数理解析研究所講究録
950巻 1996年 246-249

正則パターン言語和の包含に関する強コンパクト性

有村 博紀　　篠原 武

arim@ai.kyutech.ac.jp　　shino@ai.kyutech.ac.jp

九州工業大学, 情報工学部
〒820 飯塚市大字川津 680-4

1 パターン言語とその和言語

Angluin (1980) は，与えられた文字列の集合に共通なパターンを抽出するという機械学習の問題を考察し，**パターン言語** (pattern languages) を提案した．定数記号の有限集合を $\Sigma = \{a, b, \cdots\}$ とし，変数の可算集合を $X = \{x, y, \ldots\}$ とする．**パターン**とは，定数記号または変数からなる $aaxybbxaa$ のような文字列 $p \in (\Sigma \cup X)^*$ である．変数名をつけかえて同じになるパターンは，同一のものと考える．

本稿では，同じ変数が2回以上出現しないようなパターンである**正則パターン** (regular, or repetition-free patterns) だけをあつかう．正則パターンで空代入を許すもの全体 (erasing regular patterns) のなす族を ERP と書き，空代入を許さないもの全体 (nonerasing regular patterns) のなす族を RP と書く．正則パターンは，文字列照合におけるワイルドカード "*" (Variable Length Don't Care Symbols) を許した問い合わせパターンと同一のものである．以降では，$p, q, r, p_0, p_1, \ldots$ で正則パターンを表わす．

パターン p の意味として，パターン言語 $L(p)$ を，パターン中の変数を定数文字列でおきかえて得られる定数文字列全体のなす集合と定義する．いいかえると，$L(p)$ はパターン p に照合する定数文字列全体の集合である．パターンの集合 $\{p_1, \ldots, p_m\}$ に対しては，その言語をパターン言語の和集合 $L(\{p_1, \ldots, p_m\}) = L(p_1) \cup \cdots \cup L(p_m)$ と定義する．正整数 $k \geq 1$ に対して，ERP^k (RP^k) で空代入を許す（空代入を許さない）正則パターンの高々 k 個からなる集合の族を表す．

正則パターン言語はたいへん単純な正則言語である．例えば，パターン照合機械構成の技法を用いて，与えられた正則パターンを，受理言語が同じでサイズが線形の決定性有限オートマトン (DFA) に線形時間で変換できる (Shinohara 1982)．さらに，任意の正則パターン言語は，Kleene star $*$ を用いずに $w_0 \overline{\epsilon} w_1 \cdots w_{n_i - 1} \overline{\epsilon} w_{n_i}$ の形の拡張正則表現で表せる．

パターン間には，代入操作によって自然な半順序 $\preceq$ が定義できる：

$$p \preceq q \iff \text{ある代入 } \theta \text{ に対して } p = \theta(q).$$

これを**包摂順序** (subsumption) という．パターンの集合 P, Q に対しては，これを

$$P \sqsubseteq Q \iff (\forall p \in P)\,(\exists q \in Q)\; p \preceq q$$

とべき集合順序を用いて拡張することができる．定義より，$P \sqsubseteq Q \Rightarrow L(P) \subseteq L(Q)$ であるが，この逆は一般に成立しない．パターンの構文的順序 $P \sqsubseteq Q$ に関連した計算が容易なのに対して，意味的順序 $L(P) \subseteq L(Q)$ に関連した計算は容易ではない．

2 包含に関する強コンパクト性

機械学習や情報検索における質問最適化では，構文的順序 $P \sqsubseteq Q$ と意味的順序 $L(P) \subseteq L(Q)$ が一致するとさまざまな計算が効率的におこなえるようになり，たいへん都合がいい．そこで，本稿では両者が一致すること，すなわち $P \sqsubseteq Q \iff L(P) \subseteq L(Q)$ と等価な条件であるつぎの性質について考察する．

> **包含に関する強コンパクト性** (Compactness with respect to containment): 正整数を $k \geq 0$ とする．パターンの族 $\mathcal{C}^k$ が包含に関する強コンパクト性を満たすとは，任意の $l \leq k$ と任意のパターン $p, q_1, \ldots, q_l \in \mathcal{C}$ に対して，$L(p) \subseteq L(q_1) \cup \cdots \cup L(q_l)$ であることと，ある $1 \leq i \leq l$ に対して $p \preceq q_i$ であることが同値であることをいう．

