社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

言語測度に基づいた最適スーパバイザの強化学習

[1] 谷口和隆† 　潮　　俊光† 　山崎　達志††

† 大阪大学 大学院基礎工学研究科　〒 560-8531 大阪府豊中市待兼山町 1-3
†† 関西学院大学 理工学部
E-mail: [1]taniguti@hopf.sys.es.osaka-u.ac.jp

あらまし　最近，形式言語に対して測度の概念が導入され，その測度に基づく最適スーパバイザの設計法が提案されている．しかしこの方法を適用するためには測度に関する情報が既知でなければならない．既知でない場合には，何らかの学習方法を導入する必要がある．一方，環境への適応性，柔軟性を持った制御手法を行うために強化学習が応用されている．強化学習では，環境から受け取る報酬をもとに学習者はよりよい行動政策を獲得できるように学習を行う．本報告では，言語測度に基づくスーパバイザの強化学習を提案する．特に，事象の生起を禁止したときのコストも考慮している．

キーワード　言語測度，スーパバイザ，強化学習，離散事象システム，Q 学習

A Reinforcement Learning of Optimal Supervisor Based on Language Measure

[1]Kazutaka TANIGUTI†, Toshimitsu USHIO†, and Tatsushi YAMASAKI††

† Graduate School of Engineering Science, Osaka University　Matikaneyama-tyou 1-3, Toyonaka-shi, Osaka, 560-8531 Japan
†† School of Science and Technology, Kwansei Gakuin University
E-mail: [1]taniguti@hopf.sys.es.osaka-u.ac.jp

Abstract　This paper proposes a synthesis method of an optimal supervisor in terms of a language measure by using a reinforcement learning. Recently, a concept of the language measure is introduced to the formal languages and a synthesis method of an optimal supervisor based on the language measure has been proposed. In this paper, we apply the reinforcement learning as a learning method of the language measure, and show that the optimal supervisor in terms of the language measure can be derived through learning. By computer simulation, we examine an opimality of the obtained supervisor.

Key words　language measure, supervisor, reinforcement learning, discrete event system, Q-learning

1. ま え が き

離散事象システムに対する論理的な制御法として，スーパバイザ制御がある．スーパバイザ制御では，システムの生成言語が最大になるという意味で最適な制御パターンを指定する．ところで，最近 Ray らによって言語測度という形式言語に対する符号付の測度の概念が導入された [1,3]．この言語測度を用いることによって，離散事象システムを定量的に評価することが出来る．その言語測度に基づく最適スーパバイザの設計法 [2] が提案されているが，本報告では強化学習における Bellman 方程式の状態価値関数が言語測度におけるパフォーマンス測度と一致することを示し，言語測度に基づいて最適となるスーパバイザを強化学習によって構成する方法を提案する．また，提案手法を食事をする哲学者の問題と n 本腕バンディット問題に適用することにより，最適なスーパバイザが獲得できていることを確かめる．

2. 言語測度の定義

離散事象システム G のモデルとして，ここではオートマトン表現を用いる．

$$G = (X,\ \Sigma,\ \delta,\ x_1,\ X_m)$$

ただし，Σ は事象の集合，X は状態の集合，$\delta : \Sigma \times X \to X$ は状態遷移関数，$x_1 \in X$ は初期状態，$X_m \subseteqq X$ は目標状態を表す．空事象 ϵ を含み，Σ の要素からなるすべての事象列の集合を Σ^* とおく．Σ^* の部分集合を言語という．

ここで，$|X| = n$, $|\Sigma| = m$, $\mathcal{I} = \{1,\ 2,\ \ldots,\ n\}$ はインデックス集合である．

また，G によって生成される言語を $L(G)$，G によって受理される言語を $L_m(G)$ とおく．

$$
\begin{aligned}
L(G,\ x) &= \{s \in \Sigma^* \mid \delta(x,\ s) \in X\} \\
L_m(G,\ x) &= \{s \in \Sigma^* \mid \delta(x,\ s) \in X_m\} \\
L(G) &= L(G,\ x_1) \\
L_m(G) &= L_m(G,\ x_1)
\end{aligned}
$$

X_m を次のように分割する [2].

$$
X_m = X_m^+ \cup X_m^-, \qquad X_m^+ \cap X_m^- = \emptyset
$$

ここで，

- X_m^+ は可到達が望ましい状態の集合．
- X_m^- は可到達が望ましくない状態の集合．

を表す．以下に記号の定義をおこなう．

$$
\begin{aligned}
L(x,\ p) &= \{s \in \Sigma^* \mid \delta(p,\ s) = x \in X\} \\
L(x) &= L(x,\ x_1) \\
L_m^+ &= \bigcup_{q \in X_m^+} L(x) \\
L_m^- &= \bigcup_{q \in X_m^-} L(x) \\
L^0 &= \bigcup_{x \in X - X_m} L(x)\ (= L(G) - L_m(G))
\end{aligned}
$$

