有限個である．これは，族 $\mathcal{RP}$ が正例から推論可能であるための十分条件である「有限の厚さ」をもつことを意味する ([1])．以下，正例から効率的に正則パターンを学習するアルゴリズムを与える．

本稿で扱う正則パターン言語の所属問題は，$\mathcal{RP}_e$ や $\mathcal{RP}_{ne}$ の場合と同様に多項式時間計算可能である．

補題 1.5. 定数列 $w \in \Sigma^*$ と正則パターン $p \in \mathcal{RP}$ に対して，$w \in L(p)$ か否かは $O(|w|+|p|)$ で計算可能である．

$v, w \in \Sigma^+$ とし，$w = a_1 \cdots a_n$ とおく．ただし，a_i はそれぞれ定数であるとする．定数列 v が w の非連続部分列であるとは，ある $i_1, \cdots, i_k$ (ただし $1 \leq i_1 < \cdots < i_k \leq n$) に対して，$v = a_{i_1} \cdots a_{i_k}$ となることである．これを $v \leq w$ で表す．

$S \subseteq \Sigma^*$ とする．S の共通文字列 CS と極大共通文字列 MCS を次のように定義する．

$$\mathrm{CS}(S) = \{v \in \Sigma^* \mid \forall w \in S, v \leq w\}, \quad \mathrm{MCS}(S) = \{v \in \mathrm{CS}(S) \mid \forall w \in \mathrm{CS}(S), \neg(v < w)\}$$

Shinohara[10] は，有限集合 S の極大共通文字列 (の一つ) を $O(\sharp S \times m^3)$ の時間で計算する手続きを与えている．ただし，m は，S の最長定数列の長さとする．

補題 1.6. $S \subseteq \Sigma^*$ を空でない有限集合とし，$s \in \mathrm{MCS}(S)$ とする．次の手続き MINL は，$\mathcal{RPL}$ における S の極小言語を生成する正則パターンを $O(\sharp S \times m^2)$ の時間で計算する．ただし m は S の最長文字列の長さとする．

Procedure MINL
Inputs: 有限集合 $S \subseteq \Sigma^*$; $s \in \mathrm{MCS}(S)$;
Outputs: 正則パーダン
begin
  let s be a maximal common subsequence of S;
  let $s := a_1 a_2 \cdots a_k$; let l be the length of the shortest constant strings in S;
  let $m := l - k$; $q_0 := y_1 a_1 \cdots a_k y_{k+1}$ where $y_i \in Y$;
  for $i = 1$ **to** $k+1$ **do**
    begin if $S \subseteq L(q_{i-1}\{y_i := \varepsilon\})$ **then**
        begin $q_i := q_{i-1}\{y_i := \varepsilon\}$; **goto** E **end**;
      for $j = m$ **downto** 1 **do**
      if $S \subseteq L(q_{i-1}\{y_i := z_{i,1} \cdots z_{i,j}\})$, where $z_{i,j} \in Z$, **then**
        begin $q_i := q_{i-1}\{y_i := z_{i,1} \cdots z_{i,j}\}$; **goto** E **end** ;
    $q_i := q_{i-1}$
E: **end**;
  output q_{k+1}
end

有限な厚さ (更には一般的な有限の弾力性) を持つ族では，所属問題及び MINL 問題が多項式時間で計算可能ならば，正例から多項式時間で学習可能であることが示されている ([3])．従って，補題 1.5 及び補題 1.6 により，次の定理が成立する．

定理 1.7. 正則パターン言語は正例から多項式時間で推論可能である．

2 消去可能変数のみを持つ正則パターン言語の和の学習

本節では，$\mathcal{RPL}_e$ の正則パターンの言語の和を効率的に学習する問題を扱う．$\mathcal{RP}_e$ に含まれる高々 k 個の正則パターン集合の族を $\mathcal{RP}_e^k$ とかく．

2.1 族 $\mathcal{RP}_e^k$ のCompactness

一般の正則パターンに対しては成り立たないが、消去可能変数だけを含むパターンに限定すると，次の結果が成り立つ．

補題 2.1. $\sharp\Sigma \geq 3, p_1yp_2, q \in \mathcal{RP}_e$ とする．異なる定数 $a_1, a_2, a_3 \in \Sigma$ に対して，$p_1a_ip_2 \preceq q$ であるならば，$p \preceq q$ が成立する．

この補題から，次の結果が得られる．

定理 2.2. $p \in \mathcal{RP}_e$ とする．このとき，$S_1(p)$ は $\mathcal{RPL}_e$ における $L(p)$ の特徴集合である．

