数理解析研究所講究録
第 213 巻 1974 年 249-258

有限オートマトンの特徴集合について

北九州大　商　　　棚次　奎介
九大　理　情報研　　有川　節夫

§1. 有限オートマトンとその特徴集合

Biermann [1] は任意の有限サンプル集合をもとに，有限オートマトンの対話型学習アルゴリズムを提供した。ここでは，その学習アルゴリズムの中核をなす有限オートマトンの特徴集合と次数についての2,3の性質を調べ，正則言語族の一つの階層を与える．

$\mathfrak{A}=(S, \delta, s_0, F)$ を Σ 上の有限オートマトンとする。長さ k 以下の語 $x\in\Sigma^*$ が存在して $\delta(s,x)$ または $\delta(s',x)$ のどちらか一方だけが最終状態となるとき，二つの状態 $s, s'\in S$ は <u>k-区別可能</u>であるという。S の任意の二つの状態が k-区別可能となるような最小の k を $\mathfrak{A}$ の<u>次数</u>といい，$d(\mathfrak{A})$ で表わす．$\#S=1$ のときは $d(\mathfrak{A})=0$ と定める。$d(\mathfrak{A})$ は $\mathfrak{A}$ が既約であるとき，そのときにのみ定義される。以後 $\mathfrak{A}$ は最小オートマトンとする.

次に示す方法によって $\mathfrak{A}$ の状態遷移図から対応する木 $T(\mathfrak{A})$ を導き，$T(\mathfrak{A})$ と $d(\mathfrak{A})$ とから $\mathfrak{A}$ の<u>特徴集合</u>とよばれる $L(\mathfrak{A})$ の有限

部分集合 $C(\alpha)$ を決定できる[1]：

(1) $T(\alpha)$ の構成

1° s_0 を木の root node 名とし，木の最上位におく．それから $\#\Sigma$ 本の枝を延ばして各枝を Σ の要素で名づけ，枝 σ につながる node 名を $\delta(s_0, \sigma)$ とする．

2° s をある node 名とする．s がこの node の左方または上位にある node の名前として既にあらわれているなら，いま着目している node からは枝を張らない．そうでないなら，$\#\Sigma$ 本の枝を延ばして各枝を Σ の要素で名づけ，枝 σ につながる node 名を $\delta(s, \sigma)$ とする．

(2) $C(\alpha)$ の決定

$T(\alpha)$ において root node から任意の node へ至るまでの枝名を順に連接することによって得られる語の全体を $T^*(\alpha)$ とするとき，

$$C(\alpha) = \{uv;\ u \in T^*(\alpha),\ |v| \leq d(\alpha),\ \delta(s_0, uv) \in F\}.$$

上記方法の一例を図1に示す．ここに $S = \{1, 2, 3\}$，$s_0 = 1$，$F = \{3\}$ とする．

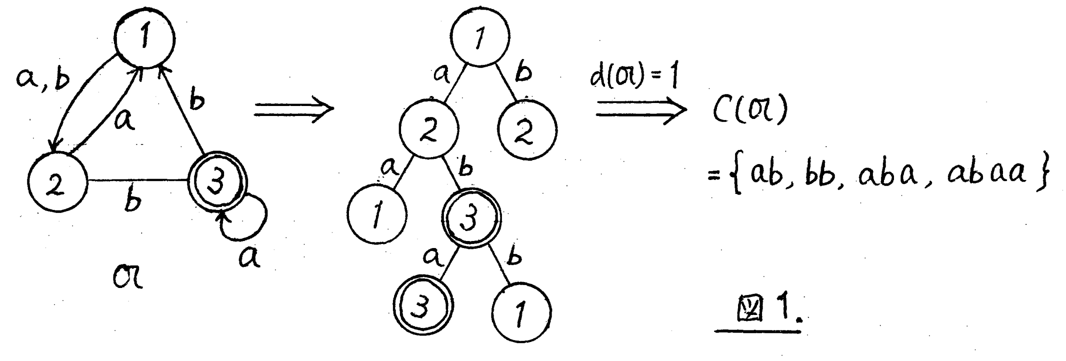


図1.

