社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－MPS－50　(8)
2004／6／22

線形言語のある部分言語族に対する多項式時間 PAC 学習可能性

但馬 康宏，小谷 善行，寺田 松昭

東京農工大学大学院 共生科学技術研究部 システム情報科学部門

あらまし　本研究において，線形言語のある部分言語族に対する所属性質問を用いた多項式時間 PAC 学習が，いくつかの条件のもとで可能であることを示す．学習対象となる言語族は，正則言語族と even-linear 言語族を真に含む．この学習対象の言語族に対して，以下のような 2 つの設定を考える．第一の設定は，学習対象を表す文法の生成規則の出現率のうち，もっとも小さな値とその文法のサイズが既知である場合である．第二の設定は，学習者は，ある小さな確率で終了しないことが許される場合である．どちらの設定においても，学習対象となる言語族は，所属性質問と代表部分集合から多項式時間厳密学習を行うアルゴリズムを用いて，効率的に PAC 学習可能である．

Polynomial time PAC learnability of a sub-class of linear languages

Yasuhiro TAJIMA, Yoshiyuki KOTANI, Matsuaki TERADA

Department of Computer, Information and Communication Sciences,
Tokyo University of Agriculture and Technology

Abstract　We show polynomial time PAC learnability of a sub-class of linear languages with membership queries in some special settings. Where the new sub-class of linear languages includes the class of regular languages and the class of even linear languages. For this sub-class, we consider two learning settings as follows. The first case is when the learner knows both of the minimum probability of appearing a specific rule in an example and the size of a grammar which generates the target language. The second case is when the learner does not have to terminate with a small probability. In both cases, the sub-class of linear languages is learnable by supervising algorithms of an exact learning algorithm via membership queries and a set of representative samples.

1　Introduction

In this paper, we show that a sub-class of linear languages is polynomial time learnable via membership queries and examples in two special settings. The sub-class of linear languages is newly defined by us such that it includes the class of regular languages and the class of even linear languages which is polynomial time learnable via queries and counterexamples[4].

In both settings, the learnability is shown by a front-end algorithm which supervises an exact learning algorithm with membership queries and a set of representative samples.

2　Preliminaries

A *context-free grammar* (CFG for short) is a 4-tuple $G = (N, \Sigma, P, S)$. Let σ be the word whose length is 0. Assume that all CFGs are σ-free. In this paper, $\gamma A \gamma' \overset{*}{\underset{G}{\Rightarrow}} \gamma\beta\gamma'$ denotes the *derivation* from $\gamma A\gamma'$ to $\gamma\beta\gamma'$ in G. The *language* generated from γ by G is denoted by $L_G(\gamma) = \{w \in \Sigma^* \mid \gamma \overset{*}{\underset{G}{\Rightarrow}} w\}$. The language generated by G is denoted by $L(G) = L_G(S)$. A nonterminal $A \in N$ is said to be *reachable* if $S \overset{*}{\underset{G}{\Rightarrow}} wA\beta$ for some $w \in \Sigma^*$, $\beta \in N^*$, and a nonterminal $D \in N$ is said to be *live* if $L_G(D) \neq \emptyset$. For two CFGs G_1 and G_2, $L(G_1)\Delta L(G_2)$ denotes the set

$\{w \in \Sigma^* \mid w \in (L(G_1) - L(G_2)) \cup (L(G_2) - L(G_1))\}$. A CFG G is a *linear grammar* iff every rule in G is one of the form $A \to aBb$, $A \to aB$, $A \to Bb$ or $A \to a$ for $a, b \in \Sigma$ and $A, B \in N$. Any other definitions about formal language theories are referred to [2].

Assuming a probability distribution D over Σ^* and let $Pr(w)$ be the probability for $w \in \Sigma^*$, a hypothesis L_h is probably approximately correct[6] (PAC for short) iff

$$Pr[P(L_h \Delta L_t) \leq \varepsilon] \geq 1 - \delta$$

holds where $P(L_h \Delta L_t)$ is the probability of difference between L_h and L_t. Any other definitions about PAC learning are referred to [3].

A membership query $MEMBER(w)$ for $w \in \Sigma^*$ on a linear language L_t replies with 1 if $w \in L_t$ or 0 otherwise.

In this paper, we assume that the learner can use membership queries and examples.

3 Mode selective linear languages

We define a new sub-class of linear languages to show the learnability via membership queries and examples.

3.1 Definitions and properties

A linear grammar $G = (N, \Sigma, P, S)$ which satisfies the following is called a *mode selective linear grammar*:

If a rule $A \to aBc$ is in P for $A, B \in N$ and $a, c \in \Sigma$, then none of $A \to aCc$, $A \to aC$ and $A \to Cc$ are in P for any $C \in N$ such that $C \neq B$.