正則パターンの集合 P に対して，$L(P) = \bigcup_{p\in P} L(p)$ 及び $S(P) = \bigcup_{p\in P} S(p)$ とする．また，$\mathcal{RPL}_e^k = \{L(P) \mid P \in \mathcal{RP}_e^k\}$ とおく．2つの正則パターンの集合 P,Q に対して，$P \sqsubseteq Q$ であるとは，任意の $p \in P$ について，$p \preceq q$ となる $q \in Q$ が存在することである．明らかに，$P \sqsubseteq Q$ ならば，$L(P) \subseteq L(Q)$ である．

定義 2.3. 族 $\mathcal{RP}_e^k$ が**包含に関するCompactness**をもつとは，任意の $P, Q \in \mathcal{RP}_e^k$ に対して，構文的包含 $P \sqsubseteq Q$ と意味的包含 $L(P) \subseteq L(Q)$ が等価であることをいう．

正則パターン p に対して，$S_2(p)$ を p の各変数に長さ1又は2の定数列を代入して得られる Σ の文字列の集合とする．また，$S_2(P) = \bigcup_{p\in P} S_2(p)$ とする．$S_1(p) \subseteq S_2(p)$ となることに注意する．

定理 2.4. $k \geq 1, P, Q \in \sharp\mathcal{RP}_e^k, \sharp\Sigma \geq k+2$ とする。このとき、下の3つは同値である。

(i) $S_2(P) \subseteq L(Q)$ (ii)$P \sqsubseteq Q$ (iii) $L(P) \subseteq L(Q)$

系 2.5. $k \geq 1, \sharp\Sigma \geq k+2$ とする．任意の $P \in \mathcal{RP}_e^k$ に対して，集合 $S_2(P)$ は $\mathcal{RP}_e^k$ における $L(P)$ の特徴集合である．

有村等[4]は、$\sharp\Sigma = k+1$ のとき，$\mathcal{RP}_e^k$ がCompactnessをもたないことを示した．その反例とは，$\Sigma = \{a_0, a_1, \cdots, a_k\}$, $p = a_0a_0y_0a_ka_k$, $q_i = xa_0a_iy_1$ $(1 \leq i \leq k)$ とする．このとき，$L(p) \subseteq \bigcup_{i=1}^k L(q_i)$ であるが，任意の i に対して，$p \not\preceq q_i$ となる．従って，次のことが成立する．

定理 2.6. $k \geq 1$ とする。族 $\mathcal{RP}_e^k$ が包含に関するCompactnessもつための必要十分条件は，$\sharp\Sigma \geq k+2$ である．

注）消去不能変数だけを含むパターン言語の和の族に関しては,「族 $\mathcal{RP}_{ne}^k$ が包含に関するCompactnessを有する必要十分条件は，$\sharp\Sigma \geq 2k-1$ である．ただし，$k \geq 3$ とする」という結果が得られている([9])．

2.2 言語族 $\mathcal{RPL}_e^k$ の効率的な学習

§1.2で議論したように，言語族 $\mathcal{RPL}$ は有限の厚さを有し，従って，有限の弾力性と呼ばれる性質をもつ．有限の弾力性は言語族の和演算に関して閉じているので(Wright[13])，任意の k に対して，言語族 $\mathcal{RPL}^k$ は正例から推論可能である．従って，その部分族である $\mathcal{RPL}_e^k$ も正例から推論可能である．

一方，Arimura et al.[3] は，本稿の(標準)正則パターンの半順序集合 $(\mathcal{RP}_e, \preceq)$ を含むより一般的なシステムを対象に，正の事例から言語の和を効率的に学習するアルゴリズムを提案した．そこで仮定された条件は，ここで扱う正則パターンを用いて記述すると下記の通りである：

1. 任意の $p,q \in \mathcal{RP}_e$ に対して，$p \preceq q$ と $L(p) \subseteq L(q)$ は等価である．
2. $\mathcal{RP}_e^k$ は包含に関する Compactness をもつ．
3. $\mathcal{RP}_e$ は，次に定義する意味で**効率的** である．
   (a) 任意の $p,q \in \mathcal{RP}_e$ に対して，$p \preceq q$ か否かを多項式時間で判定できる．
   (b) $\mathcal{RPL}_e$ の MINL 問題は多項式時間で計算可能である．
   (c) 多項式時間で計算可能な関数 size : $\mathcal{RP}_e \to N$ が存在し，次の条件を満たす：
      i. $p \prec q$ ならば，size$(p) <$ size(q) である．
      ii. 有限個を除く全ての $p \in \mathcal{RP}_e$ に対して，計算可能な関数 h, h' が存在して size$(p) \leq h(|p|)$ かつ $|p| \leq h'($size$(p))$ となる．
      iii. 集合 $\{p \in \mathcal{RP}_e \mid$ size$(p) \leq n\}$ は，有限かつ計算可能である．

