2009－MPS－73 (20)
2009／3／5

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

若月 光夫 †,‡ 富田 悦次 †,‡

† 電気通信大学 電気通信学部 情報通信工学科 〒182–8585 東京都調布市調布ケ丘 1–5–1
E-mail: ‡{wakatuki,tomita}@ice.uec.ac.jp

あらまし 比較的単純な言語クラス中の言語を準同型写像によって変換して得られるような，幾つかの拡張された言語クラスが正例から極限同定可能であることが報告されている．われわれは，これらの言語クラスの極限同定を統一的に扱える枠組みを提案し，このような言語クラスが正例から極限同定可能であるための十分条件を与え，その同定アルゴリズムを提案した．本稿では，そのような言語クラスのある新たな真部分クラスに対して，上記の手法に基づいた正例からの多項式時間極限同定アルゴリズムを提案する．

An Identification in the Limit from Positive Data for Some Subclasses of Languages Extended by Homomorphisms

Mitsuo WAKATSUKI†,‡ and Etsuji TOMITA†,‡

†Department of Information and Communication Engineering, Faculty of Electro-Communications, The University of Electro-Communications, Chofugaoka 1–5–1, Chofu, Tokyo 182-8585, Japan
E-mail: ‡{wakatuki,tomita}@ice.uec.ac.jp

Abstract. Many language classes can be regarded as language classes extended by applying homomorphisms and it is known that they are identifiable in the limit from positive data. We have been concerned with a unified algorithm for extending classes of languages identifiable in the limit from positive data. We gave a sufficient condition for the language classes to be identifiable in the limit from positive data and presented a unified identification algorithm for them. In this note, we show that some new proper subclasses of the language classes are polynomial time identifiable in the limit from positive data in the sense of Yokomori.

1 まえがき

形式言語の帰納推論の研究において，Gold[4] は極限同定の概念を定義し，正則言語のクラスでさえ正例から極限同定できないことを示したが，その後の研究によって，正則言語の多くの真部分クラスが正例から極限同定可能であることが報告されている [1, 3, 5, 6, 8]．

それらの中で，Yokomori[8] は strictly deterministic automaton (SDA) と呼ばれる有限オートマトンによって受理される言語 (strictly regular 言語 (SRL)) のクラスが，正例から多項式時間極限同定可能であることを証明した．この SRL は，線形文法の Szilard 言語 [6] を，後述する strict prefix コードで規定した準同型写像によって変換された言語とみなせる．また，Kobayashi and Yokomori[5] は，k-reversible 言語 ($k \geq 0$)[1] をコード [2] で規定した準同型写像によって変換された言語クラスが，正例から極限同定可能であることを証明した．ただし，そこではその言語クラスに対する具体的な同定アルゴリズムは提案されていなかった．

われわれは，このような比較的単純な言語クラス中の言語を準同型写像によって変換して得られるような，拡張された言語クラスに対する正例からの極限同定を統一的に扱える枠組みを提案した [7]．元となる Σ' 上の言語クラスを $\mathcal{L}$ とし，Σ 上の記号列の有限集合のクラスを $\mathcal{X}$ とする．ここで，ある $X \in \mathcal{X}$ に対して準同型写像 $\varphi : \Sigma'^* \to X^*$ が存在するとする．$\mathcal{L}$ の拡張されたクラスは $\mathcal{L}$ と $\mathcal{X}$ によって定義され，$\mathcal{C}(\mathcal{L}, \mathcal{X})$ で表される．文献 [7] では，言語クラス $\mathcal{C}(\mathcal{L}, \mathcal{X})$ が正例から極限同定可能であるための十分条件を示し，その条件を満たす言語クラスに対する統一的な同定アルゴリズムを提案した．

本稿では，文献 [7] の手法に基づいた，$\mathcal{C}(\mathcal{L}, \mathcal{X})$ の新たな真部分クラス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提案する．