If a rule $A \to aB$ is in P for $A, B \in N$ and $a \in \Sigma$, then

1. neither $A \to aCc$ nor $A \to aC$ is in P for any $c \in \Sigma$ and any $C \in N$ such that $C \neq B$, and
2. there is no rule in P such as $A \to Db$ for any $D \in N$ and any $b \in \Sigma$.

If a rule $A \to Ba$ is in P for $A, B \in N$ and $a \in \Sigma$, then

1. neither $A \to cCa$ nor $A \to Ca$ is in P for any $c \in \Sigma$ and any $C \in N$ such that $C \neq B$, and
2. there is no rule in P such as $A \to bD$ for any $D \in N$ and any $b \in \Sigma$.

In words, when the derivation for $w \in \Sigma^*$ is proceeded to $S \underset{G}{\overset{*}{\Rightarrow}} uAv$ where $u, v, z \in \Sigma^*$, $a, b \in \Sigma$ and $uazbv = w$, we can select a rule for the next derivation in deterministic with a and b.

Throughout this paper, we assume that the target language is a mode selective linear language denoted by L_t and G_t denotes some mode selective linear grammar such that $L(G_t) = L_t$.

Theorem 1 The class of mode selective linear languages is incomparable to the class of simple deterministic languages and contains the class of regular languages and the class of even-linear languages. □

Theorem 2 Let $G = (N, \Sigma, P, S)$ be a mode selective linear grammar and $w \in L(G)$. Consider a derivation such that

$$S \underset{G}{\overset{*}{\Rightarrow}} w_1 A w_2$$

where $w_1 aubw_2 = w$ for $w_1, w_2 \in \Sigma^*$, $a, b \in \Sigma$ and $u \in \Sigma^+$. Then, there is exactly one rule in P whose left-hand side is A and which can be used in a derivation of $A \underset{G}{\overset{*}{\Rightarrow}} aub$. □

From this theorem, $aub \in L(A)$ iff $w_1 aubw_2 \in L_t$. That is, we can observe behavior of a nonterminal by membership queries for L_t.

3.2 Representative samples

We define a set of representative samples of a linear language L for learning algorithms shown in later.

Definition 3 Let $G = (N, \Sigma, P, S)$ be a linear grammar such that every $A \in N$ is reachable and live. Let Q be a finite subset of $L(G)$. Then Q is a set of *representative samples* (RS for short) of G iff the following holds.

- For any $A \to aBc$ in P, there exists a word $w \in Q$ such that $S \underset{G}{\overset{*}{\Rightarrow}} xAy \underset{G}{\Rightarrow} xaBcy \underset{G}{\overset{*}{\Rightarrow}} w$ for some $x, y \in \Sigma^*$. □

From this definition, for any linear grammar $G = (N, \Sigma, P, S)$, there exists a set of RS Q such that $|Q| \leq |P|$.

Definition 4 For a linear language L, a finite set $Q \subseteq L$ is a set of RS iff there exists a linear grammar $G = (N, \Sigma, P, S)$ such that $L(G) = L$ and Q is a set of RS of G. □

We can find an RS of a linear grammar $G = (N, \Sigma, P, S)$ in time of a polynomial of $|N|$, $|\Sigma|$.

4 PAC learnability

We consider that how many examples are needed to construct a set of RS. For every rule $A \to \beta$ where $\beta \in (N_t \cup \Sigma)^+$ in P_t, let

$$Z(A \to \beta) = \{w \in \Sigma^* \mid S_t \underset{G_t}{\overset{*}{\Rightarrow}} \alpha_1 A \alpha_2 \underset{G_t}{\Rightarrow} \alpha_1 \beta \alpha_2 \underset{G_t}{\overset{*}{\Rightarrow}} w\}$$

for α_1, $\alpha_2 \in (N_t \cup \Sigma)^*$. Then, a probability $Pr(A \to \beta)$ is defined as follows;

$$Pr(A \to \beta) = \sum_{u \in Z(A \to \beta)} Pr(u).$$

It is to say that $Pr(A \to \beta)$ is an appearing probability of $A \to \beta$ when a sample word is given. Now, let $d = \min\{Pr(A \to \beta) \mid A \to \beta \ in \ P_t\}$, then the probability that the rule $A \to \beta$ does not appear in derivations of m samples is bounded by $(1-d)^m$. There are $|P_t|$ rules, thus a set of m samples which satisfies

$$|P_t|(1-d)^m < \delta$$

is a set of RS with a probability at least $1-\delta$. Let $m > -\frac{1}{d}\log(\frac{\delta}{|P_t|})$, then

$$|P_t|(1-d)^m \leq |P_t|e^{-dm} < \delta.$$

On the other hand, it has been proved that equivalence checking between a hypothesis and the target language can be replaced by checking polynomial examples on the PAC criterion[1]. Now, let $n_i \geq \frac{1}{\varepsilon}\left(\log(\frac{1}{\delta}) + (\log 2)(i+1)\right)$ for the learner's i-th guess, then the hypothesis which is consistent to n_i examples satisfies the PAC criterion.