利用できる定数記号の数に上限がないとき ($|\Sigma| = \infty$) には，強コンパクト性は自明な性質である．しかし，Σ が有限な場合，一般にはこの性質は成立しないことを，Angluin (1980) が指摘している．彼女は，$\Sigma = \{\mathbf{0}, \mathbf{1}\}$ のとき，$p = \mathbf{00}x\mathbf{11}$ と $q = x\mathbf{01}y$ に対して，$L(p) \subseteq L(q)$ である一方で，$p \not\preceq q$ であることを示した．これは，$|\Sigma| = 2$ のとき，正則パターンの族 $ERP = ERP^1$ と $RP = RP^1$ は，包含に関する強コンパクト性をもたないことを意味する．

一方で，有限だが十分多くの定数記号が利用できるとき，包含に関する強コンパクト性が成立する場合があることもわかってきた．Shinohara (1982) は，$|\Sigma| \geq 3$ のとき，族 ERP の包含に関する強コンパクト性を証明した．同様の手法で，Mukouchi (1992) は，$|\Sigma| \geq 3$ のとき，族 RP の強コンパクト性を示した．

パターン言語の和に関しても，十分多くの定数記号が利用できるときには，包含に関する強コンパクト性が成立する場合があることもわかっている．Wright (1989) は，$|\Sigma| \geq k+1$ のとき，k 個の 1 変数パターン言語の和のなす族 P_1^k の包含に関する強コンパクト性を示した．有村他 (1992) は，$|\Sigma| \geq k+1$ のとき，1 階論理項によって定義される言語の k 個の和がなす族 TP^k の包含に関する強コンパクト性を示した．しかし，正則パターン言語の和の族 ERP^k, RP^k については，変数の出現数を定数 m で限定した場合に，$|\Sigma| \geq 2km+1$ ならば強コンパクト性が成立するという結果しか得ていなかった．このことは，機械学習や質問最適化への応用においては，大きな制限である．

本稿の主結果はつぎの定理である．これは，$2k+1$ 個以上の定数記号が利用できれば，変数の出現数に依存せずに，族 ERP^k と RP^k が強コンパクト性をもつことを示している．

Theorem 1 (RP^k **and** ERP^k **の強コンパクト性**) 正整数を $k \geq 1$ とし，アルファベットを Σ，正則パターンを $p, q_1, \ldots, q_l$ $(l \leq k)$ とする．このとき，$|\Sigma| \geq 2k+1$ ならば次が成立する:

$$L(p) \subseteq L(q_1) \cup \cdots \cup L(q_l) \iff \text{ある } 1 \leq i \leq l \text{ に対して } q_i \preceq p.$$

つぎに，包含に関する強コンパクト性がなりたつために必要な定数記号数の下限について考える．任意の正整数 $k \geq 1$ に対して，アルファベットを $\Sigma = \{\mathbf{0}, \mathbf{1}, \ldots, \mathbf{k}\}$ とし，正則パターンを $p = \mathbf{00}x\mathbf{kk}$, and $q_i = x\mathbf{0i}y$ $(1 \leq i \leq k)$ とする．このとき，$L(p) \subseteq L(q_1) \cup \cdots \cup L(q_k)$ だが，どんな $1 \leq i \leq k$ に対しても $p \not\preceq q_i$ である．空代入を許さない正則パターンに対しても，同一の Σ と，$p, q_1, \ldots, q_k \in RP$ に対して同じ性質が成立する．このことから，次の定理が成立する．

Theorem 2 (強コンパクト性のために必要な $|\Sigma|$ の下限) 任意の正整数を $k \geq 1$ とする．このとき，和言語族 ERP^k と RP^k のどちらも包含に関する強コンパクト性をもたないような，要素数 $k+1$ のアルファベット Σ が存在する．

現在のところ，上限 $2k+1$ と下限 $k+2$ の見積もりにはかなりのへだたりがある．

3 パターン言語和の包含性判定問題への応用

決定性有限オートマトン (DFA) の部分族 $\mathcal{C}$ に対して，言語和の包含性判定問題とは，つぎの問題 $\mathbf{NCP}(\mathcal{C}^k)$ である．DFA の個数制限 k がないときには，$\mathbf{NCP}(\mathcal{C}^*)$ と書くとする．

The Non-Containment Problem for k DFA ($\mathbf{NCP}(\mathcal{C}^k)$)
Instance: DFA $N, M_1, \ldots, M_m \in \mathcal{C}$, where $m \leq k$.
Question: whether $L(N) \subseteq L(M_1) \cup \cdots \cup L(M_m)$?