2 諸定義および記法

紙数の都合上より，本稿では文献 [7] を基礎として議論を展開するので，本稿で特に明記されていない定義

および記法については，文献 [7] を参照して頂きたい．

アルファベット Σ は記号の有限集合である．空記号列 (長さ 0 の記号列) を ε で表す．記号列 w の長さを $|w|$ で表し，集合 S の要素数を $|S|$ で表す．Σ^* の任意の部分集合 L を Σ 上の言語という．ある記号列 $w \in \Sigma^*$ に対して，$\mathrm{alph}(w)$ は w 中に出現している記号の集合を表す．言語 $L \subseteq \Sigma^*$ に対して，$\mathrm{alph}(L) = \cup_{w \in L}\mathrm{alph}(w)$ と定義する．

続いて，正例からの極限同定の概念について述べる．本稿における多項式時間極限同定の概念は Yokomori[8] の定義に従うものとする．

アルファベット Σ 上の言語クラス $\mathcal{L}$ の要素を表現するクラスを $\mathcal{R}$ で表す．表現の実例としてはオートマトンや文法等がある．$r \in \mathcal{R}$ によって表現される言語を $L(r)$ で表す．記号列 $w \in L$ を L の正例といい，$\{w_i \mid i \geq 1\} = L$ を満たす L の正例の無限系列 $w_1, w_2, \ldots$ を L の正提示という．

言語 $L \in \mathcal{L}$ を表現のクラス $\mathcal{R}$ を用いて推論を行うアルゴリズムを $\mathcal{A}$ とする．$\mathcal{A}$ は，入力として L の任意の正提示が与えられると，$\mathcal{R}$ の要素の無限系列 $r_1, r_2, \ldots$ を出力する．このとき十分大きいある整数 $m\,(\geq 1)$ に対して $L(r_m) = L$ で，かつ $r_m = r_{m+1} = \cdots$ であるならば，$\mathcal{A}$ は L を正例から極限において同定するという．任意の正提示によって任意の言語 $L \in \mathcal{L}$ を正例から極限において同定するアルゴリズム $\mathcal{A}$ が存在するとき，かつそのときに限り，$\mathcal{L}$ は正例から極限同定可能であるといい，このときの $\mathcal{A}$ を $\mathcal{L}$ に対する正例からの極限同定アルゴリズムと呼ぶ．

アルゴリズム $\mathcal{A}$ がある時点 $i\,(\geq 1)$ において推測 r_i を出力したとする．このとき，$i+1$ 番目の正例 w_{i+1} に対して $w_{i+1} \notin L(r_i)$ ならば，$\mathcal{A}$ は時点 i において間接的な誤予測をするという．

定義 1([8]) 以下の性質を満たす，$\mathcal{L}$ に対する正例からの極限同定アルゴリズム $\mathcal{A}$ が存在するとき，かつそのときに限り，$\mathcal{L}$ が正例から多項式時間極限同定可能であるという：表現のサイズが n である任意の $L \in \mathcal{L}$ と，L の任意の正提示に対して，$i\,(\geq 1)$ 番目の正例 w_i を受け取ってから i 番目の推測 r_i を出力するまでに要する時間が高々$p(n, \sum_{j=1}^{i}|w_j|)$ で，かつ $\mathcal{A}$ が間接的な誤予測をする回数が高々$q(n, l)$(ここで，$l = \max\{|w_j| \mid 1 \leq j \leq i\}$) であるような多項式 p, q が存在する． □

3 言語クラスのコードによる拡張

X, Y を Σ 上の記号列の有限集合 (すなわち，$X, Y \subseteq \Sigma^*$) とし，$w \in \Sigma^*$ とする．$w = v_1 v_2 \cdots v_n$ なる任意の系列 $(v_1, v_2, \ldots, v_n) \in (\Sigma^*)^n\ (n \geq 0)$ を w の因子分解という．X が $w \in X^*$ を満たすとき，X は w を因子分解できるという．また，X が $Y \subseteq X^*$ を満たすとき，X は Y を因子分解できるという．

