P-モデル定常環境のための新しい可変階層構造学習アルゴリズム

最上 義夫*・馬場 則夫**・松下 正樹***

New Algorithm of Variable Hierarchical Structure Learning Automata for P-model Stationary Random Environment

Yoshio MOGAMI*, Norio BABA** and Masaki MATSUSHITA***

In this paper, a new learning algorithm for variable hierarchical structure learning automata operating in a P-model stationary random environment is constructed by extending the estimator algorithm proposed by Thathachar and Sastry. The learning property of our algorithm is considered theoretically, and it is proved that the optimal path probability can be approached 1 as much as possible by using our algorithm. In numerical simulation, the average number of iterations of our algorithm is compared with that of the variable hierarchical structure learning algorithm of L_{R-I} type proposed by Mogami and Baba, and it is shown that our algorithm can find the optimal path after the smaller number of iterations than that of the algorithm of L_{R-I} type.

Key Words: variable hierarchical structure learning automata, estimator algorithm, reward parameter, sample estimate of reward probability, stationary random environment

1. ま え が き

学習オートマトンは、システムそのものに関する情報やその動作している環境に関する情報が先験的にはほとんど明らかでない問題、すなわち、不確実性が大きい問題に対して適用される。このとき、システムそのものとそれが動作する環境とを合わせて改めて環境と呼ぶことにすると、この環境はそれに関する情報がほとんど明らかでないことから未知環境ということになる。そして、学習オートマトンは、この未知環境に対してある動作を出力し、その出力とそれに対して未知環境から与えられる評価 (reward) とに基づいて学習を行い、この学習サイクルを繰り返すことによって最終的に未知環境から与えられる reward が出来るだけ大きい動作 (最適動作) を見い出そうとするものである。すなわち、学習オートマトンは未知環境の自律的学習システムであるといえる。

学習オートマトンの研究は、Tsetlin¹⁾によって最初に行われ、有限決定性オートマトンが学習特性を持つことが発見された。そして、その後数年して、Varshavskii and Vorontsova²⁾により確率オートマトンも学習特性を持つことが指摘され、それ以後、学習オートマトンに関する研究が活発に行われてき

た^{3)~7)}。そして、Thathachar and Ramakrishnan⁸⁾および Ramakrishnan⁹⁾が、学習オートマトンを木構造に組み合わせた階層構造学習オートマトンを提案し、P-モデル定常環境中で良好な学習特性を持つことを示した。また、Thathachar and Sastry は、P-モデル定常環境の reward 確率のサンプル平均値を用いる Estimator Algorithm を提案し¹⁰⁾、さらに、S-モデル定常環境中で動作する階層構造学習オートマトンへの拡張を行った¹¹⁾。

ところで、ある問題に学習オートマトンを適用するときはその出力となり得るものを学習対象と呼び、また、それらの集合を対象集合と呼ぶことにすると、従来行われてきた研究は、対象集合の要素が学習開始時にすべて与えられて確定しており、その reward 確率のみが未知である場合、すなわち、所定の学習を遂行するためには、学習オートマトンにどのような出力を合計何個持たせればよいかが事前に明確である場合について考察している。しかし、実際の問題に学習オートマトンを適用することを考えるとき、reward 確率のみならず対象集合の要素数および内容もが未知であるような周囲環境中において動作する学習オートマトンについても考察することが必要である。このような環境の例は、移動ロボットの迷路通過問題に見られる¹²⁾。

このような場合に対して、最上¹²⁾が、P-モデル定常環境中で動作する可変階層構造学習オートマトンを提案し、その学習特性の考察を行った。また、最上、馬場¹³⁾が、P-モデル非定常複数教師環境に対しても可変階層構造学習オートマトンを適用した。

* 徳島大学 徳島市南常三島町 2-1

** 大阪教育大学 柏原市旭ヶ丘 4

*** 三菱電機 (株) 姫路市千代田町 840

* The University of Tokushima, Tokushima

** Osaka Educational University, Kashiwara

*** Mitsubishi Electric Co., Himeji

(Received May 27, 1996)

(Revised December 15, 1997)

ところで、一般に学習を行う場合には、学習の結果獲得されるべき目標物（学習目標）が設定される。それらの集合として目標集合を定義すると、対象集合の要素数および内容が未知である場合には、対象集合の中に学習目標ではないものが含まれることがある。すなわち、目標集合は対象集合の部分集合となる。また、環境から成功の評価を最も得やすい学習目標を最良目標と呼ぶ。さらに、いま、何らかの方法で対象集合が事前に確定できたとすると、全ての学習対象にそれぞれパス^(注1)を割り当てた完全な形の階層構造が構成できる。本論文では、このような階層構造を完全構造と呼ぶ。また、完全構造から1個以上の学習オートマトンが欠けた階層構造を部分構造と呼ぶ。さらに、完全構造上のパスを完全パス、部分構造上のパスで完全パスではないものを部分パス、学習目標を出力とするパスを目標パス、最良目標を出力とするパスを最良目標パスと呼ぶ（目標パスは、全て完全パスである）。そして、つぎの条件が成立するものとする。

[C.1] 対象集合は、学習開始時には確定していない。

[C.2] 最良目標は唯一に存在し、最良目標パスは完全構造の最適パスである。

[C.3] 可変階層構造学習オートマトンは、収束したパスが目標パスであるかどうかは認識できる。

本論文では、条件 C.1~C.3 が成り立つ場合において、対象集合の要素数および内容とそれに対する reward 確率との両者が未知であるような P-モデル定常環境中で動作する可変階層構造学習オートマトンについて考察する。このような可変階層構造学習オートマトンについては文献 12) において報告されているが、これは現在の入出力のみに基づいて学習を進めていくものである。ところで、学習オートマトンを実問題に適用するときには、出力数の増加につれてその学習回数が飛躍的に増加するという問題点が存在する。この問題点を克服するために、現在の入力のみならず過去に得た入力をも学習に利用することによって、学習回数の減少を計ることが考えられる。そこで、このような新しい可変階層構造学習アルゴリズムが、Thathachar and Sastry によって提案された Estimator Algorithm を拡張することによって得られることを示す。さらに、その学習特性についても考察する。