逆に $C(\mathfrak{a})$ と $d(\mathfrak{a})$ が既知であれば，もとの $\mathfrak{a}$ を再構成できる．これを一般に $C \subseteq \Sigma^*$，整数 $k\ (\geq 0)$ に対する方法として次に与える：

いま，$g(w, C, k) = \{x;\ wx \in C,\ |x| \leq k\}$ とする．そのとき，前記方法 (1) の $1°$ における s_0 を $g(\varepsilon, C, k)$，$\delta(s_0, \sigma)$ を $g(\sigma, C, k)$ で，$2°$ における s を $g(w, C, k)$，$\delta(s, \sigma)$ を $g(w\sigma, C, k)$ でおきかえることによって，まず C と k に対応する木 $T(C, k)$ が得られる．$T(C, k)$ の各 node 名を状態名とすることにより，最小有限オートマトン $\mathfrak{a}(C, k)$ が容易に定まる．すなわち $\mathfrak{a}(C, k) = (S, \delta, s_0, F)$ において

$$S = \{g(x, C, k);\ x \in T(C, k)\}$$

$$s_0 = g(\varepsilon, C, k)$$

$$F = \{g(x, C, k);\ \varepsilon \in g(x, C, k)\} \subseteq S$$

$$\delta:\ \delta(g(x, C, k), a)$$

$$= \begin{cases} g(xa, C, k); & x \in T^*(C, k) - T_*(C, k) \text{ のとき} \\ g(x'a, C, k); & x \in T_*(C, k) \text{ のとき} \end{cases}$$

である．ここに $T_*(C, k)$ は $T(C, k)$ において root node から lief node へ至るまでの枝名を順に連接することによって得られる語の全体とし，$x' \in T^*(C, k) - T_*(C, k)$ かつ $g(x, C, k) = g(x', C, k)$，$a \in \Sigma$ とする．

上記方法は C が有限であるなしにかゝわらず唯一つの $\mathfrak{a}(C, k)$ を決定する．$L(C, k) = L(\mathfrak{a}(C, k))$ とするとき，さらに次の命題が成り

たつ．

命題．　Lが正則言語ならば $L=L(C,k)$ なるような $C\subseteq\Sigma^*$，整数 $k(\geq 0)$ が存在する．

この命題は $\mathfrak{A}$ を $L=L(\mathfrak{A})$ なる最小オートマトンとするとき，$C=C(\mathfrak{A})$，$k=d(\mathfrak{A})$ において成りたつが，後に示すようにその場合だけに限らない．

§2．正則言語族の階層化

$C\subseteq\Sigma^*$ と整数 $k(\geq 0)$ を任意に与えたとき，それらが有限オートマトン決定に及ぼす影響を考察する．

補題1．　$w\in T^*(C,k)$ のとき $g(w,C,k)=g(w,L(C,k),k)$．

（証明）　$x\in g(w,C,k)$ とする．　$wx\in T^*(C,k)$ の場合は $\varepsilon\in g(wx,C,k)$ より $wx\in L(C,k)$．　そこで $wx\notin T^*(C,k)$ とする．　そのとき 次の $(*)$ を満たす $w_1,w_2,\cdots,w_{\ell-1}\in T^*(C,k)-T_*(C,k)$ が存在し，$\varepsilon\in g(w_{\ell-1}x_\ell,C,k)$ となる．ここに，$\ell\geq 1$，$w_{\ell-1}x_\ell\in T^*(C,k)$．

$$(*)\begin{cases} x=x_1x_2\cdots x_\ell \ (x_1,x_2,\cdots,x_\ell\neq\varepsilon) \\ x_1\in T_*(C,k),\ x_2\cdots x_\ell\in g(w_1,C,k)=g(x_1,C,k) \\ w_1x_1\in T_*(C,k),\ x_3\cdots x_\ell\in g(w_2,C,k)=g(w_1x_1,C,k) \\ \vdots \\ w_{\ell-2}x_{\ell-1}\in T_*(C,k),\ x_\ell\in g(w_{\ell-1},C,k)=g(w_{\ell-2}x_{\ell-1},C,k) \end{cases}$$