Σ 上の記号列の有限集合 X について，任意の $w \in X^*$ に対する因子分解が一意に定まるとき，X はコード (code) であるという [2]．

$X \subseteq \Sigma^*$ 中のどの要素についても他の X 中の要素の接頭辞でないならば，X は prefix 集合であるという．また，X 中のどの要素についても他の X 中の要素の接尾辞でないならば，X は suffix 集合であるという．更に，X が prefix 集合かつ suffix 集合であるとき，X は biprefix 集合であるという．$\{\varepsilon\}$ を除く，任意の prefix(suffix, biprefix) 集合はコードである [2]．

prefix 集合 X において，X 中の全ての要素に対して先頭記号が互いに異なるとき，X は strict prefix 集合であるという．同様に，suffix 集合 X 中の全ての要素に対して最後尾の記号が互いに異なるとき，X は strict suffix 集合であるという [7]．

$\mathcal{X}$ を Σ 上のコードのクラス，S を空でない Σ^* の部分集合，$X \in \mathcal{X}$ を S を因子分解できるコードとする．S を因子分解できる任意の $X' \in \mathcal{X}$ が X を因子分解できるとき，かつそのときに限り，X は S を因子分解できる最も粗いコードであるという．S を因子分解できる $\mathcal{X}$ 中の最も粗いコードは高々1 つしか存在しない [7]．

Σ' 上の言語クラスを $\mathcal{L}$，Σ 上のコードのクラスを $\mathcal{X}$ とする．コードを用いることによって拡張された Σ 上の言語クラス $\mathcal{C}(\mathcal{L}, \mathcal{X})$ を，次のように定義する．

定義 2 $\mathcal{C}(\mathcal{L}, \mathcal{X})$ は，以下の手続きに従って得られる全ての言語からなるクラスである．(1) 言語 $L \in \mathcal{L}$ に対して $\Sigma'_L = \mathrm{alph}(L) (\subseteq \Sigma')$ とおく．(2) $|X| = |\Sigma'_L|$ なるコード $X \in \mathcal{X}$ に対して，Σ'_L から X への全単射 φ を定める．(3) 各 $L \in \mathcal{L}$，および $X = \varphi(\Sigma'_L)$ なる各 $\varphi : \Sigma'^* \to \Sigma^*$ に対して，Σ 上の言語 $\varphi(L) \in \mathcal{C}(\mathcal{L}, \mathcal{X})$ を $\varphi(L) = \{\varphi(w) \in X^* \mid w \in L\}$ で定義する．このときの φ を coding morphism と呼ぶ． □

定義 3 $\mathcal{L}$ を Σ 上の言語クラスとし，$\Sigma' \subseteq \Sigma$ とする．任意の $L \in \mathcal{L}$ と，L を因子分解できる任意のコード $X \subseteq \Sigma^*$ に対して，$\varphi^{-1}(L) \in \mathcal{L}$ が成立するとき，かつそのときに限り，$\mathcal{L}$ は inverse coding morphism に関して閉じているという．ここで，$\varphi : \Sigma'^* \to \Sigma^*$ は $X = \varphi(\Sigma')$ なる coding morphism である． □

以降では，次の条件 1 および 2 を満たす言語クラス $\mathcal{C}(\mathcal{L}, \mathcal{X})$ を対象とする．

条件 1 言語クラス $\mathcal{L}$ は以下の条件を満たす：(1) $\mathcal{L}$ が inverse coding morphism に関して閉じている．(2) $\mathcal{L}$ は正例から極限同定可能である． □

条件 2 Σ 上のコードのクラス $\mathcal{X}$ は以下の条件を満たす：(1) S をある $X \in \mathcal{X}$ に対して $S \subseteq X^+$ を満た

す，空でない任意の記号列集合とするとき，S を因子分解できる最も粗いコード $X_S \in \mathcal{X}$ が存在する．(2) S の任意の正提示から X_S を極限同定するアルゴリズムが存在する． □