2. 可変階層構造学習オートマトンと未知環境

可変階層構造学習オートマトンの概略を Fig. 1 に示す。その構造であるが、各学習オートマトンはそれぞれ r 個の動作を持ち、レベル 1 には 1 個の学習オートマトン Z が存在する。そして、その動作 α_{i_1} ($i_1 = 1, 2, \dots, r$) の下には、それぞれ 1 個の学習オートマトン Z_{i_1} が結合可能であり、よって、レベル 2 には最大 r 個の学習オートマトンが存在可能である。同様に、レベル 2 の各動作 $\alpha_{i_1 i_2}$ ($i_1, i_2 = 1, 2, \dots, r$) にも、それぞれ 1 個の学習オートマトン $Z_{i_1 i_2}$ が結合可能

Fig. 1 Variable Hierarchical Structure Learning Automata

であり、その結果、レベル 3 には最大 r^2 個の学習オートマトンが存在可能となる。このようにして、一般に、レベル s には r^{s-1} 個の学習オートマトンが存在可能であり、レベル s の学習オートマトンを $Z_{i_1 i_2 \dots i_{s-1}}$ 、その動作を $\alpha_{i_1 \dots i_{s-1} i_s}$ ($i_s = 1, 2, \dots, r$) と表記する。また、レベル 1 から最下レベル N に至る動作の列 $\{\alpha_{i_1}, \alpha_{i_1 i_2}, \dots, \alpha_{i_1 i_2 \dots i_N}\}$ をパスと呼んで $\phi_{i_1 i_2 \dots i_N}$ と記述する。

可変階層構造学習オートマトンにおいては、ある階層構造での学習と階層構造の更新との 2 つの作業が行われるが、この 2 つの作業がどのように進められるのかについて Fig. 1 に従って以下に述べる。(a) 既知の情報に対応する初期構造 Σ^0 として、 n_0 レベル階層構造学習オートマトンを構成し学習を開始する (Fig. 1 では、 $n_0 = 1$)。初期時刻 $t = 0$ における動作確

(注 1) 階層構造のレベル 1 の学習オートマトンから最下レベルの学習オートマトンに至る動作の列をパスと呼ぶ。

率は、すべてのレベルの学習オートマトンについて等しいとする(つまり、各学習オートマトンの動作数を r とすると $1/r$ となる)。さて、今、レベル1の学習オートマトン Z が時刻 t において、動作確率分布 $(p_1(t), p_2(t), \dots, p_r(t))'$ に基づいて動作 α_{j_1} を選択したとする。すると、 $\phi(t) = \phi_{j_1} (= \alpha_{j_1})$ が環境に出力される。この出力に対して、P-モデル未知環境は reward 確率 d_{j_1} に従って reward $\beta(t) = 1$ (nonpenalty) を、また、確率 $1 - d_{j_1}$ に従って $\beta(t) = 0$ (penalty) を学習オートマトンに与える。そして、この返答を用いて Z は学習アルゴリズムに従って動作確率分布を更新し、時刻を $t+1$ として新たなパスを選択する。このような学習を繰り返した結果、ある時刻において Σ^0 の最適パスとして ϕ_2 が選択されたとする。(b) ここで、 ϕ_2 が目標パスではないとすると、可変階層構造学習オートマトンは、 α_2 の下に新たに1個の学習オートマトン Z_2 を追加して構造 Σ^1 を構成し、学習を継続する。このとき、レベル2の学習オートマトン Z_{j_1} (ここでは Z_2) は、レベル1で動作 α_{j_1} (ここでは α_2) が選択されたときのみ動作し、 Z と同様、時刻 t における動作確率分布 $(p_{j_11}(t), p_{j_12}(t), \dots, p_{j_1r}(t))'$ に基づいて動作 $\alpha_{j_1j_2}$ を選択する。ここで、レベル1から最下レベルに至る動作の列 $\{\alpha_{j_1}, \alpha_{j_1j_2}\}$ 、すなわち、パス $\phi_{j_1j_2}$ が選択されたとする。すると、このパスに対応する出力が環境に与えられ、それに対して、環境は reward 確率 $d_{j_1j_2}$ に基づいて reward $\beta(t) = 1$ を、また、確率 $1 - d_{j_1j_2}$ に基づいて $\beta(t) = 0$ を学習オートマトンに与える。そして、この返答に基づいて学習オートマトンはその動作確率分布を更新し、時刻を $t+1$ として新たなパスを選択する。このような学習を繰り返した結果、ある時刻において Σ^1 の最適パスとして ϕ_{21} が選択されたとする。(c) ϕ_{21} も目標パスではないとすると、可変階層構造学習オートマトンは α_{21} の下に新たに1個の学習オートマトン Z_{21} を追加して構造 Σ^2 を構成し、学習を継続する。

以上のように学習が進められて行き、ある構造 Σ^M で最適パスとして目標パスが得られると、可変階層構造学習オートマトンは学習を終了する。ここで、初期構造から m 回の変更を経た構造を Σ^m ($m = 0, 1, \dots$) と記述し、構造 Σ^m のパスの総数を Ψ^m 、Fig. 1 のように階層構造を図示したときに、左から数えて k 番目のパスを $\phi_{i_1i_2 \dots i_{N_k^m}}$ ($k = 1, 2, \dots, \Psi^m$) と記述する(ただし、 N_k^m はこのパスのレベル数である)。そして、これに対応するパス確率は

$$\pi_{i_1i_2 \dots i_{N_k^m}}(t) = p_{i_1}(t)p_{i_1i_2}(t) \dots p_{i_1i_2 \dots i_{N_k^m}}(t) \quad (1)$$