そのとき $\mathfrak{A}(C,k)$ において

$$\delta(g(\varepsilon, C, k), wx) = \delta(g(wx_1, C, k), x_2\cdots x_\ell)$$
$$= \delta(g(w_1x_2, C, k), x_3\cdots x_\ell) = \cdots = \delta(g(w_{\ell-2}x_{\ell-1}, C, k), x_\ell)$$
$$= g(w_{\ell-1}x_\ell, C, k) \in F.$$

したがって $wx \in L(C,k)$，すなわち $x \in g(w, L(C,k), k)$。

以上から $g(w, C, k) \subseteq g(w, L(C,k), k)$． 逆も同様． ⊠

補題1により $T(C,k)$ と $T(L(C,k),k)$ とは全く同一の木となるのでただちに次の定理を得る．

定理1. $L(C,k) = L(L(C,k), k)$．

補題2. $w, w' \in T^*(C,k)$，$L(C,k) = C$ のとき

$$g(w, C, k+1) = g(w', C, k+1) \Leftrightarrow g(w, C, k) = g(w', C, k).$$

（証明）（$\Rightarrow$） 明らか．

（$\Leftarrow$） $g(w, C, k) = g(w', C, k)$ であって $x \in g(w, C, k+1) \triangle g(w', C, k+1)$ なる x が存在すると仮定する．一般性を失うことなく $wx \in C$，$w'x \notin C$ とできる． $wx \in C = L(C,k)$ より $\mathfrak{A}(C,k)$ において

$$\delta(g(\varepsilon, C, k), w'x) = \delta(g(w', C, k), x)$$
$$= \delta(g(w, C, k), x) = \delta(g(\varepsilon, C, k), wx) \in F.$$

したがって $w'x \in L(C,k) = C$ となり矛盾を生ずる． ⊠

定理2. $L(C,k) = C \Rightarrow L(C,k+1) = C$

（証明） 補題2より $T(C,k+1)$ は $T(C,k)$ と同型となる．

また，任意の $w \in \Sigma^*$ に対して

$$\varepsilon \in g(w, C, k) \Leftrightarrow \varepsilon \in g(w, C, k+1)$$

だから，$\mathfrak{A}(C, k+1)$と$\mathfrak{A}(C, k)$とは同等になる． $\boxtimes$

上の定理は，Cを固定してkを0から1ずつ増やしていくとき，一度$L(C, k)$がCに一致すればそれ以降はつねにCであり続けることを示している．

さて $\mathcal{L}_k = \{ L(C, k) ; C \subseteq \Sigma^* \}$ としよう．

<u>定理 3.</u> $\mathcal{L}_k \subsetneq \mathcal{L}_{k+1}$

（証明） 任意の$L(C, k) \in \mathcal{L}_k$に対して定理1から$L(C, k) = L(L(C, k), k)$，定理2から$L(C, k) = L(L(C, k), k+1)$を得る．すなわち，$L(C, k) = L(C', k+1)$なる$C' \subseteq \Sigma^*$が存在する．ゆえに$\mathcal{L}_k \subseteq \mathcal{L}_{k+1}$．

次に$\mathcal{L}_k \neq \mathcal{L}_{k+1}$を示す．いま$C = \{\sigma^{k+2}\}$ ($\sigma \in \Sigma$)としよう．明らかに$\min\{|x| ; x \in L(C, k+1)\} = k+2$である．$L(C, k+1) = L(C', k)$なる$C' \subseteq \Sigma^*$が存在すると仮定し，$C'$の最小長の語を$w$としよう．$|w| \leq k+1$ならば$w \in L(C', k)$だから$|w| < k+2$となり矛盾．したがって$|w| \geq k+2$．そのとき$g(\varepsilon, C', k) = \phi$，任意の$\sigma \in \Sigma$に対して$g(\sigma, C', k) = \phi$より$L(C', k) = \phi$となる．これまた矛盾を導く． $\boxtimes$

この定理3によって図2のような正則言語族の非負整数kによる無限の階層が与えられる．

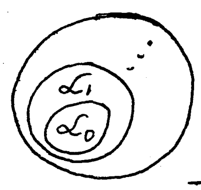


図2.

§3. $\mathcal{L}_k$の性質

定理3によって得られた階層に関連して以下の命題が成りたつ.