Thus, if there exists an exact learning algorithm via membership queries and a set of RS then the following two learning algorithm can be thought. Let A_q be such an exact query learning algorithm.

Algorithm 1
INPUT : δ, d and $|P_t|$;
OUTPUT: a hypothesis G_h;
begin
  take m examples;
  (let M be the set of example words)
  $Q := Q \cup \{w \in M \mid w \in L_t\}$;
  execute A_q with Q as a set of RS;
  (let G_h be the hypothesis)
  output G_h and terminate;
end.

With this algorithm, we can obtain the following theorem.

Theorem 5 The class of mode selective linear languages is PAC learnable with $\varepsilon = 0$ if the learner knows δ, $|P_t|$ and d. Where the time complexity is bounded by a polynomial of δ, $|P_t|$, d and the maximum length of example words. □

Algorithm 2
INPUT : δ;
OUTPUT: a hypothesis G_h;
begin
  $i := 1$, $Q := \emptyset$;
  repeat
    take i examples;
    (let M be the set of example words)
    $Q := Q \cup \{w \in M \mid w \in L_t\}$;
    execute A_q with Q as a set of RS;
    (let G_h be the hypothesis)
    take n_i examples;
    (let K be the set of example words)
    if $(\forall w \in K, w \in L_t \iff w \in L(G_h))$
    then
      output G_h and terminate;
    fi
    $i := i + 1$;
  until (forever)
end.

With this algorithm, we can obtain the following theorem.

Theorem 6 Algorithm 2 terminates with a probability at least $1-\delta$. If it terminates then the hypothesis is PAC and the time complexity is bounded by a polynomial of δ, $|P_t|$, d and the maximum length of examples. □

We note that Algorithm 2 runs forever with a probability δ. Thus, this is not precise PAC learning algorithm.

5 The query learning algorithm

In this section, we describe the exact learning algorithm for mode selective linear languages via membership queries and a set of RS. This algorithm is based on the algorithm for simple deterministic languages[5]. Let Q be the given set of RS. The following R is the set of candidates for nonterminals.

$$R = \{(x, y, z) \mid x, z \in \Sigma^*, y \in \Sigma^+, x \cdot y \cdot z \in Q\} \cup \{(\sigma, w, \sigma) \mid w \in Q\}$$

and we define $T : R \times \Sigma^* \to \{0, 1\}$ as

$$T((u, v, w), x) = MEMBER(u \cdot x \cdot w).$$

Assume that $W \subseteq \Sigma^*$ is a set of words for partitioning R. At the beginning of the learning algorithm, $W = \emptyset$ and it grows up step by step. We define an equivalence relation $\overset{\pi}{=}$ over R such that

$$r \overset{\pi}{=} r' \iff T(r, w) = T(r', w)$$

for any $w \in W$ where $r, r' \in R$. In addition, let $B_\pi(r) = \{r' \in R \mid r' \overset{\pi}{=} r\}$.

Now, a CFG $G_\pi = (R/\pi, \Sigma, P_{\rm all}/\pi, S_\pi)$ is defined as follows;

$$R/\pi = \{B_\pi(r) \mid r \in R\},$$

$$S_\pi = B_\pi((\sigma, w, \sigma)),$$

where $w \in Q$, and

$$\begin{aligned}
P_{\rm all}/\pi = \{&B_\pi((u_1, a, u_3)) \to a \mid a \in \Sigma, \\
&(u_1, a, u_3) \in R\} \\
\cup\{&B_\pi((u_1, au_2, u_3)) \to aB_\pi((u_1a, u_2, u_3)) \\
&\mid (u_1, au_2, u_3), (u_1a, u_2, u_3) \in R, a \in \Sigma\} \\
\cup\{&B_\pi((u_1, u_2a, u_3)) \to B_\pi((u_1, u_2, au_3))a \\
&\mid (u_1, u_2a, u_3), (u_1, u_2, au_3) \in R, a \in \Sigma\} \\
\cup\{&B_\pi((u_1, au_2b, u_3)) \to aB_\pi((u_1a, u_2, bu_3))b \\
&\mid (u_1, au_2b, u_3), (u_1a, u_2, bu_3) \in R, \\
&\quad a, b \in \Sigma\}.
\end{aligned}$$

From this CFG G_π, the learner deletes some rules according to following conditions. Now, let $A, B \in R/\pi$ and $a \in \Sigma$.