である。また、構造 Σ^m 上のパス確率の最大値を Σ^m 上の最大パス確率と呼び、これに対応するパスを Σ^m 上の最大パスと呼ぶ。

つぎに、このような可変階層構造学習オートマトンが動作する未知環境としてはつぎのような特性を持つP-モデル定常環境を考える。

[N.1] 構造 Σ^m において、つぎの不等式を満たすパス $\phi_{j_1^*j_2^* \dots j_{N_k^m}^*}$ が唯一に存在する。

$$d_{j_1^*j_2^* \dots j_{N_k^m}^*} > d_{i_1i_2 \dots i_{N_k^m}}$$

ただし、 $k = 1, 2, \dots, \Psi^m$ 、かつ、 $k \neq a$ である。このとき、パス $\phi_{j_1^*j_2^* \dots j_{N_k^m}^*}$ が構造 Σ^m の最適パスである。

[N.2] 構造 Σ^m のパス $\phi_{j_1j_2 \dots j_{N_k^m}}$ に学習オートマトンを追加するとする。このとき、あらゆる $j_{N_k^m+1}$ ($= 1, 2, \dots, r$) に対して、

$$d_{j_1j_2 \dots j_{N_k^m}} \geq d_{j_1j_2 \dots j_{N_k^m}j_{N_k^m+1}}$$

となる。

ここで、完全構造を Σ^P とし最良目標パスを $\phi_{h_1^*h_2^* \dots h_{N_a^P}^*}$ とすると、

$$d_{h_1^*h_2^* \dots h_{N_a^P}^*} > d_{i_1i_2 \dots i_{N_k^P}} \quad (k = 1, 2, \dots, \Psi^P; k \neq a) \quad (2)$$

が成り立つ。そして、可変階層構造学習オートマトンの学習目的はこの最良目標パスを見出すことである。

このとき、任意の部分構造における最適パスに関してつぎの定理を得る。

《定理1》 条件C.1~C.3を満たす問題において、特性N.1およびN.2が成り立つものとする。このとき、任意の部分構造において最適パスとなるのは、部分パスか、あるいは、最良目標パスかのいずれかである。

(証明) 文献12)と同様にして出来る。

3. reward パラメータと学習アルゴリズム

可変階層構造学習オートマトンに Estimator Algorithm を適用するわけであるが、このとき、Estimator Algorithm は各動作に対応する reward 確率のサンプル平均値に基づいて学習を進める学習アルゴリズムであること¹⁰⁾に注意すると、階層構造上のすべての学習オートマトンのすべての動作に reward 確率のサンプル平均値を対応させることが必要である。しかしここで考察している可変階層構造学習オートマトンは全体入出力であるので、1つのパスに対して1つの reward 確率に対応しているだけであって、各動作に reward 確率は対応していない。このことから、すべての動作に reward 確率のサンプル平均値を対応させることはできない。そこで、パスに対応している reward 確率のサンプル平均値を用いてすべての動作に対して reward パラメータ $\hat{e}(t)$ を生成し、この値を reward 確率のサンプル平均値の代わりに用いることにする。

いま、任意のパス $\phi_{i_1i_2 \dots i_{N_k^m}}$ に対応している reward 確率 $d_{i_1i_2 \dots i_{N_k^m}}$ のサンプル平均値を $\hat{d}_{i_1i_2 \dots i_{N_k^m}}(t)$ と記述することになると、これは次式で計算される。

時刻 t までに任意のパス $\phi_{i_1i_2 \dots i_{N_k^m}}$ が選択された回数を $z_{i_1i_2 \dots i_{N_k^m}}(t)$ とし、得た reward の合計を $R_{i_1i_2 \dots i_{N_k^m}}(t)$ とする。そして、時刻 t において、あるパス $\phi_{j_1j_2 \dots j_{N_k^m}}$ が選択されて、未知環境から reward $\beta(t)$ を得たとすると、次式

が成り立つ.

$$z_{j_1 j_2 \dots j_{N_k^m}}(t+1) = z_{j_1 j_2 \dots j_{N_k^m}}(t) + 1 \quad (3)$$

$$z_{i_1 i_2 \dots i_{N_l^m}}(t+1) = z_{i_1 i_2 \dots i_{N_l^m}}(t) \quad (l = 1, 2, \dots, \Psi^m; l \neq k) \quad (4)$$

$$R_{j_1 j_2 \dots j_{N_k^m}}(t+1) = R_{j_1 j_2 \dots j_{N_k^m}}(t) + \beta(t) \quad (5)$$

$$R_{i_1 i_2 \dots i_{N_l^m}}(t+1) = R_{i_1 i_2 \dots i_{N_l^m}}(t) \quad (l = 1, 2, \dots, \Psi^m; l \neq k) \quad (6)$$

これらの値を用いて $\hat{d}_{i_1 i_2 \dots i_{N_l^m}}(t+1)$ がつぎのように計算される.

$$\hat{d}_{i_1 i_2 \dots i_{N_l^m}}(t+1) = \frac{R_{i_1 i_2 \dots i_{N_l^m}}(t+1)}{z_{i_1 i_2 \dots i_{N_l^m}}(t+1)} \quad (l = 1, 2, \dots, \Psi^m) \quad (7)$$

そこでこの値を用いて, 本論文では reward パラメータ $\hat{e}(t)$ をつぎのように与えることにする.