4 同定アルゴリズム

条件1を満たす $\mathcal{L}$ に対する正例からの極限同定アルゴリズム $\mathcal{A}_{\mathcal{L}}$ が存在し，なおかつ条件2の $\mathcal{X}$ 中の最も粗いコードを極限同定するアルゴリズム $\mathcal{A}_{\mathcal{X}}$ が存在するとする．$\mathcal{A}_{\mathcal{L}}$，$\mathcal{A}_{\mathcal{X}}$ の具体的な流れは，文献 [7] 参照．

$\mathcal{A}_{\mathcal{L}}$ 中の関数として CONSTRUCT(r', w') が定義される．これは，入力として目標とする言語 L' の正例 w' および $\mathcal{L}$ 中のある言語の表現 r' を受け取り，新たに得た w' によって更新された $\mathcal{L}$ 中のある言語の表現を出力する．また，$\mathcal{A}_{\mathcal{X}}$ 中の関数として UPDATE(X', w') が定義される．これは，入力として目標とするコード $X \in \mathcal{X}$ に対する正例 $w' \in X^+$ およびコード $X' \in \mathcal{X}$ を受け取り，$X' \cup \{w'\}$ を因子分解できる $\mathcal{X}$ 中の最も粗いコードを出力する．

条件1および2を満たす言語クラス $\mathcal{C}(\mathcal{L}, \mathcal{X})$ に対する極限同定アルゴリズム $\mathcal{A}_{\mathcal{C}(\mathcal{L},\mathcal{X})}$ を以下に示す．

Identification Algorithm $\mathcal{A}_{\mathcal{C}(\mathcal{L},\mathcal{X})}$
Input: a positive presentation $w_1, w_2, \ldots$ of a target language in $\mathcal{C}(\mathcal{L}, \mathcal{X})$
Output: a sequence of pairs $(r_1, X_1), (r_2, X_2), \ldots$ such that $\varphi_i(L(r_i)) \in \mathcal{C}(\mathcal{L}, \mathcal{X}) (i \geq 1)$, where r_i is a representation for a language $L(r_i)$ and X_i is a code such that $X_i = \varphi_i(\Sigma'_i)$ for a coding morphism $\varphi_i : \Sigma'^*_i \to \Sigma^*$
Procedure

```
begin
  S_0 := ∅;   Σ_0 := ∅;   S'_0 := ∅;   Σ'_0 := ∅;
  X_0 := ∅;
  initialize φ_0 so that φ_0(Σ'_0) = X_0;
  initialize r_0 so that L(r_0) = ∅;
  i := 1;   read the next positive example w_i;
  while  w_i = ε do
    S_i := {ε};   Σ_i := ∅;   S'_i := {ε};   Σ'_i := ∅;
    X_i := ∅;   φ_i := φ_{i-1};
    if  L(r_{i-1}) = {ε} then  r_i := r_{i-1}
    else  r_i := CONSTRUCT(r_{i-1}, ε)   fi
          /* Call the function in A_L. */
    output  (r_i, X_i) as a conjecture for a
    language {ε};
    i := i + 1;
    read the next positive example w_i   od
  repeat  (forever)
    S_i := S_{i-1} ∪ {w_i};   Σ_i := Σ_{i-1} ∪ alph(w_i);
    if  w_i ≠ ε then  X_i := UPDATE(X_{i-1}, w_i)
              /* Call the function in A_X. */
    else  X_i := X_{i-1}   fi
    if  X_i ≠ X_{i-1} then
      let a set Σ'_i be given as |Σ'_i| = |X_i|,
      where X_i ⊂ Σ*_i;
      let φ_i be a bijection of Σ'_i onto X_i;
      W := φ_i^{-1}(S_i);
      S'_0 := ∅;   reset r'_0 so that L(r'_0) = ∅;
      j := 1;
      repeat
        w'_j := φ_i^{-1}(w_j);   W := W − {w'_j};
        S'_j := S'_{j-1} ∪ {w'_j};
        if  w'_j ∈ L(r'_{j-1}) then  r'_j := r'_{j-1}
        else  r'_j := CONSTRUCT(r'_{j-1}, w'_j)   fi
              /* Call the function in A_L. */
        j := j + 1
      until  W = ∅;   /* When i = j − 1,
                        it holds that S'_i = φ_i^{-1}(S_i). */
      r_i := r'_i
    else
      Σ'_i := Σ'_{i-1};   φ_i := φ_{i-1};   w'_i := φ_i^{-1}(w_i);
      S'_i := S'_{i-1} ∪ {w'_i};
      /* It holds that S'_i = φ_i^{-1}(S_i). */
      if  w'_i ∈ L(r_{i-1}) then  r_i := r_{i-1}
      else  r_i := CONSTRUCT(r_{i-1}, w'_i)   fi   fi
            /* Call the function in A_L. */
    output  (r_i, X_i) as a conjecture for a
    language φ_i(L(r_i));
    i := i + 1;   read the next positive example w_i
end
```