DFA の個数を限定しない場合に $\mathbf{NCP}(\text{DFA}^*)$ は PSPACE 完全である．一方で，DFA の個数を定数 k に限定した場合，$\mathbf{NCP}(\text{DFA}^k)$ は NLOGSPACE 完全となる $(k \geq 1)$．先に述べたように，正則パターン言語はたいへん単純な正則言語である．まず個数限定がない場合に，正則パターン言語和の包含性判定問題の計算量は，以下のようになることがわかった．

Theorem 3 問題 $\mathbf{NCP}(ERP^*)$ と $\mathbf{NCP}(RP^*)$ は NP 完全である．

NP 困難性を示すのは容易である．証明が難しいのは，NP に入ることを示す部分である．包含が成りたたないことを示すには，集合 $L(N) - L(M_1) \cup \cdots \cup L(M_m)$ に属するある文字列 w が存在することを言えばいい．しかし，一般の DFA に対しては証拠となる文字列 w はオートマトンのサイズの総和の指数長になりうる．正則パターンについては，つぎの補題が成立する．補題の証明には，Aho-Corasick による複数文字列パターン照合の技法を用いる．

Lemma 4 もし $L(p) - (L(q_1) \cup \cdots \cup L(q_m)) \neq \emptyset$ ならば，集合 $w \in L(p) - (L(q_1) \cup \cdots \cup L(q_m))$ に属する文字列で，長さが高々 $O(n^2)$ のものが存在する．ここに，n は $p, q_1, \ldots, q_m$ の長さの総和である．

パターンの個数制限 k がある場合，一般の DFA に関する結果から，$\mathbf{NCP}(ERP^k) \in$ NLOGSPACE が容易に導かれる．前節の結果から，定数記号が十分に多い場合には包含に関する強コンパクト性が成り立つので，包含性判定はパターン照合問題に帰着できる．

Theorem 5 $|\Sigma| \geq 2k+1$ のとき，$\mathbf{NCP}(ERP^k)$ と $\mathbf{NCP}(RP^k)$ は $DLOGSPACE$ に属する．

$|\Sigma| \leq 2k$ の場合，とくに $\Sigma = \{0, 1\}$ の場合の $\mathbf{NCP}(ERP^k)$ と $\mathbf{NCP}(RP^k)$ の計算量は今後の課題である．

4 パターン言語和の正例からの推論への応用

正例からの多項式時間帰納推論 (Angluin 1980, Shinohara 1982) は，与えられた未知概念の正例だけの枚挙から，効率よく仮説を更新しながら，極限において未知概念を同定する問題である．この学習モデルにおいて，和言語族 ERP^k と RP^k が多項式時間推論可能かどうかは未解決の問題であった．

正例からの帰納推論では，例の過剰汎化を避ける必要があることから，与えられた例をすべて覆う極小言語の計算を効率よくおこなう必要がある．しかし，意味的順序である言語の包含関係を直接あつかうことは難しいので，ほとんどの正例からの効率的学習アルゴリズムは構文的な極小仮説を計算する．有村他 (1994) は，族 ERP^k と RP^k の両方に対して，正例をすべて覆い，包摂順序 $\sqsubseteq$ に関して極小であるような仮説 $H \in ERP^k$ (ERP^k) を多項式時間で計算するアルゴリズムを示している．これと，定数記号が十分に多い場合の包含に関する強コンパクト性を合わせると，つぎの定理が導かれる．

Theorem 6 $|\Sigma| \geq 2k+1$ のとき，族 ERP^k と RP^k は，*Angluin (1980)* の意味で正例から多項式時間推論可能である．

参考文献

[1] D. Angluin, Finding patterns common to a set of strings, *J. Comput. System Sci.* 21 (1980) 46–62.

[2] H. Arimura, T. Shinohara, and S. Otsuki. Polynomial time inference of unions of two tree pattern languages. *IEICE Transactions on Information and Systems*, E75-D(7):426–434, 1992.

[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, in: *Proc. 11th Symp. on Theoretical Aspects of Computer Science*, Lecture Notes in Computer Science, Vol. 775 (Springer, Berlin, 1994) 649–660.

[4] Y. Mukouchi. Characterization of pattern languages. *IEICE Transactions on Information and Systems*, E75-D(4):420–425, 1992.

[5] T. Shinohara. Polynomial time inference of extended regular pattern languages. In *RIMS Symposia on Software Science and Engineering*, pp. 115–127. LNCS 147, Springer, 1982.

[6] T. Shinohara. Polynomial time inference of pattern languages and its applications. In *Proceedings of the 7th IBM Symposium on Mathematical Foundations of Computer Science*, pp. 191–209, 1982.

[7] K. Wright. *Inductive Inference of Pattern Languages.* PhD thesis, University of Pittsburgh, 1989.