

整合関係を考慮した記号列パターン識別オートマトンの 構成および適応的修正

東京工業大学 工学部 榎本 肇
富田悦次

1. ま え が き

系列パターン認識システムに関する研究の一環として、我々はこれまでに、カテゴリ名の付与された有限のサンプル記号列が与えられたときに、これらを正しく識別するようなオートマトンを構成し、さらに必要に応じて、それを教師との質問応答によって、適応的に修正し、ゆくゆくアルゴリズムを考えてきた。本稿においては、簡単のために、識別機構を決定性有限オートマトンに限って、これらの方法をまとめて記した。

2. 諸定義・記法

[1] 決定性有限オートマトン(以下では、単に FSA と記す)を $M = (K, \Sigma, \delta, S_0, F)$ (K : 状態の集合, Σ : 入力記号の集合, $\delta: K \times \Sigma \rightarrow K$ への写像, S_0 : 初期状態, F : 最終状態の集合) を表す。

[2] FSA M において、 $\delta(s_1, a) = s_2 \in \delta$ であるとき、この写像規則による推移を、 $(s_1) \xrightarrow[M]{a} (s_2)$ と表す。さらに、 $i = 1, 2, \dots, m$ に対して $(s_i) \xrightarrow[M]{a_i} (s_{i+1})$ なる推移が可能であるときには、 $(s_1) \xrightarrow[M]{x} (s_{m+1})$ ただし、 $x = a_1 a_2 \dots a_m$ と表す。さもなくば、 $(s_1) \not\xrightarrow[M]{x} (s_{m+1})$ と表す。

入力記号列 u に対し、 M の決定性推移 $(s_0) \xrightarrow[M]{u} (p)$ によって一意的に定まる状態 p を、 $C_M(u)$ で表す。

[3] 入力記号には、適当にアルファベット順序を設ける。入力記号列の間には、短いものから長いものへと順に、また同じ長さのもの同士では、初めて入力記号の異なる位置における入力記号のアルファベット順に従ってその辞書式順序を定める。このような入力記号列集合 π のうちで、最も辞書式順序が先行する記号列を $\text{First}(\pi)$ で表す。

3. 整合関係を考慮した記号列パターン識別 FSA の構成⁽¹⁾

3-1 サンプル記号列集合に対する前提

ある正規集合をなすパターン全体の各カテゴリから、いくつかの有限長、有限個のサンプル記号列が抽出されているとき、それらのサンプルから元のパターン集合全体を推論する方法を考える。ここで、元のパターン集合の記号列は終止記号($\#$)付きのものであり、異ったカテゴリ集合間の重なりはなく、その集合を支配する規則は、最終状態名の区別によって

カテゴリを識別するような FSA $M_D = (K_D, \Sigma_D, \delta_D, S_{0D}, F_D)$ として記述されているとする。

このような推論の第一近似として、先ず少くともこれらのサンプルだけは正しく識別するような FSA M' の構成を計る。ところで、このようなサンプルの間には、その取得過程に応じて何らかの関係が生じていることが普通であり、それに関する予備知識も導入することによって、より正しい推論を達成することができる。ここで本節においては、サンプル記号列集合は、次の条件を満足するようなものとなっていることが予めわかっている場合を扱うことにする。

〔サンプルに対する前提条件〕カテゴリ i に属するサンプル記号列の集合を π_{ai} としたとき、 $uv \in \pi_{ai}$ であるならば、 $\forall$ $u \cdot \text{First}\{v \mid C_{M_D}(u) \xrightarrow{v}_{M_D} (F_i), F_i \in F_D \text{ はカテゴリ } i \text{ 用の最終状態}\} \in \pi_{ai}$ である。すなわち、ある状態からの推移を示す記号列としては、最も辞書式順序の先行するものをさしおいて、それよりも順序が後の記号列だけがサンプル中に採られているというようなことはない。

3-2 部分記号列の整合関係を考慮した FSA 構成.

前項の前提条件を満足するようなサンプル記号列の集合に対して、それらの部分記号列間の整合関係を解析しながら FSA の構成を行う方法を示す。

先ず、サンプルに対する前提条件のいかんにかかわらず、次の関係が成立する。

[命題1] (i) M_D において、 F_i, F_j をそれぞれカテゴリ i, j ($i \neq j$) 用の最終状態としたとき、 $F_i \neq F_j$ 。(ii) サンプル記号列中の2箇所共通な部分記号列 v が存在し、 $(S_h) \xrightarrow{v}_{M_D} (S_k)$ 、 $(S_h) \xrightarrow{v}_{M_D} (S_{k'})$ であるとしたとき、 $S_k \neq S_{k'}$ ならば、 $S_h \neq S_{h'}$ 。

