

## 線形文法の推定

北九州大 商 棚次奎介

任意の線形言語  $L$  に対し、 $L$  に属するかが否かだけが知られている語のある有限系列を入力として  $L = L(G)$  なる線形文法  $G$  を推定するための手続きを示す。この方法は Biermann が提示した正規言語に対する学習方法<sup>(1)</sup> の自然な一般化である。

### §1. 言語 $L$ からの推定

アルファベット  $\Sigma$  上の言語  $L$  に対し、 $u, v \in \Sigma^*$  による  $L$  の導言語を  $\{u; uv \in L\}$  とし、 $uL\bar{v}$  で表わす。

$L$  に対しいつでも構造図  $D(L) = (N, f)$  が一意に定まる<sup>(3)</sup>。こゝに  $N$  は節の集合で  $\{uL\bar{v}; u, v \in \Sigma^*\}$  を表わす。  $f$  は節から節へのラベルつき矢印を与える写像で、 $a \in \Sigma$  に対し

$$f(X, Y) = \begin{cases} a \cdot & (Y = \bar{a}X \text{ のとき}) \\ \cdot a & (Y = X\bar{a} \text{ のとき}) \\ \text{定義されない} & (\text{その他のとき}) \end{cases}$$

と定める。  $L$  が正規であれば  $D(L)$  は  $L$  を受理する有限オートマトンを与えることになる。 そうでないときは  $D(L)$  は無限構造図となるが、この場合でも  $L$  の文法的構造を十分に内包しており、それを

どうとり出すかが、ここでの中心的課題となる。

いま長さ  $k$  以下の  $L$  の語全体を  $(L)_k$  で表わす。  $\Sigma = \{a_1, a_2, \dots, a_n\}$  として  $L$  の深さ  $k$  の構造木  $T(L, k)$  を次のように帰納的に構成する：

ステップ 0: 根節  $(\varepsilon, \varepsilon)$  に  $(L)_k$  をラベルせよ。

ステップ  $n$ : 現時点で得られている木において、まだチェックされていない葉のうち、最も早く生成された節  $(u, v)$  を選べ。

同じラベルをもつ他のいかなる節からもまだ枝が張られていないなら、節  $(u, v)$  の下に左から右へ順に節  $(ua_1, v), (u, a_1v), (ua_2, v), (u, a_2v), \dots, (ua_n, v), (u, a_nv)$  を並べ、節  $(u, v)$  とそれぞれラベル  $a_1, \cdot a_1, a_2, \cdot a_2, \dots, a_n, \cdot a_n$  をもつ枝で結べ。新しい各節  $(u', v')$  には  $(u'Lv')$  をラベルせよ。節  $(u, v)$  と同じラベルをもつ他の節から既に枝が張られているなら節  $(u, v)$  をチェックせよ。

$(\Sigma^*)_k$  が有限なので上の手続きはいつでも有限ステップで終了する。

いま、 $T(L, k)$  における節の全体を  $T^*(L, k)$  とし、 $\{uLv; (u, v) \in T^*(L, k)\}$  を  $N(L, k)$  で表わす。  $k \leq n$  として、 $L \subset \Sigma^*$  の深さ  $(k, n)$  の構造図  $D(L, k, n) = (N, f)$  を次のように定める：

$$N = \{(X)_k; X \in N(L, k)\}$$

$$f(X, Y) = \alpha \Leftrightarrow T(L, k) \text{ において } (X)_k \text{ とラベルされた節から } (Y)_k \text{ とラベルされた節へラベル } \alpha \text{ をもつ枝が張られている。}$$



(4) 文法  $G(L, 1, 2) = (I, \Sigma, P, S)$ :  $I = \{ \{b\}, \{ba\}, \{ab\}, \{\epsilon\}, \{a\} \}$ ,  $\Sigma = \{a, b\}$ ,  $S = \{b\}$ ,  $P = \{ \{b\} \rightarrow a\{ba\} / \{ab\}a / b\{\epsilon\} / \{\epsilon\}b, \{ba\} \rightarrow \{b\}a / b\{a\}, \{\epsilon\} \rightarrow \epsilon, \{a\} \rightarrow a\{\epsilon\} / \{\epsilon\}a \}$ .