(i) 最下レベルの動作 $\alpha_{i_1 i_2 \dots i_{N_l^m}}$ に対するもの:

$$\hat{e}_{i_1 i_2 \dots i_{N_l^m}}(t) = \hat{d}_{i_1 i_2 \dots i_{N_l^m}}(t) \quad (8)$$

(ii) 最下レベルにない動作 $\alpha_{i_1 i_2 \dots i_s}$ に対するもの:

$$\hat{e}_{i_1 i_2 \dots i_s}(t) = \max_{i_{s+1}} \{ \hat{e}_{i_1 i_2 \dots i_s i_{s+1}}(t) \} \quad (9)$$

(注意) (8) 式および (9) 式より, $\hat{e}_{i_1 i_2 \dots i_s}(t)$ は動作 $\alpha_{i_1 i_2 \dots i_s}$ を通るすべてのパスを考えたとき, これらのパスに対応する reward 確率のサンプル平均値の中の最大値ということになる.

このような reward パラメータ $\hat{e}$ を, Estimator Algorithm における reward 確率のサンプル平均値 $\hat{d}$ の代わりに用いることによって, Thathachar and Sastry¹⁰⁾ の Estimator Algorithm を拡張した可変階層構造学習アルゴリズムが以下のように構成される.

《可変階層構造学習アルゴリズム》

Step 1. 初期時刻 $t = 0$ として, 初期構造 Σ^0 (n_0 レベル階層構造学習オートマトン) を構成する ($m = 0, \Psi^0 = r^{n_0}, p_{j_1 j_2 \dots j_s}(0) = 1/r$ ($j_k = 1, 2, \dots, r; k = 1, 2, \dots, s; s = 1, 2, \dots, n_0$)). また, 収束判定値 ρ_0 (1 に近い小数) を与える.

Step 2. 最大パス $\tilde{\phi}_{i_1 i_2 \dots i_{N_l^m}}$ 上の各動作の動作確率を $\tilde{p}_{i_1 i_2 \dots i_s}(t)$ ($s = 1, 2, \dots, N_l^m$) とする.

(i) $\tilde{p}_{i_1 i_2 \dots i_s}(t) > \rho_m$ ($s = 1, 2, \dots, N_l^m$) のとき: 最大パスが目標パスであるならば, 本アルゴリズムを終了する. それ以外の時は, Step 4 へ.

(ii) (i) 以外のとき: Step 3 へ.

Step 3. 構造 Σ^m のもとで, 下記の強化法を用いて学習を行う.

[強化法] 時刻 t において, パス $\phi_{j_1 j_2 \dots j_s \dots j_{N_k^m}}$ が選択され, これに対して未知環境から $\beta(t)$ なる返答が与えられたとする.

(i) パス $\phi_{j_1 j_2 \dots j_s \dots j_{N_k^m}}$ 上の各学習オートマトン $Z_{j_1 j_2 \dots j_{s-1}}$ ($s = 1, 2, \dots, N_k^m$) の動作確率分布を次式によって更新する.

$$p_{j_1 j_2 \dots j_{s-1} i_s}(t+1) = p_{j_1 j_2 \dots j_{s-1} i_s}(t) - \lambda_s [f(\hat{e}_{j_1 j_2 \dots j_s}(t)) - f(\hat{e}_{j_1 j_2 \dots j_{s-1} i_s}(t))] \times \left[u_{j_1 j_2 \dots j_{s-1} i_s}^{j_1 j_2 \dots j_s}(t) p_{j_1 j_2 \dots j_{s-1} i_s}(t) + \frac{u_{j_1 j_2 \dots j_{s-1} i_s}^{j_1 j_2 \dots j_s}(t) (1 - p_{j_1 j_2 \dots j_{s-1} i_s}(t)) p_{j_1 j_2 \dots j_s}(t)}{r - 1} \right] \quad (i_s \neq j_s) \quad (10)$$

$$p_{j_1 j_2 \dots j_s}(t+1) = 1 - \sum_{i_s=1, i_s \neq j_s}^r p_{j_1 j_2 \dots j_{s-1} i_s}(t+1) \quad (11)$$

ただし, λ_s は $0 < \lambda_s < 1$ を満たす実数であり, f は $[0, 1] \rightarrow [0, 1]$ なる単調増加関数, u は

$$u_{j_1 j_2 \dots j_{s-1} i_s}^{j_1 j_2 \dots j_s}(t) = \begin{cases} 1 & \hat{e}_{j_1 j_2 \dots j_s}(t) > \hat{e}_{j_1 j_2 \dots j_{s-1} i_s}(t) \text{ のとき} \\ 0 & \hat{e}_{j_1 j_2 \dots j_s}(t) \leq \hat{e}_{j_1 j_2 \dots j_{s-1} i_s}(t) \text{ のとき} \end{cases} \quad (12)$$

なる関数である.

(ii) (3) 式~(9) 式に従って reward パラメータ

$\hat{e}_{j_1 j_2 \dots j_s}(t+1)$ を計算する.

(iii) $t = t+1$ として, Step 2 へ.

(ただし, パス $\phi_{j_1 j_2 \dots j_{N_k^m}}$ 上にない学習オートマトンの動作確率分布および reward パラメータの値はまったく変更されない.)

Step 4. Σ^m 上の最大パスの下に学習オートマトンを 1 個追加することによって, 構造を Σ^m から Σ^{m+1} に更新する.

Step 5. 新構造 Σ^{m+1} 上の全てのパスのパス確率を $1/\Psi^{m+1}$ と置く.

Step 6. パス確率をもとに, パス上の各学習オートマトンの動作確率を計算し, この値を新構造での初期動作確率とする. $m = m+1$ とし, 収束判定値 ρ_m (1 に近い小数) を与えて, Step 2 へ.