さらに、サンプルに対する前提条件のもとでは、

[命題2] サンプル記号列の prefix u_1, u_2 に対して、
 $\text{First}\{v_1 | u_1 v_1 \in \pi_{a_i}\} \neq \text{First}\{v_2 | u_2 v_2 \in \pi_{a_i}\}$ for some i であるならば、 $CM_D(u_1) \neq CM_D(u_2)$ 。

この2命題中の M_D を M' で置き換えて得られるような関係は、 M_D とできるだけ近いものとして構成しようとする FSA M' においても当然満足されるべき関係であり、FSA 構成におけるこのような条件を、suffix 整合条件 と名付ける。特に、命題2に対する条件は、前提付き suffix 整合条件 と呼ぶ。

FSA の構成の仕方としては、サンプル記号列において、その先端位置よりの辞書式順序が後の方の位置から順に、suffix 整合条件 に反しない限りにおいて逐次状態の統合を計ってゆくという方法をとる。このような統合の可能性は一般に幾通りか考えられるが、ある段階でその可能性を調べるのは各個先の範囲内のものまでに留め、そこで統合が不可能で

あるようなものに対しては新しい状態を割り当てる⁽²⁾。なお、前提付き suffix 整合条件を考慮せずに、 $k=\infty$ あるいは $k=0$ とおいた場合は、文献(3)の最小状態法、および(単純)階層法に相当する。

前述のサンプルに対する前提条件は、いつの場合においても完全に満たされているという保証は勿論ないが、ほぼ満足されているような場合に対しては、この前提におけるのと同様にして FSA の構成を行い、然る後にその前提の妥当性をチェックし、必要な修正を行うようにしてもよい。(次節参照)
また、サンプルの母集団が正規集合でない場合にも、この構成法は、推論の近似的解を与える。

4. FSA の適応的修正

前節で得られるような FSA は、一般にはまだ誤りを含んでいる可能性があるため、本節では、それをさらに適応的に修正して目的のものへと到達させる方法を示す。なお、以下では簡単のために対象とする FSA は非冗長であり、パターンの(受理)カテゴリは単一であるとする。ここで、教師に対してある特定の記号列がそのカテゴリに属するものであるか否かの質問を發したときには、その答は誤りなく与えられるものとする。

4-1 FSA の代表記号列集合^{(4), (5)}

受理または非受理の指示（それぞれ，○印，×印で表す）が付けられた有限のサンプル記号列集合 $\pi_c(M)$ に対して，それらを正しく識別するような FSA は一般にいくつも構成されるが，構成において満足されるべきある基準を設けたとき，そのもとでは特定の FSA M （あるいは，それと同形のもの）が一意的に対応付けられるとき，記号列集合 $\pi_c(M)$ は，その基準のもとでは M の特性を完全に代表して表現しているといえることができ，このような $\pi_c(M)$ を M の代表記号列集合と名付ける。ここで，FSA M' の構成の基準は次のように定める。

〔FSA 構成基準〕 ○ $w \in \pi_c(M)$ ならば， $(s_0) \xrightarrow{w}_M (s_f)$ ， s_f は M の最終状態。× $\bar{w} \in \pi_c(M)$ ならば， $(s_0) \xrightarrow{\bar{w}}_M (s_f)$ であり，かつ，前節の前提付き suffix 整合条件を満足するような，状態数および写像規則数が最少の FSA M' を構成する。（たとえば，前節で $k = \infty$ とした場合の構成法によればよいが，具体的構成法は問わない）

以下に，代表記号列集合の求め方を示す。

〔基本記号列集合 $\pi_F(M)$ 〕 $\pi_F(M) = \{uav \mid (s_0) \xrightarrow{u}_M (p) \xrightarrow{a}_M (q) \xrightarrow{v}_M (s_f), \delta(p, a) = q \in \delta, u = \text{First}\{u' \mid (s_0) \xrightarrow{u'}_M (p)\}, v = \text{First}\{v' \mid (q) \xrightarrow{v'}_M (s_f)\}\}$

〔副次記号列集合 $\pi_S(M)$ 〕 $u_i v', u_j v'' \in \pi_F(M), \text{First}\{v' \mid C_M(u_i) \xrightarrow{v'}_M (s_f)\} = \text{First}\{v'' \mid C_M(u_j) \xrightarrow{v''}_M (s_f)\}$ であり，か