<u>命題1.</u> $\forall k (\geq 0), \exists n_k : \# \mathcal{L}_k \leq n_k$

(証明) $\Sigma^{(k)} = \{x ; x \in \Sigma^*, |x| \leq k\}$とするとき, 任意の$w \in T^*(c, k)$に対して$g(w, c, k) \subseteq \Sigma^{(k)}$. すなわち, $\mathfrak{a}(c, k)$の状態の個数はたかだか$2^{\# \Sigma^{(k)}}$である. ☒

<u>例1.</u> $\Sigma = \{a\}$のとき $\mathcal{L}_0 = \{\phi, \varepsilon, a^*, aa^*, (a^2)^*, a(a^2)^*\}$. $\# \Sigma = 2$のとき $\# \mathcal{L}_0 = 26$..

<u>命題2.</u> $L \subseteq \Sigma^{(k)} \Rightarrow L \in \mathcal{L}_k$

(証明) $g(\varepsilon, L, k) = L$ より $L = L(L, k) \in \mathcal{L}_k$. ☒

<u>命題3.</u> (1) $\mathcal{L}_k$は complementation に関して閉じている.

(2) $\mathcal{L}_k$は$\cup$, $\cap$, 連接に関して閉じていない.

(証明) (1) $g(w, \bar{c}, k) = \Sigma^{(k)} - g(w, c, k)$. したがって

$g(w, \bar{c}, k) = g(w', \bar{c}, k) \Leftrightarrow g(w, c, k) = g(w', c, k)$.

すなわち, $T(\bar{c}, k)$と$T(c, k)$とは同型である. また

$$\varepsilon \in g(w, \bar{c}, k) \Leftrightarrow \varepsilon \notin g(w, c, k).$$

ゆえに $\overline{L(c, k)} = L(\bar{c}, k) \in \mathcal{L}_k$.

(2) 演算$\circ$に関して閉じているなら

$$\forall c_1, c_2 \subseteq \Sigma^*, \exists c : \quad L(c_1, k) \circ L(c_2, k) = L(c, k).$$

定理1より $L(L(c_1, k) \circ L(c_2, k), k) = L(c_1, k) \circ L(c_2, k)$となる.

(i) $\cup$ について: $C_1=\{\varepsilon\}$, $C_2=\{a^{k+1}\}$ とすれば $a^{2k+2}\notin L(C_1,k)\cup L(C_2,k)$, $a^{2k+2}\in L(L(C_1,k)\cup L(C_2,k),k)$.

(ii) $\cap$ について: $C_1=\{a^{k+1},a^{k+2}\}$, $C_2=\{\varepsilon,a^{k+2}\}$ とすれば $a^{k+2}\in L(C_1,k)\cap L(C_2,k)$ だが, $L(L(C_1,k)\cap L(C_2,k),k)=\phi$.

(iii) 連接について: $C_1=C_2=\{a^{k+1}\}$ とすれば $a^{2k+2}\in L(C_1,k)\cdot L(C_2,k)$ だが, $L(L(C_1,k)\cdot L(C_2,k),k)=\phi$. ⊠

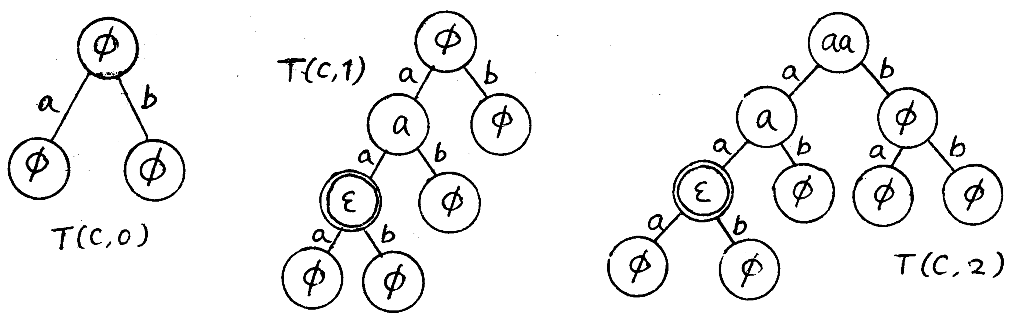
正則言語(有限オートマトン)の学習という立場からすれば, L からのいかなるサンプル集合と k の値の組み合わせによってもとの L を同定できるかという問題がある. この問題に関連して以下の命題が成りたつ.