(注意) Step 6 における階層構造上の各学習オートマトンの動作確率を計算する方法を述べる. いま, レベル s の学習オートマトン $Z_{j_1 j_2 \dots j_{s-1}}$ のある動作 $\alpha_{j_1 j_2 \dots j_s}$ に対する動作確率 $p_{j_1 j_2 \dots j_s}(t)$ を計算するとしよう. まず, 動作 $\alpha_{j_1 j_2 \dots j_s}$ を通るパスの数を $u_{j_1 j_2 \dots j_s}$ 本とし, また, 学習オートマトン $Z_{j_1 j_2 \dots j_{s-1}}$ を通るパスの数を $v_{j_1 j_2 \dots j_{s-1}}$ 本とする. このとき, 動作確率 $p_{j_1 j_2 \dots j_s}(t)$ はつぎのように計算する.

$$p_{j_1 j_2 \dots j_s}(t) = \frac{u_{j_1 j_2 \dots j_s}}{v_{j_1 j_2 \dots j_{s-1}}} \quad (13)$$

4. 学習特性

reward 確率のサンプル平均値に関して、文献 10), 11), 14) と同様にしてつぎの2つの補題が得られる.

[補題 1] 構造 Σ^m において、任意の実数 $\delta^m > 0$ と正整数 $K^m < \infty$ に対して、可変階層構造学習アルゴリズムのもとで次式を満たす正の実数 $\{\lambda_s^* : s = 1, 2, \dots, N_l^m, l = 1, 2, \dots, \Psi^m\}$ および時刻 $t^m < \infty$ が存在する. すなわち、任意の $\lambda_s \in (0, \lambda_s^*)$ に対して、

$$\text{Prob}\{\text{すべてのパスが時刻 } t \text{ までにそれぞれ少なくとも } K^m \text{ 回選択される}\} > 1 - \delta^m \quad \text{for } \forall t \geq t^m \quad (14)$$

[補題 2] 構造 Σ^m における最適パスに対する reward 確率を $d_{j_1^* j_2^* \dots j_{N_a^m}^*}$ とする. このとき、任意の実数 $\Delta^m > 0$ および $\delta^m > 0$ に対して、可変階層構造学習アルゴリズムのもとで次式を満たす正の実数 $\{\lambda_s^* : s = 1, 2, \dots, N_l^m, l = 1, 2, \dots, \Psi^m\}$ および時刻 $t^m < \infty$ が存在する. すなわち、任意の $\lambda_s \in (0, \lambda_s^*)$ に対して、

$$\text{Prob}\{\hat{d}_{j_1^* j_2^* \dots j_{N_a^m}^*}(t) > \hat{d}_{i_1 i_2 \dots i_{N_l^m}}(t)\} > 1 - \delta^m \quad \text{for } \forall t \geq t^m \quad (l \neq a) \quad (15)$$

このとき、補題 1 および補題 2 のもとでつぎの定理が成り立つ.

《定理 2》 reward パラメータ $\hat{e}$ が (8) 式および (9) 式で与えられるものとする. このとき、構造 Σ^m の最適パス上の各学習オートマトン $Z_{j_1^* j_2^* \dots j_{s-1}^*}$ ($s = 1, 2, \dots, N_a^m$) において、任意の実数 $\delta^m > 0$ に対して、次式を満たす正の実数 $\{\lambda_s^* : s = 1, 2, \dots, N_a^m\}$ および時刻 $t^m < \infty$ が存在する. すなわち、任意の $\lambda_s \in (0, \lambda_s^*)$ に対して、つぎの関係式が成立する.

$$\text{Prob}\{\hat{e}_{j_1^* j_2^* \dots j_{s-1}^* j_s^*}(t) > \hat{e}_{j_1^* j_2^* \dots j_{s-1}^* i_s^*}(t)\} > 1 - \delta^m \quad \text{for } \forall t \geq t^m \quad (16)$$

ただし、 $i_s = 1, 2, \dots, r; i_s \neq j_s^*$.

(証明) 最適パス $\phi_{j_1^* j_2^* \dots j_{N_a^m}^*}$ 上の学習オートマトンに着目しよう. まず、最下レベル N_a^m の学習オートマトン $Z_{j_1^* j_2^* \dots j_{N_a^m-1}^*}$ においては、(8) 式より、動作 $\alpha_{j_1^* j_2^* \dots j_{N_a^m-1}^* j_{N_a^m}^*}$ に対する reward パラメータ $\hat{e}_{j_1^* j_2^* \dots j_{N_a^m-1}^* j_{N_a^m}^*}(t)$ として、パス $\phi_{j_1^* j_2^* \dots j_{N_a^m-1}^* j_{N_a^m}^*}$ に対応する reward 確率のサンプル平均値が格納されるため、

$$\hat{e}_{j_1^* j_2^* \dots j_{N_a^m-1}^* j_{N_a^m}^*}(t) = \hat{d}_{j_1^* j_2^* \dots j_{N_a^m-1}^* j_{N_a^m}^*}(t) \quad (17)$$

が成立する. また、 $Z_{j_1^* j_2^* \dots j_{N_a^m-1}^*}$ の $\alpha_{j_1^* j_2^* \dots j_{N_a^m-1}^* j_{N_a^m}^*}$ 以外の動作 $\alpha_{j_1^* j_2^* \dots j_{N_a^m-1}^* l_1}$ ($l_1 = 1, 2, \dots, r; l_1 \neq j_{N_a^m}^*$) に対しても、(8) 式および (9) 式より、次式が成立する.