つ、 $C_M(U_i) \neq C_M(U_j)$ であるとある。 $(\exists w_0)[w_0 = \text{First}$
 $\{w \mid C_M(U_i) \xrightarrow[M]{w} (S_f), C_M(U_j) \xrightarrow[M]{w} (S_f)\}]$ であるならば、 $\circ U_i w_0$
 $\in \pi_{sa}(M)$, $\times U_j w_0 \in \pi_{sr}(M)$ 。さらに、 $C_M(U_i) \xrightarrow[M]{w'} (Q) \xrightarrow[M]{w''} (S_f)$,
 $w_0 = w' w''$ なる任意の状態 Q に関して、 $w_0'' = \text{First}\{w'' \mid (Q) \xrightarrow[M]{w''} (S_f)\}$
 としたとき、 $\circ U_i w' w_0'' \in \pi_F(M) \cup \pi_{sa}(M)$ とする。い、
 の立場を逆にした場合も同様。以上の操作を、前記関係にある
 基本記号列中のあらゆる prefix U_i, U_j に対して適用した
 結果得られる記号列集合 $\pi_{sa}(M)$, $\pi_{sr}(M)$ の和集合を副次記
 号列集合 $\pi_s(M)$ とする。」

$\pi_F(M) \cup \pi_s(M) = \pi_c(M)$ とおいたとき、この $\pi_c(M)$ は、前節
 の、サンプル記号列に対する前提条件を満足している。

[定理] $\pi_c(M) = \pi_F(M) \cup \pi_s(M)$ から前記構成基準に従
 って得られる FSA M' は、もとの FSA M と同形である。そ
 れ故、この $\pi_c(M)$ は、前記構成基準のもとで、 M の代表記号
 列集合をなす。

[例題1] 図1の FSA M_D に対して、

$$\pi_F(M_D) = \{ \circ ab\#, \circ aac\#, \\ \circ aabbb\#, \circ aabdb\# \}$$

$$\pi_s(M_D) = \{ \circ aac \cdot acb\#, \times aab\cdot acb\#, \\ \times aabd \cdot acb\# \}$$

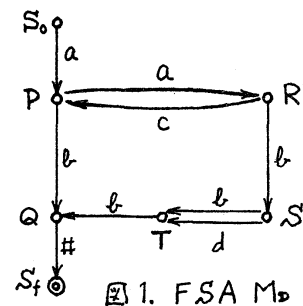


図1. FSA M_D

4-2 代表記号列集合を利用したFSAの適応的修正⁽⁶⁾

最初に与えられたFSAを M_0 としたとき、前項に示したように、代表記号列集合 $\pi_c(M_0)$ はその構成基準のもとにおいて M_0 を完全に表現しており、 M_0 が誤りをもっている場合には、目的とあるFSA M_D に対する代表記号列集合 $\pi_c(M_D)$ とは異なったものとなっている。従って、FSAの適応的修正の過程は、先ず教師への質問によって、 $\pi_c(M_0)$ の記号列に対する受理・非受理の指示を全て正しい応答に訂正したような集合 $\pi_{c_D}(M_0)$ を求め、それでも $\pi_c(M_D) \neq \pi_{c_D}(M_0)$ である場合には、さらに積極的に $\pi_c(M_D) = \pi_F(M_D) \cup \pi_S(M_D)$ の要素として不足している可能性のある記号列を、 $\pi_c(M_D)$ の各要素の定義に応じて生成し、その応答を教師に質問して新しい確認情報を得、 $\pi_c(M_D) - \pi_{c_D}(M_0)$ を逐次的により小さくしてゆく過程として考えることができる。

以下、具体例に沿って、このような考えに従った適応的修正の過程を示す。(一般アルゴリズムは、文献(6)参照)

【例題2】最初に与えられたFSAは、図

2の M_0 、目的とあるFSAは図1の M_D とある。

[0] $\pi_F(M_0) = \{ \circ a a b b \# , \circ a a c b \# , \circ a a b d \# , \circ a a b b b \# \}$

$\pi_S(M_0) = \varnothing$

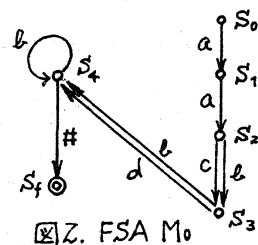


図2. FSA M_0

従って、 $\pi_{C_D}(M_0) = \pi_{C_{Da}}(M_0) \cup \pi_{C_{Dr}}(M_0)$

ただし、 $\pi_{C_{Da}}(M_0) = \{ \circ aacbb\#, \circ aabbbb\# \}$,

$\pi_{C_{Dr}}(M_0) = \{ x aabbb\#, x aabdb\# \}$

$\pi_{C_{Da}}(M_0)$ の記号列だけでは、まだ $(S_3) \xrightarrow{d}_{M_0} (S_4)$ の路が未通過であるので、これを通過するような受理記号列を探すと、