アルゴリズム $\mathcal{A}_{\mathcal{C}(\mathcal{L},\mathcal{X})}$ の正当性は文献 [7] で証明されており，次の定理を得る．

定理 1([7]) 言語クラス $\mathcal{C}(\mathcal{L}, \mathcal{X})$ が条件1および2を満たすとき，正例から極限同定可能である． □

5 適用例

広義の決定性有限オートマトン (DFA) を $M = (Q, \Sigma, \delta, q_0, F)$ で表し，慣例に従って状態推移関数を $\delta : Q \times \Sigma^* \to Q$ と拡張して扱うこととする．非負の整数 k と DFA の状態 $p, q \in Q$ に対して，$\delta(p, u) = q$ を満たす記号列 $u \in \Sigma^k$ を，q の k-leader という．

任意の $a \in \Sigma$ に対して，$\delta(p, a) = q$ なる状態対 $(p, q) \in Q \times Q$ が高々1つしか存在しないとき，かつそのときに限り，M を restricted strict DFA (RSDA) と呼び，RSDA が受理する言語を restricted strict 正則言

語 (RSRL) と呼ぶ [7]．RSRL の言語クラスは線形文法の Szilard 言語 [6] のクラスと一致する．

k を正の整数とする．任意の相異なる $q_1, q_2 \in Q$ に対して q_1 と q_2 に共通の k-leader が存在しないとき，かつそのときに限り，M を k-definite DFA(k-DDFA) と呼び，k-DDFA が受理する言語を k-definite 正則言語 (k-DRL) と呼ぶ [7]．

$\mathcal{RSR}$，$\mathcal{DR}_k$ $(k \geq 1)$，$\mathcal{R}ev_k$ $(k \geq 0)$ をそれぞれ，RSRL，k-DRL，k-reversible 言語 [1] のクラスとする．これらの言語クラス間には，$\mathcal{RSR} \subset \mathcal{DR}_k \subset \mathcal{R}ev_k (k \geq 1)$ なる関係があるため，これらの言語クラスは正例から極限同定可能である．また，これらの言語クラスは inverse coding morphism に関して閉じており [7]，条件 1 を満たしている．

strict prefix コード，strict suffix コードのクラスをそれぞれ，$\mathcal{SP}$，$\mathcal{SS}$ で表す．言語クラス $\mathcal{C}(\mathcal{RSR}, \mathcal{SP})$ は SRL[8] の言語クラスと一致する．文献 [7] では，$\mathcal{RSR}$，$\mathcal{DR}_k$，$\mathcal{R}ev_k$ のいずれかと，$\mathcal{SP}$，$\mathcal{SS}$ のいずれかとの対による各言語クラスの極限同定可能性について議論した．本稿では，$\mathcal{RSR}$，$\mathcal{DR}_k$，$\mathcal{R}ev_k$ のいずれかと biprefix コードのクラス $\mathcal{BP}$ との対による各言語クラスを対象として考える．