$\alpha_{j_1^* j_2^* \dots j_{N_a^m-1}^* l_1}$ が最下レベルの動作のとき:

$$\hat{e}_{j_1^* j_2^* \dots j_{N_a^m-1}^* l_1}(t) = \hat{d}_{j_1^* j_2^* \dots j_{N_a^m-1}^* l_1}(t) \quad (18)$$

$\alpha_{j_1^* j_2^* \dots j_{N_a^m-1}^* l_1}$ が最下レベルの動作でないとき:

$$\hat{e}_{j_1^* j_2^* \dots j_{N_a^m-1}^* l_1}(t) = \max_{l_2} \{\hat{e}_{j_1^* j_2^* \dots j_{N_a^m-1}^* l_1 l_2}(t)\} \quad (19)$$

ここで、(19) 式の右辺において、 $\hat{e}_{j_1^* j_2^* \dots j_{N_a^m-1}^* l_1 l_2}(t)$ が最下レベルの動作に対するものであるならば、(8) 式より、

$$\hat{e}_{j_1^* j_2^* \dots j_{N_a^m-1}^* l_1 l_2}(t) = \hat{d}_{j_1^* j_2^* \dots j_{N_a^m-1}^* l_1 l_2}(t) \quad (20)$$

となり、また、最下レベルでない動作に対するものであるならば、(9) 式より、

$$\hat{e}_{j_1^* j_2^* \dots j_{N_a^m-1}^* l_1 l_2}(t) = \max_{l_3} \{\hat{e}_{j_1^* j_2^* \dots j_{N_a^m-1}^* l_1 l_2 l_3}(t)\} \quad (21)$$

となる. このような考察を最下レベルの動作に至るまで繰り返す. ここで、動作 $\alpha_{j_1^* j_2^* \dots j_{N_a^m-1}^* l_1 l_2 l_3}$ が最下レベルの動作であるとする、(8) 式および (21) 式より、(19) 式はつぎのようになる.

$$\begin{aligned} \hat{e}_{j_1^* j_2^* \dots j_{N_a^m-1}^* l_1}(t) &= \max_{l_2, l_3} \{\hat{e}_{j_1^* j_2^* \dots j_{N_a^m-1}^* l_1 l_2 l_3}(t)\} \\ &= \max_{l_2, l_3} \{\hat{d}_{j_1^* j_2^* \dots j_{N_a^m-1}^* l_1 l_2 l_3}(t)\} \end{aligned} \quad (22)$$

よって、(15) 式、(17) 式、(18) 式および (22) 式に注意すると、つぎの関係が成立する.

$$\text{Prob}\{\hat{e}_{j_1^* j_2^* \dots j_{N_a^m-1}^* j_{N_a^m}^*}(t) > \hat{e}_{j_1^* j_2^* \dots j_{N_a^m-1}^* l_1}(t)\} > 1 - \delta^m \quad \text{for } \forall t \geq t^m \quad (l_1 = 1, 2, \dots, r; l_1 \neq j_{N_a^m}^*) \quad (23)$$

つぎに、学習オートマトン $Z_{j_1^* j_2^* \dots j_{N_a^m-1}^*}$ が結合されている動作 $\alpha_{j_1^* j_2^* \dots j_{N_a^m-1}^*}$ を持つレベル ($N_a^m - 1$) の学習オートマトン $Z_{j_1^* j_2^* \dots j_{N_a^m-2}^*}$ について考える. (8) 式、(9) 式および (23) 式より、動作 $\alpha_{j_1^* j_2^* \dots j_{N_a^m-1}^*}$ に対して、つぎの関係が $1 - \delta^m$ 以上の確率で成立する.

$$\begin{aligned} \hat{e}_{j_1^* j_2^* \dots j_{N_a^m-2}^* j_{N_a^m-1}^*}(t) &= \max_{l_1} \{\hat{e}_{j_1^* j_2^* \dots j_{N_a^m-1}^* l_1}(t)\} \\ &= \hat{e}_{j_1^* j_2^* \dots j_{N_a^m-1}^* j_{N_a^m}^*}(t) \\ &= \hat{d}_{j_1^* j_2^* \dots j_{N_a^m-1}^* j_{N_a^m}^*}(t) \quad \text{for } \forall t \geq t^m \end{aligned} \quad (24)$$

また、 $Z_{j_1^* j_2^* \dots j_{N_a^m-2}^*}$ の $\alpha_{j_1^* j_2^* \dots j_{N_a^m-2}^* j_{N_a^m-1}^*}$ 以外の動作 $\alpha_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1}$ ($l_1 = 1, 2, \dots, r; l_1 \neq j_{N_a^m-1}^*$) に対しても、(8) 式および (9) 式より、次式が成立する.

$\alpha_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1}$ が最下レベルの動作のとき:

$$\hat{e}_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1}(t) = \hat{d}_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1}(t) \quad (25)$$

$\alpha_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1}$ が最下レベルの動作でないとき:

$$\hat{e}_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1}(t) = \max_{l_2} \{\hat{e}_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1 l_2}(t)\} \quad (26)$$

ここで、(26) 式の右辺において、(20) 式および (21) 式と同様な式が成り立ち、動作 $\alpha_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1 l_2 l_3}$ が最下レベルの動作であるとする、(8) 式および (26) 式より次式を得る.

$$\hat{e}_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1}(t) = \max_{l_2, l_3} \{\hat{e}_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1 l_2 l_3}(t)\}$$

$$= \max_{l_2, l_3} \{ \hat{d}_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1 l_2 l_3}(t) \} \quad (27)$$

よって, (15) 式に注意すると, (24) 式, (25) 式および (27) 式より, レベル $(N_a^m - 1)$ の学習オートマトン $Z_{j_1^* j_2^* \dots j_{N_a^m-2}^*}$ についても, つぎの関係が成立する.

$$\text{Prob}\{\hat{e}_{j_1^* \dots j_{N_a^m-2}^* j_{N_a^m-1}^*}(t) > \hat{e}_{j_1^* \dots j_{N_a^m-2}^* l_1}(t)\} > 1 - \delta^m \\ \text{for } \forall t \geq t^m \quad (l_1 = 1, 2, \dots, r; l_1 \neq j_{N_a^m-1}^*) \quad (28)$$

以下同様にして, 一般にレベル s の学習オートマトン $Z_{j_1^* j_2^* \dots j_{s-1}^*}$ において, 次式が成立する.

$$\text{Prob}\{\hat{e}_{j_1^* j_2^* \dots j_{s-1}^* j_s^*}(t) > \hat{e}_{j_1^* j_2^* \dots j_{s-1}^* i_s}(t)\} > 1 - \delta^m \\ \text{for } \forall t \geq t^m$$

ただし, $i_s = 1, 2, \dots, r, i_s \neq j_s^*$.