$aabdb\#$, $aabbb\cdot b\#$ の prefix, suffix の連接より成る $aabdb\cdot b\#$ が見出せ、これを基本記号列の補充分 $\pi'_F(M_0)$ とする。

さらに、以後のFSA構成時においては、前提付き suffix 整合条件を常に考慮に入れることにするため、その妥当性チェックを行う。すなわち $\pi_{C_{Da}}(M_0) \cup \pi'_F(M_0)$ 中の記号列で、その suffix をより辞書式順序の先行する suffix で置換したものが新たな受理記号列とならないかを調べると、 $\circ a \cdot ac \cdot b\#$ に対して、 $a \cdot b\#$ が受理であることがわかり、これを $\pi_A(M_0)$ とおく。

以上の $\pi_{C_D}(M_0) \cup \pi'_F(M_0) \cup \pi_A(M_0)$ より前項の構成基準に従って FSA を構成すると、図3の M_1 を得る。

[1] $\pi_{C_D}(M_1) = \{ \circ ab\#, \circ aacbb\#, \circ aabbbb\#, \circ aabdbb\# \}$

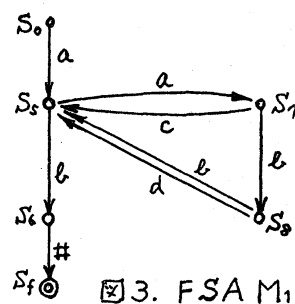


図3. FSA M_1

複数の入力推移のある状態 S_5 において、

$aac\cdot b\#, aabbb\cdot b\#, aabdb\cdot b\# \in \pi_F(M_1)$ の共通 suffix

$b\#$ における推移が同じとなっていることが正しいかどうかのチェックとして、副次記号列の定義に応じて、記号列 $aac\cdot acb\#$,

$aabb \cdot acb \#$, $aabd \cdot acb \#$ を生成して、その正しい応答を質問すると、これらはそれぞれ、受理、非受理、非受理である。従って、 $\pi'_S(M_1) = \{ \circ aacace \#, x a a b b a c b \#, x a a b d a c b \# \}$ を副次記号列集合の補充分とある。ここで、 $\pi_{C_D}(M_1) \cup \pi'_S(M_1)$ は $\pi_C(M_D)$ と一致し、基準に従った構成によって図1の FSA M_D が得られ、修正は完了。

5. あとがき

以上の方法は、ある種の決定性フォッシュダウンオートマトンに対して拡張することも可能である^{(1),(4),(6)}。また、これまでのオートマトンでは、最終的な受理または非受理の応答を考えているだけであるが、さらに推移の途中における出力応答も考えて、それによって自己の応答、あるいは他の部分構造のふるまいを制御し合うような機構への拡張も重要と考えられる^{(7)~(9)}。

最後に、本研究に対して熱心に御討論、協力をいただきました堂下修司教授（現、京都大学）、米崎直樹君はじめ、当研究室の皆様には感謝いたします。また、本研究の一部は、文部省科学研究費の援助を受けていることを記し、謝意を表します。

文 献

- (1) 榎本, 富田, 米崎; "サンプル記号列の整合関係を考慮したオートマトンの一構成法", 信学会オートマトン研資(1973-11)

- (2) 榎本, 富田, 堂下, 牧野; "サンフォル記号列を識別ある決定性FSAの構成法— k 階層法— 昭48. 信学全大 1313 (1973-03)
- (3) 榎本, 堂下, 富田, 山口; "句構造サンフォルパターンを識別あるオートマトンの構成について," 信学会オートマトン研資 (1970-03)
- (4) H. Enomoto, E. Tomita and S. Doshita; "Synthesis of automata that recognize given strings and characterization of automata by representative sets of strings." 1-st USA-Japan Computer Conf. Proc. p.21 (Oct. 1972)
- (5) A.W. Biermann; "An interactive finite-state language learner": *ibid.* p.13
- (6) 榎本, 富田, 堂下; "オートマトンのチェック修正用記号列の生成", 信学会オートマトン・パターン研資 (1973-01)
- (7) 榎本, 富田, 小谷野; "一般化順序機械の代表入出力記号列集合", 昭49. 信学全大 1491 (1974-07)
- (8) 榎本, 富田, 根岸; "相互作用有限オートマトンシステムのテキスト表現", 昭49 信学全大 1492 (1974-07)
- (9) 阿部; "記号系列パターンを識別あるあるオートマトンの構成法について," 信学会パターン研資 (1974-05)