定理 1.3，定理 2.4，補題 1.5 及び補題 1.6 より，上記の 1, 2, 3-(a), 3-(b) が満たされる．3-(c) については，任意の (標準) 正則パターン $p \in \mathcal{RP}_e$ に対して，size$(p) = 3|p|_c - |p|_v + 1$ が成立することによる．ただし，$|p|_c, |p|_v$ はそれぞれ p に出現する定数と変数の個数とする．また，$p \prec q$ ならば size$(p) <$size(q) である k とは，容易に示される．$h(x) = 3x+1, h'(x) = x+1$ と定義すると，h, h' は上記の条件をすべて満たすことがわかる．従って，論文 [3] から，次の結果で得られる．

定理 2.7. 言語族 $\mathcal{RPL}_e^k$ は，正例から多項式時間で推論可能である．

3 結び

本稿では消去可能変数及び消去不能変数を含む正則パターンを導入し、このような一般化正則パターンで生成される言語の族が効率的に帰納学習可能であることを証明した。パターン言語の和の効率的な学習アルゴリズムを構築する際、言語の意味的包含問題をパターン集合に関する構文的包含問題に帰着させる、いわゆる「包含に関する Compactness」の成立が重要な役割を演ずる。本稿では、消去可能な変数のみを許す正則パターンに制限し、高々 k 個の正則パターン集合からなる族が「包含に関する Compactness」を有する必要十分条件が得られた。その結果を用いて、消去可能なパターン言語の和を正例から効率的に推論可能であることを示した。

しかし、本稿で導入した一般化正則パターンに関しては、高々 2 個のパターン言語の和の族でさえ、「包含に関する Compactness」は成立しない。例えば、意味的包含関係

$$L(bayb) \subseteq L(z_1az_2b) \cup L(zab)$$

が成り立つが、構文的包含関係 $bayb \preceq z_1az_2b$ 及び $bayb \preceq zab$ は共に成立しない。一般化正則パターン言語の和の学習問題は、今後の課題である。

References

[1] D.Angluin: *Finding patterns common to a set of strings*, Information and Control, vol. **21**, 46-62, (1980).

[2] D. Angluin: *Inductive inference of formal languages form positive data*, Information and Control, vol. **45**, 117-135, (1980).

[3] H. Arimura, T. Shinohara and S. Otsuki: *Finding minimal generalizations for unions of pattern languages and its application to inductive inference from positive data*, Lecture Notes in Computer Science, vol. **775**, 646-660, (1994).

[4] 有村博紀・篠原武:正則パターン言語和の包含に関する強コンパクト性, 京都大学数理解析研究所講究録, **950**, 246-249, (1996).

[5] E. M. Gold: *Language identification in the limit*, Information and Control, vol. **10**, 447-474, (1967).

[6] T. Moriyama and M. Sato: *Properties of language classes with finite elasticity*, IEICE Transactions on Information and Systems, **E78-D**(5), 532-538, (1995).

[7] Y.Mukouchi: *Containment problems for pattern languages*, IEICE Trans. Inf. & Syst., vol. **E75-D**, No. **4**, 420-425, (1992).

[8] Y.Mukouchi and M.Sato: *Language Learning with a Neibor System*, in Proceedings of the 3rd International Conference on Discovery Science,(2000) to appear

[9] M. Sato, Y. Mukouchi and D. Zheng: *Characteristic sets for unions of regular pattern languages and compactness*, Lecture Notes in Artificial Intelligence, **1501**, 220-233, (1998).

[10] T.Shinohara: *Polynomial time inference of extended regular pattern languages*, RIMS Symposia on Software Science and Engineering, Kyoto, 1982, Proceedings, Lecture Notes in Computer Science **147**, 115-127, (1982)

[11] T. Shinohara and H. Arimura: *Inductive inference of unbounded unions of pattern languages from positive data*, Proc. the 7th International Workshop on Algorithmic Learning Theory, Lecture Notes in Artificial Intelligence, **1160**, 256-271(1996).

[12] T. Shinohara and H. Arimura: *Inductive inference of unbounded unions of pattern languages from positive data*, Lecture Notes in Artificial Intelligence, **1160**, 256-271, (1996)

[13] K.Wright: *Identification of unions of languages drawn from positive data*, Proc. the 2nd Annual Workshop on Computational Learning Theory, 328-333, (1989)