Q.E.D.

このことから文献 11) の定理 2 と同様にして, 本可変階層構造学習アルゴリズムに対して, 下記の定理が成立する.

《定理 3》 構造 Σ^m において, 特性 N. 1 が成り立つものとする. このとき, 任意の $\varepsilon^m > 0$ および $\delta^m > 0$ に対して,

$$\text{Prob}\{|\pi_{j_1^* j_2^* \dots j_{N_a^m}^*}(t) - 1| < \varepsilon^m\} > 1 - \delta^m \\ \text{for } \forall t \geq t^m, \forall \lambda_s \in (0, \lambda_s^*) \quad (s = 1, 2, \dots, N_a^m) \quad (29)$$

なる正実数 $\lambda_1^*, \lambda_2^*, \dots, \lambda_{N_a^m}^*$ および t^m が存在する.

さらに, 定理 1 および定理 3 より, つぎの定理が得られる.

《定理 4》 条件 C. 1~C. 3 および特性 N. 1, N. 2 が成立するものとする. このとき, 初期構造として部分構造を採用するならば, 可変階層構造学習アルゴリズムによって最良目標パスが選択される確率は, $\{\lambda_s\}$ を適切に選ぶことによって, いくらでも 1 に近くすることができる.

(証明) いま, 構造 Σ^m の最適パスが選択される確率を Q^m と記述すると, 定理 3 の (29) 式より, 十分小さな ε^m と δ^m に対して, $\{\lambda_s\}$ を適切に選ぶことによって,

$$Q^m > 1 - \delta^m \quad (30)$$

を得る.

ところで, 可変階層構造学習アルゴリズムの初期構造として部分構造を採用するならば, つぎのように学習が進むとき, 最良目標パスが選択されることになる. すなわち, 「初期構造 Σ^0 (部分構造) から学習を開始し, Σ^0 の最適パスが選択される. このパスは, 定理 1 より, 最良目標パスかあるいは部分パスであるから, これが (最良) 目標パスなら学習は終了し, 部分パスならばオートマトンが追加されて, 構造 Σ^1 が構成される. そして, 学習が継続され, Σ^1 の最適パスが選択される. このパスも, 定理 1 より, 最良目標パスかあるいは部分パスであり, (最良) 目標パスなら学習は終了し, 部分パスならつぎなる構造が構成される. このように, 最適パスとしての部分パスの選択と構造の更新とを繰り返しながら, ついには, ある構造 Σ^M (部分構造と完全構造のいずれ

でもよい) において最良目標パスを最適パスとして選択して学習を終了する.]

よって, 可変階層構造学習アルゴリズムによって最良目標パスが選択される確率を Q と書くと, (30) 式より,

$$Q \triangleq \text{Prob}\{\text{構造 } \Sigma^0 \text{ の最適パスが選択される}\} \\ \times \text{Prob}\{\text{構造 } \Sigma^1 \text{ の最適パスが選択される}\} \\ \vdots \\ \times \text{Prob}\{\text{構造 } \Sigma^{M-1} \text{ の最適パスが選択される}\} \\ \times \text{Prob}\{\text{構造 } \Sigma^M \text{ の最適パスが選択される}\} \\ = Q^0 Q^1 \dots Q^M \\ > (1 - \delta^0)(1 - \delta^1) \dots (1 - \delta^M) \\ \geq 1 - (M + 1)\delta \quad (31)$$

となる. ただし, $\delta = \max\{\delta^0, \delta^1, \dots, \delta^M\}$ である. ところで, δ の値はいくらでも小さくとれることに注意しよう. すると, (31) 式の右辺はいくらでも 1 に近くすることが出来る. よって, 確率 Q の値はいくらでも 1 に近くすることが出来る.

Q.E.D.

5. 数値シミュレーション

本節では, P-モデル定常環境が 1 個存在する場合の可変階層構造学習オートマトンについて, 本論文において構築した可変階層構造学習アルゴリズムを適用した場合の学習速度と, 最上¹²⁾によって提案された L_{R-I} タイプの可変階層構造学習アルゴリズムを適用した場合の学習速度とを数値シミュレーションによって比較検討する.

まず, 各学習オートマトンは 2 個の動作を持っており, 完全構造は 6 から 10 の種々のレベル数を持つパス 944 本から成っていて, 最良目標パスは $\phi_{22222111212}$ とする. そして, 任意の部分構造における各パスに対する reward 確率は, 階層構造上のすべの動作に確率を対応させパス上の動作に対するこの確率の積として与えた. これを最良目標パス (Optimal Objective Path) $\phi_{22222111212}$ および第 2 位の目標パス (Second Objective Path) $\phi_{2222211}$ について示すと Table 1 のようになる. この Table から, 目標パス $\phi_{22222111212}$ および $\phi_{2222211}$ に対する reward 確率は, それぞれ $d_{22222111212} = 0.2496$ および $d_{2222211} = 0.1503$ となる. この他のパスに対する reward 確率もこれらと同様にして構成している. そして, 学習開始時には, レベル 1 の 2 個の動作の存在のみがわかっているものとする. 初期構造はレベル数 1 の階層構造学習オートマトンとなる.