$P$  を単純化すると  $P$  と等価な規則  $P' = \{ S \rightarrow aSa / b \}$  を得る.

明らかに  $L(L, 1, 2) = L$  である.

## §2. 手続きの完全性

任意の線形言語  $L$  に対して適当な非負整数の組  $(k, r)$  を選べば  $L(L, k, r) = L$  とできることを示そう.

いま線形文法  $G = (I, \Sigma, P, S)$  に対して,  $I \cup \Sigma = V$  とし

$T_G = \{(u, v); \text{導出 } S \xrightarrow{*} x \text{ には各非端末記号はたかだか一度しか現われない, } x \Rightarrow uyv, x, y \in V^*, u, v \in \Sigma^*\}$

とし, 言語  $L$  の次数を

$$d(L) = \min_{L=L(G)} (\min \{k; T^*(L, k) \supset T_G\})$$

と定める. また

$$d(L, k) = \max_{\substack{L_1, L_2 \in N(L, k) \\ L_1 \neq L_2}} (\min \{i; (L_1)_i \neq (L_2)_i\})$$

とおく.

命題.  $L \subset \Sigma^*$ ,  $X, Y \in N(L, k)$  とするとき,  $G(L, k, k)$  において  $(X)_k \xrightarrow{*} u(Y)_k v$  なる導出が存在するための必要十分条件は  $Y = \bar{u} X \bar{v}$  である.  $i = 1 - k = d(L, k)$  とする.

証明.  $D(L, k, k) = (N, f)$ ,  $G(L, k, k) = (I, \Sigma, P, S)$

とする.  $N(L, k)$  の要素は互いにたかだか  $k$  の長さで異なるので

$X, Y \in N(L, k)$  とし

$$((X)_k \rightarrow a(Y)_k) \in P \Leftrightarrow f((X)_k, (Y)_k) = a \Leftrightarrow Y = \bar{a}X$$

$$((X)_k \rightarrow (Y)_k a) \in P \Leftrightarrow f((X)_k, (Y)_k) = \cdot a \Leftrightarrow Y = X\bar{a}$$

が成り立つ.  $L$  かつ

$$\neg (X)_k = (Y)_k \Leftrightarrow X \neq Y$$

よって  $(X)_k \xrightarrow{G} u(Y)_k v$  なる導出が存在する  $\Leftrightarrow Y = \bar{u}X\bar{v}$  であることが同等になる.  $\square$

補題 1. 任意の  $L \subset \Sigma^*$  に対し,  $L(L, k, d(L, k)) \subset L$ .

証明.  $d(L, k) = k$ ,  $G(L, k, k) = (I, \Sigma, P, S)$  とおく.

$w \in L(L, k, k)$  とすると命題から

$$S = (L)_k \xrightarrow{G} u(\bar{u}L\bar{v})_k v \Rightarrow uv = w$$

なる導出が存在する. 導出の最後には生成規則  $(\bar{u}L\bar{v})_k \rightarrow \varepsilon$  が適用されており,  $\varepsilon \in \bar{u}L\bar{v}$  となる. ゆえに  $w = uv \in L$ .  $\square$

補題 2.  $G$  をある正規形文法とし,  $L(G) = L$  とおく.

$T^*(L, k) \supset T_G$  ならば  $L(L, k, d(L, k)) = L$  である.