最も粗い biprefix コードを極限同定するアルゴリズムを $\mathcal{A}_{\mathcal{BP}}$ とする．$\mathcal{A}_{\mathcal{BP}}$ 中の関数 UPDATE は次のように書けるので，$\mathcal{BP}$ は条件 2 を満たす．

Function UPDATE(X', w')
begin
  $X := X'$; $Q := \{w'\}$;
  $/ * Q$ is a queue of strings to be parsed $*/$
  let w be the element of Q removed from the top;
  $/ * w$ holds the current string being parsed $*/$
  flag := false;
  repeat
    if $w = \varepsilon$ **then**
      if $Q = \emptyset$ **then** flag := true
      else let w be the element of Q removed
        from the top **fi**
    else
      case w **of**
      (1) $w = yz$ or $w = zy$ $(z \in \Sigma^*)$ for some
      $y \in X$: $w := z$;
      (2) $y = wz$ or $y = zw$ $(z \in \Sigma^+)$ for some
      $y \in X$: $X := (X - \{y\}) \cup \{z\}$;
        put z at the end of Q;
      (3) **else** $X := X \cup \{w\}$; $w := \varepsilon$
      end /* of case */ **fi**
  until flag = true;
  output X
end

文献 [7] と同様の解析を行うことによって，$\mathcal{A}_{\mathcal{BP}}$ の 1 回当りの更新時間および間接的誤予測の回数が，正例の長さに関する多項式オーダで与えられることを示せる．よって，次の定理を得る．

定理 2 言語クラス $\mathcal{C}(\mathcal{RSR}, \mathcal{BP})$ は正例から多項式時間極限同定可能である．また，$k \geq 1$ を定数とみなせば，各言語クラス $\mathcal{C}(\mathcal{DR}_k, \mathcal{BP})$ は正例から多項式時間極限同定可能である． □

上記の $\mathcal{C}(\mathcal{RSR}, \mathcal{BP})$ は，文献 [3] で学習対象としている言語クラスを真に含む．更に，上述の関数 UPDATE を用いたアルゴリズム $\mathcal{A}_{\mathcal{C}(\mathcal{R}ev_k, \mathcal{BP})}$ は，言語クラス $\mathcal{C}(\mathcal{R}ev_k, \mathcal{BP})$ $(k \geq 0)$ に対する正例からの極限同定アルゴリズムとなっている．

6 むすび

本稿では，準同型写像によって拡張された言語クラスに対する極限同定の統一的な枠組みに基づき，biprefix コードによって拡張された言語クラス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提案した．

謝辞 本研究は科学研究費補助金基盤研究 (C) の支援を受けている．

参考文献

[1] Angluin, D., *Inference of reversible languages*, J. ACM **29** (1982), 741-765.

[2] Berstel, J. and D. Perrin, "Theory of Codes", Academic Press, Inc., 1985.

[3] Emerald, J. D., K. G. Subramanian and D. G. Thomas, *Learning code regular and code linear languages*, ICGI'96, LNAI **1147** (1996), 211-221.

[4] Gold, E. M., *Language identification in the limit*, Inform. and Control **10** (1967), 447-474.

[5] Kobayashi, S. and T. Yokomori, *Identifiability of subspaces and homomorphic images of zero-reversible languages*, ALT'97, LNAI **1316** (1997), 48-61.

[6] Mäkinen, E., *The grammatical inference problem for the Szilard languages of linear grammars*, Inform. Process. Lett. **36** (1990), 203-206.

[7] Wakatsuki, M., E. Tomita and G. Yamada, *A unified algorithm for extending classes of languages identifiable in the limit from positive data*, ICGI2006, LNAI **4201** (2006), 161-174.

[8] Yokomori, T., *On polynomial-time learnability in the limit of strictly deterministic automata*, Machine Learning **19** (1995), 153-179.