この問題に本可変階層構造学習アルゴリズムおよび L_{R-I} タイプの可変階層構造学習アルゴリズムをそれぞれ適用した結果を Table 2 に示す. ここで, Our algorithm が本可変階層構造学習アルゴリズムによる結果であり, L_{R-I} algorithm が L_{R-I} タイプの可変階層構造学習アルゴリズムによる結

Table 1 Probabilities Corresponding to Each Action

Level	Optimal Objective Path	Second Objective Path
1	0.875	0.875
2	0.813	0.813
3	0.803	0.803
4	0.899	0.899
5	0.845	0.500
6	0.769	0.812
7	0.865	0.721
8	0.964	
9	0.974	
10	0.921	

Table 2 Simulation Results

Automaton	Average Number of Iterations	Ratio
Our algorithm	4419.76	0.63
L_{R-I} algorithm	7017.04	1

果である。また、Average Number of Iterations はそれぞれの学習オートマトンに対して 50 回づつ実験を行い、50 回すべてで最良目標パスを見出したときの平均学習回数を示している。この結果から、本可変階層構造学習アルゴリズムは L_{R-I} タイプの可変階層構造学習アルゴリズムと比較して 0.63 倍の学習回数で最良目標パスを見出すことが出来る。

6. む す び

本論文では、P-モデル定常環境中で動作する全体入出力の可変階層構造学習オートマトンについて考察し、reward 確率のサンプル平均値を用いる可変階層構造学習アルゴリズムを提案した。そして、特性 N. 1 および N. 2 を満たす環境のもとでは、初期構造として部分構造を採用するならば、本学習アルゴリズムによって最良目標パスが選択される確率は、パラメータ $\{\lambda_s\}$ を適切に選ぶことによっていくらかでも 1 に近くすることができることを理論的に示した。

また、数値シミュレーションによって、本可変階層構造学習アルゴリズムによる学習が、現在時刻の出力と reward とのみに基づいて学習を行う L_{R-I} タイプの可変階層構造学習アルゴリズムと比較して少ない学習回数で最良目標パスを見出すことを確認した。このことによって、学習の最初から現在時刻に至るまでの reward に基づいて得られた reward 確率のサンプル平均値を用いることが、可変階層構造学習オートマトンによる学習の高速化にとって有用であることがわかった。

参 考 文 献

- 1) M.L. Tsetlin: On the Behavior of Finite Automata in Random Media, *Avtomatika i Telemekhanika*, **22**-10, 1345/1354 (1961)
- 2) V.I. Varshavskii and I.P. Vorontsova: On the Behavior of Stochastic Automata with a Variable Structure, *Automation and Remote Control*, **24**, 327/333 (1963)
- 3) K.S. Narendra and S. Lakshmivarahan: Learning Automata - A Critique, *Journal of Cybernetics and Informa-*

tion Science - Special Issue on Learning Automata, 1-2-4, 53/65 (1977)

- 4) K.S. Narendra: Recent Developments in Learning Automata - Theory and Applications, *Proc. of the 3rd Yale Workshop on Applications of Adaptive Systems Theory*, 90/99 (1983)
- 5) N. Baba: New Topics in Learning Automata Theory and Applications, *Lecture Note in Control and Information Sciences*, Springer-Verlag (1985)
- 6) K.S. Narendra and R.M. Wheeler Jr.: Recent Developments in Learning Automata, in *Adaptive and Learning Systems*, 197/211, Plenum Press (1986)
- 7) K.S. Narendra and M.A.L. Thathachar: *Learning Automata An Introduction*, Prentice-Hall, Inc. (1989)
- 8) M.A.L. Thathachar and K.R. Ramakrishnan: A Hierarchical System of Learning Automata, *IEEE Trans. Systems, Man, and Cybernetics*, **SMC-11**-3, 236/241 (1981)
- 9) K.R. Ramakrishnan: Hierarchical Systems and Cooperative Games of Learning Automata, Ph.D. Thesis, Indian Institute of Science, Bangalore, India (1982)
- 10) M.A.L. Thathachar and P.S. Sastry: A Class of Rapidly Converging Algorithms for Learning Automata, *IEEE International Conference on Cybernetics and Society*, 602/606 (1984)
- 11) M.A.L. Thathachar and P.S. Sastry: A New Approach to the Design of Reinforcement Schemes for Learning Automata, *IEEE Trans. Systems, Man, and Cybernetics*, **SMC-15**-1, 168/175 (1985)
- 12) 最上義夫: 一人教師定常環境中で動作する可変階層構造学習オートマトン, 計測自動制御学会論文集, **25**-9, 1108/1025 (1989)
- 13) 最上義夫, 馬場則夫: 非定常複数教師環境中において動作する可変階層構造学習オートマトン, 電子情報通信学会論文誌, **J74-D-I**-7, 401/408 (1991)
- 14) M.A.L. Thathachar and P.S. Sastry: A Hierarchical System of Learning Automata That Can Learn the Globally Optimal Path, *Information Sciences*, **42**, 143/166 (1987)

[著 者 紹 介]

最 上 義 夫 (正会員)

1975 年, 京都工芸繊維大学大学院工芸学研究科修士課程修了。同年徳島大学工学部情報工学科助手, 1997 年同大学知能情報工学科講師, 現在に至る。学習オートマトン, 強化学習, 確率分布定数システムの観測器・制御器の最適配置問題の研究に従事。京都大学博士(工学)。計測自動制御学会, システム制御情報学会, 電子情報通信学会会員。

馬 場 則 夫 (正会員)

1974 年, 京都大学大学院工学研究科博士課程数理工学専攻修了。同年徳島大学工学部情報工学科助手, 講師、助教授を経て, 88 年同大学知能情報工学科教授, 90 年大阪教育大学教育学部教養学科情報科学専攻教授, 現在に至る。ニューラルネット, ゲームング, 強化学習, GAs, 数理計画法, 学習オートマトン等の研究に従事。工学博士。IEEE, INNS, 計測自動制御学会, システム制御情報学会, 日本シミュレーションゲーミング学会, 日本機械学会, オペレーションズ・リサーチ学会等の会員。

松 下 正 樹

1996 年，徳島大学工学研究科博士前期課程知能情報工学専攻修了。同年三菱電機(株)入社，現在に至る。可変階層構造学習オートマトンの研究に従事。

|||||