証明.  $G = (I, \Sigma, P, S)$  とする.  $(X \rightarrow uXv) \in P$  ( $u, v \in \Sigma^*$ ,  $X \in V^*$ ) に対し, 各非端末記号がたかだか一度しか現われないような適当な導出  $S \xrightarrow{G} u'Xv'$  ( $u', v' \in \Sigma^*$ ) が存在し,  $u'Xv' \xrightarrow{G} u'uXvv'$  となるので  $T_G$  の定義から  $(u', v')$ ,  $(u'u, vv') \in T_G \subset T^*(L, k)$ .  $L$  かつ  $\bar{u}L\bar{v}$ ,  $\overline{u'u}L\overline{vv'}$

$\in N(L, k)$  である。そのとき命題1によって

$$(\bar{u}L\bar{v})_k \xrightarrow{*} u(\bar{u}uL\bar{v}v')_k v$$

なる導出が  $G(L, k, k)$  において存在する。こゝに  $k = d(L, k)$

とする。また、 $w \in \Sigma^*$  で  $(x \rightarrow w) \in P$  に対しても、上の生成規

則で  $x = \varepsilon$ ,  $uv = w$  とした場合に相当し、 $\varepsilon \in (\bar{u}uL\bar{v}v')_k$  とな

るので  $G(L, k, k)$  において

$$(\bar{u}L\bar{v})_k \xrightarrow{*} u(\bar{u}uL\bar{v}v')_k v \xrightarrow{*} uv = w$$

なる導出が存在することになる。

以上から、 $G$  によって生成される語は  $G(L, k, k)$  によっても生成される。

一方、補題1から  $L \supseteq L(L, k, k)$  である。  $\square$

補題2と  $d(L)$  の定義からただちに次の定理を得る。

定理1. 任意の線形言語  $L$  に対して

$$L(L, d(L), d(L, d(L))) = L$$

### §3. サンプルからの推定

$L \subset \Sigma^*$  から  $L = L(G)$  なる文法  $G$  を推定する場合、 $L$  全体を入力とすることは非現実的である。実は  $L$  の適当な有限部分集合によって  $G$  を推定することが可能である。こゝではそのような部分集合の条件とそれからの推定手続きを示す。

いま  $L$  の特徴集合を

$$C(L) = \{ w \in u(\bar{u}L\bar{v})_{d(L, d(L))} v ; (u, v) \in T^*(L, d(L)) \}$$

とする.  $C(L)$  は  $L$  の有限部分集合である.

定理 2.  $L$  を線形言語とする.  $C(L) \subset C \subset L$  のとき, 適当な非負整数  $k, h$  が存在して  $L(C, k, h) = L$  とできる.

証明.  $k = d(L)$ ,  $h = d(L, d(L))$  とする.  $(u, v) \in T^*(L, k)$  に対して  $C(L)$  の定義から  $u(\bar{u}L\bar{v})_k v \subset C(L) \subset C$  を得る. ところで  $(\bar{u}L\bar{v})_k \subset \bar{u}C\bar{v}$  となり,  $(\bar{u}L\bar{v})_k \subset (\bar{u}C\bar{v})_k$  が導かれる. 一方,  $C \subset L$  だから  $(\bar{u}C\bar{v})_k \subset (\bar{u}L\bar{v})_k$ . したがって  $(\bar{u}L\bar{v})_k = (\bar{u}C\bar{v})_k$  を得る.  $k \leq h$  だから  $(\bar{u}L\bar{v})_k = (\bar{u}C\bar{v})_k$  となって  $T(L, k)$  と  $T(C, k)$  は全く一致し,  $D(L, k, h) = D(C, k, h)$ , すなわち

$$L(C, k, h) = L(L, k, h) = L. \quad \square$$

上の定理の応用として線形文法  $G$  を推定するためのアルゴリズムを図に示す. これは言語  $L$  に属するか否かは知られているものとして語の系列  $w_1, w_2, \dots, w_n$  を次々に入力とし,  $L$  の文法を逐次的に推定していく方法である.  $I_+$  は  $L$  に属する語の集合を表わし, 直接推定に用いられる.  $I_-$  は  $L$  に属さない語の集合を表わし推定された文法が妥当かどうかの判断にのみ用いられる.

例 2.  $L = \{a^n b a^n; n \geq 0\}$  とする.  $L$  の特徴文集合は  $C(L) = \{b, aba\}$  である.  $L$  の  $n=7$  の入力として  $a, b, ab, aba$  をこの順に与えると, 各段階で次のような文法を推定する:

$$(1) I_+ = \{b\}, I_- = \{a\}: G = G(I_+, 0, 1), L(G) = \{b\},$$