Condition 7 Let $u_1, u_2 \in \Sigma^*$, $B_\pi(r_A)$, $B_\pi(r_B) \in R/\pi$ and $B_\pi(r_A) \to u_1B_\pi(r_B)u_2$ be in $P_{\rm all}/\pi$. If there exists $w \in W$ such that $T(r_A, u_1wu_2) = T(r_B, w)$ then delete $B_\pi(r_A) \to u_1B_\pi(r_B)u_2$ from $P_{\rm all}/\pi$. □

Condition 8 Let $u_1, u_2 \in \Sigma^*$, $B_\pi(r_A)$, $B_\pi(r_B) \in R/\pi$ and $B_\pi(r_A) \to u_1B_\pi(r_B)u_2$ be in $P_{\rm all}/\pi$. If there exists $w \in W$ such that $T(r_B, w) = 1$ and $w \notin L_{G_\pi}(B_\pi(r_B))$ then delete $B_\pi(r_A) \to u_1B_\pi(r_B)u_2$ from $P_{\rm all}/\pi$. □

When the above deletions are repeated $|P_{\rm all}/\pi|$ times, there is no rule which satisfies both of the above conditions. Such a set of rules $P_{\rm all}/\pi$ is called *reduced*. The learner selects *base grammars* $\mathbf{G}$ from G_π.

1. Let $P_0 = P_\Sigma$.
2. For every $A \in R/\pi$ and every pair of $a \in \Sigma$ and $b \in \Sigma$ including $a = b$, select a rule which is of the form $A \to aBb$ from $P_{\rm all}/\pi$ arbitrarily, then add them to P_0. Now, $|P_0|$ is at most $|R/\pi||\Sigma|^2$.
3. For every rule in $P_{\rm all}/\pi$ which is of the form $A \to aB$ or $A \to Bc$, the rule is added to P_0 if P_0 still holds a rule set of some mode selective linear grammar. Such addition is independent of the order of rule selection from $P_{\rm all}/\pi$.
4. Let G_0 be a mode selective linear grammar such that $G_0 = (R/\pi, \Sigma, P_0, S_\pi)$.
5. For a rule $A \to u_1Br_2$ in $P_{\rm all}/\pi$, let $P(A \to u_1Bu_2)$ be a set of rules constructed by deleting all inappropriate rules but $A \to u_1Bu_2$ from $P_0 \cup \{A \to u_1Bu_2\}$ to be a rule set of some mode selective linear grammar. We note that any results of these deletions are in the same set, thus $P(A \to u_1Bu_2)$ is a unique set of rules.
6. Let $\mathbf{G} = \{G(A \to u_1Bu_2) \mid G(A \to u_1Bu_2) = (R/\pi, \Sigma, P(A \to u_1Bu_2), S_\pi), (A \to u_1Bu_2) \in P_{\rm all}/\pi\}$.

For base grammars $\mathbf{G}$, the learner check the following equivalence.

- For every $A \in R/\pi$ and every pair of $G_1 \in \mathbf{G}$ and $G_2 \in \mathbf{G}$ such that $G_1 \neq G_2$, check whether $L_{G_1}(A) = L_{G_2}(A)$ or not.

If it holds that all the above grammars are equivalent then the learner outputs any $G \in \mathbf{G}$ and terminates. Otherwise, adding all sub-words of $w \in \Sigma^*$ such that $w \in L_{G_1}(A) \Delta L_{G_2}(A)$ into W for a next hypothesis.

References

[1] D. Angluin. Learning regular languages from queries and counterexamples. Inf. & Comp., 75:87–106, 1987.

[2] J. E. Hopcroft, J. D. Ullman. Introduction to Automata Theory, Languages, and Computation. Addison-Wesley, Reading, MA, 1979.

[3] B. K. Natarajan. Machine Learning : A Theoretical Approach. Morgan, Kaufmann Publishers, San Mateo, CA, 1991.

[4] Y. Takada. A hierarchy of language families learnable by regular language learning. Inf. & Comp., 123:138–145, 1995.

[5] Y. Tajima and E. Tomita. A polynomial time learning algorithm of simple deterministic languages via membership queries and a representative sample. LNAI 1891:284–297, 2000.

[6] L. G. Valiant. A theory of the learnable. Comm. of the ACM 27:1134–1142, 1984.