412　確率的学習オートマトンの機械制御への応用
— 倒立振子の振り上げ制御 —

北大工　○坂東　友則，　三上　貞芳，　嘉数　侑昇

要旨

確率的学習オートマトン(SLA)は，環境からの賞罰信号によって，その行動選択確率を変え，環境に適応する確率的システムである．そこで，本論文では，SLAの機械制御への応用を考え，その具体例として，倒立振子の振り上げ制御を試みる．

1　はじめに

一般の機械制御問題において，非線形要素を含む系は制御が困難である．そのため，このような系を簡単に制御する方法を確立することは重要である．一方，確率的学習オートマトン(SLA)は環境からの賞罰信号によって，行動選択確率を変え，環境に適応していく確率的システムである．そこで，SLAの機械制御への応用を考え，その具体例として，倒立振子を取り上げる．

倒立振子の制御を取り上げたのは，その性質が二足歩行と類似しており [BAR92]，ロボティックスへの応用が期待でき，また，近年，様々な方法 [小池 91],[松浦 91] を用いたアプローチが行われているためである．

2　制御システム

制御システムは図1のような構成で，制御対象，状態観測関数，評価者，SLAの四つの部分よりなる．

制御対象である倒立振子系$\boldsymbol{IP}$は次のような運動方程式 [松浦 91] に従う．

$$\dot{\theta} = \omega \tag{1}$$

$$\dot{x} = v \tag{2}$$

$$\begin{aligned}(M_c + M_p)\dot{v} + M_p L_p \dot{\omega}\cos\theta \\ = -C_c v + M_p L_p \omega^2 \sin\theta + F\end{aligned} \tag{3}$$

$$\begin{aligned}(I_p + M_p L_p^2)\dot{\omega} + M_p L_p \dot{v}\cos\theta \\ = -C_p \omega + M_p L_p g \sin\theta\end{aligned} \tag{4}$$

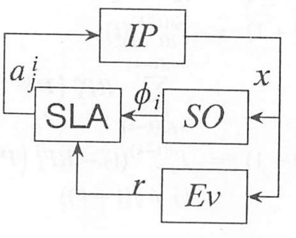

図1　システム構成

ここで，M_cは台車の質量，C_cは台車の粘性抵抗係数，M_pは棒の重心質量，L_pは重心と結合部との距離，I_pは重心まわり回転慣性モーメント，C_pは結合部まわりの粘性係数，gは重力加速度であり，xは台車の基準位置からの変位，vは台車の速度，θは直立状態を基準とした棒の傾斜角，ωは棒の角速度である．

状態観測関数$\boldsymbol{SO}$は制御対象の状態変数$\boldsymbol{x}$を観測し，それを，離散状態$\phi_i (i = 1, \ldots, n)$に変換しSLAに伝達する．

評価者$\boldsymbol{Ev}$は$\boldsymbol{x}$を観測し，それが，制御目的に近づいているかどうかを評価し，それによって，SLAに強化信号rを与える．

SLAは$\boldsymbol{SO}$よりϕ_iをうけとり，そこで取り得る行動$\{a_1^i, \ldots, a_{m_i}^i\}$のなかから，行動選択確率$P_j^i = \Pr(a_j^i|\phi_i)$に従い，確率的に行動を選択する．また，$\boldsymbol{Ev}$より$r$を受け取り，再強化学習則に従い$P_j^i$を変更する．

3　再強化学習則の拡張

離散時間マルコフ環境においては，SLA を複数組み合わせることで，適応最適制御が実現される．しかしながら，倒立振子のような動的な系は，現在の状態が過去に関係があり，マルコフ性を持たないと考えられる．そこで，このような系をSLAによって制御するために，再強化学習則の拡張を考えた．

この拡張は次のように行う．

離散時刻 t での行動と状態をそれぞれ $a(t) = a^{i(t)}_{j(t)}$, $\phi(t) = \phi_{i(t)}$とする．

$$
\begin{aligned}
1 \leq k \leq l \qquad & \text{について} \\
P^{i(t-k)}_{j(t-k)}(t+1) \;=\;& P^{i(t-k)}_{j(t-k)}(t) \\
&+ \sum_{\substack{h=1 \\ h \neq j(t-k)}}^{m_i(t-k)} RP^h_k\left(\boldsymbol{P}^{i(t-k)}(t), r\right) \quad (5) \\
P^{i(t-k)}_{j}(t+1) \;=\;& P^{i(t-k)}_{j}(t) - RP^j_k\left(\boldsymbol{P}^{i(t-k)}(t), r\right) \\
&(j \neq j(t-k)) \quad (6)
\end{aligned}
$$

ただし，$0 < RP^j_k(\boldsymbol{P},0) < 1, -1 < RP^j_k(\boldsymbol{P},1) < 0$ で，それぞれ，kについて単調減少，単調増加，とする．

4　計算機実験

計算機実験では状態数 $n = 49$, 行動数 $m_i = m = 7$として行った．また，$l = 1$ として，RPを次のように定義した．

$$
RP^j_1(\boldsymbol{P}^i, r) = \begin{cases} aP^i_j & (r=0) \\ -\frac{b}{m-1} + bP^i_j & (r=1) \end{cases} \quad (7)
$$

タイムステップ 0.01s で 60s を一試行として，100回学習を行い，その後学習をしない 10 試行について，$-0.2\pi < \theta < 0.2\pi$であった継続時間について，図 2, 図 3 のような結果が得られた．

このように，学習において賞罰の割合をかえることにより，倒立状態にある時間が変化することから，適当な関数 RPを選ぶことによって，倒立状態にある時間を延長することが期待できる．このことから，拡張した学習則が有効であると考えられる．

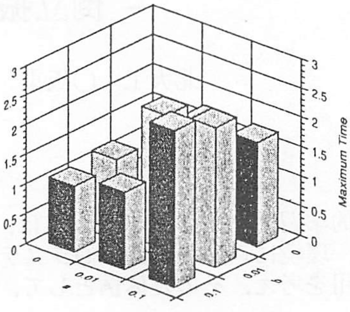

図 2　最大継続時間

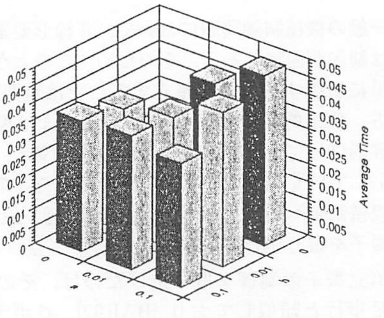

図 3　平均継続時間

5　おわりに

SLA の機械制御への応用を考え，その具体例として，倒立振子の振り上げ制御を試みた．

参考文献

[小池 91] 小池ら,(1991), “遺伝的アルゴリズムによる不安定系の一制御法”, システム合同シンポジウム．

[松浦 91] 松浦,(1991), “カート／ポール系のファジー制御”, 日本機械学会 (No.910-70)FAN シンポジウム講演論文集,pp283-288.

[BAR92] Borut,M.,(1992), “Dynamic versus Genetic versus Chaotic Programming”, *DYNAMIC,GENETIC,and CHAOTIC PROGRAMMING*,pp501-533.