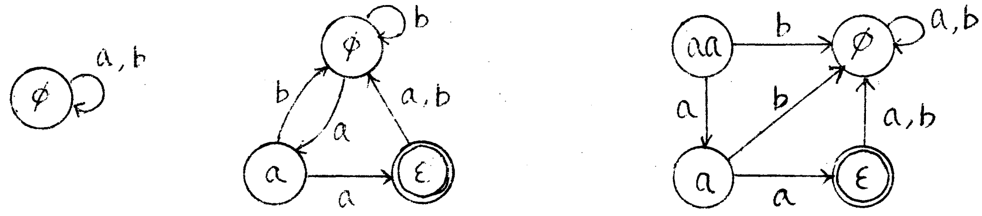
いま, $m(C)=\min\{|x|;\ x\in C\}$, $M(C)=\max\{|x|;\ x\in C\}$, $d(C)=\min\{k;\ C\subseteq L(C,k)\}$, $D(C)=\min\{k;\ C=L(C,k)\}$ としよう.

<u>命題 4.</u> (1) $m(C)-1\leq d(C)\leq M(C)-1$

(2) $D(C)\leq M(C)$

<u>例 2.</u> $C=\{aa\}$ のとき 図3より $L(C,0)\not\supseteq C$, $L(C,1)\supsetneq C$, $L(C,2)=C$. したがって $d(C)=m(C)-1=M(C)-1$, $D(C)=M(C)$.





$\mathfrak{A}(C,0)$ $\mathfrak{A}(C,1)$ $\mathfrak{A}(C,2)$

図3. $C=\{aa\}$ による学習

命題5. $\forall k(\geq 0),\ \exists C\subseteq \Sigma^*:\ d(C)=D(C)=k$

命題6. $\mathfrak{A}$を最小有限オートマトンとし，$L=L(\mathfrak{A})$とする. $C\subseteq L$ のとき，

$$C(\mathfrak{A})\subseteq C \iff \exists k:\ L(C,k)=L$$

(証明)($\Rightarrow$) $L(C(\mathfrak{A}),k)=L$ とする. $C(\mathfrak{A})\subseteq C\subseteq L$ より

$$\forall w\in\Sigma^*:\ g(w,C(\mathfrak{A}),k)\subseteq g(w,C,k)\subseteq g(w,L,k).$$

ところで $w\in T^*(C(\mathfrak{A}),k)$ のとき 補題1 から

$$g(w,C(\mathfrak{A}),k)=g(w,L(C(\mathfrak{A}),k),k)=g(w,L,k).$$

すなわち，$g(w,C(\mathfrak{A}),k)=g(w,C,k)$. したがって $T(C,k)$と$T(C(\mathfrak{A}),k)$とは同一である. ゆえに $L(C,k)=L$.

($\Leftarrow$) $L(C,k)=L$, $L(C(\mathfrak{A}),k)=L$ とする. $\mathfrak{A}(C,k)$と$\mathfrak{A}(C(\mathfrak{A}),k)$とは同等な最小オートマトンとなるから $T(C,k)$と$T(C(\mathfrak{A}),k)$とは同型で最終状態も同じ位置にある. したがって

$$C(\mathfrak{A})=\bigcup_{x\in T^*(C,k)} x\cdot g(x,C,k)\subseteq C.$$ ⊠

〈注〉 $C(\mathfrak{A})$は$L(\mathfrak{A})$を復元できる$L(\mathfrak{A})$の部分集合のう

ちで最小のものといえる。

命題 7. 任意の整数 $k\ (\geq 0)$, $C \subseteq \Sigma^*$ に対して, $L(C,k) = L(C',k)$ なる $C' \subseteq \Sigma^*$ は無数に存在する。

参考文献

[1] Biermann, A. W., An interactive finite-state language learner, 1st USA-JAPAN Comp. Conf. Proc. (1972), 13-20.