とおくと，以下の性質が成り立つ．

$$
\begin{aligned}
L(G) &= \bigcup_{x \in X} L(x)\ (= \bigcup_{i \in \mathcal{I}} L(x_i)) \ = L^0 \cup L_m^+ \cup L_m^- \\
L_m(G) &= \bigcup_{x \in X_m} L(x) \ = \ L_m^+ \cup L_m^-
\end{aligned}
$$

$\mathcal{L}$ を $L(G)$ の σ 代数とおく．集合値関数 $\mu : \mathcal{L} \to \Re(= (-\infty,\ \infty))$ が以下の 2 条件を満たすとき，符号付実測度であるという．

(1) $\mu(\emptyset) = 0$

(2) $\mu(\bigcup_{j=1}^{\infty} K_j) = \sum_{j=1}^{\infty} \mu(K_j) \quad \forall K_j \in \mathcal{L}$ s. t. $K_i \cap K_j = \emptyset$ if $i \neq j$.

符号付実測度 $\mu : 2^{L(G)} \to \Re$ を以下のように構成する．

$$
\mu(L(x)) \begin{cases} = & 0 & \text{if } x \notin X_m \\ > & 0 & \text{if } x \in X_m^+ \\ < & 0 & \text{if } x \in X_m^- \end{cases}
$$

特性関数 $y : X \to \Re$ を以下のように定義する．

$$
y(x_i) = y_i \in \begin{cases} \{\ 0\ \} & \text{if } x_i \notin X_m \\ (\ 0,\ 1\ ] & \text{if } x_i \in X_m^+ \\ [\ -1,\ 0\ ) & \text{if } x_i \in X_m^- \end{cases} \qquad i \in \mathcal{I}
$$

状態重みベクトル $Y = [y_1,\ y_2,\ \ldots,\ y_n]^T$ を Y ベクトルという．

任意の $x_i \in X$, $\sigma_j \in \Sigma$, $s \in \Sigma^*$ に対して以下の 3 条件を満たす関数 $\tilde{\pi} : \Sigma^* \times X \to [\ 0,\ 1\ ]$ を G の事象コストという．

(1) $\tilde{\pi}[\sigma_j \,|\, x_i] = \tilde{\pi}_{ij} \in [\ 0,\ 1\ ) \qquad \sum_j \tilde{\pi}_{ij} < 1$ for any $i \in \mathcal{I}$

(2) $\tilde{\pi}[\sigma_j \,|\, x_i] = \tilde{\pi}_{ij} = \begin{cases} 0 & \text{if } \delta(x_i,\ \sigma_j) \text{ not defined.} \\ 1 & \text{if } \sigma_j = \epsilon \end{cases}$

(3) $\tilde{\pi}[\sigma_j s \,|\, x_i] = \tilde{\pi}[\sigma_j \,|\, x_i]\ \tilde{\pi}[s \mid \delta(x_i,\ \sigma_j)]$

事象のインデックスの集合 σ_i^k を以下のように定義する．

$$
\sigma_i^k = \{\, j \,|\, \delta(x_i,\ \sigma_j) = x_k \,\}
$$

G の状態遷移コスト $\pi : X \times X \to [0, 1)$ を以下のように定義する．

$$
\pi[x_k \mid x_i] = \pi_{ik} = \begin{cases} \displaystyle\sum_{j \in \sigma_i^k} \tilde{\pi}[\sigma_j \,|\, x_i] & \text{if } \sigma_i^k \neq \emptyset \\ 0 & \text{otherwise.} \end{cases}
$$

この π_{ik} を用いて，状態遷移コスト行列 Π（以下，Π 行列と呼ぶ.）を以下のようにおく．

$$
\Pi = \begin{bmatrix} \pi_{11} & \pi_{12} & \ldots & \pi_{1n} \\ \pi_{21} & \pi_{22} & \ldots & \pi_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{n1} & \pi_{n2} & \ldots & \pi_{nn} \end{bmatrix}
$$

明らかに，

$$
\sum_{k=1}^{n} \pi_{ik} < 1 \qquad \forall i \in \mathcal{I}
$$

となる．

事象列 $s \in L(x_k,\ x_i)$ に対して，

$$
\mu(\{\, s \,\}) = \tilde{\pi}[s \mid x_i]\ y(x_k)
$$

と定義すると，

$$
\mu(\{\, s \,\}) = \begin{cases} = & 0 & \text{if } x_k \notin X_m \\ > & 0 & \text{if } x_k \in X_m^+ \\ < & 0 & \text{if } x_k \in X_m^- \end{cases}
$$

である．そして，

$$
\mu(L(x_k,\ x_i)) = \sum_{s \in L(x_k,\ x_i)} \mu(\{\, s \,\}) \tag{1}
$$

$$
\begin{aligned}
\mu(L_m(G,\ x_i)) &= \sum_{x \in X_m} \mu(L(x,\ x_i)) \\
&= \sum_{x \in X} \mu(L(x,\ x_i)) \qquad (2)
\end{aligned}
$$

となる．

3. 言語測度の計算

以下，各状態 $x_i \in X$ に対して，

$$
L_i = L_m(G,\ x_i)
$$

とおく．このとき式が成立する [1].

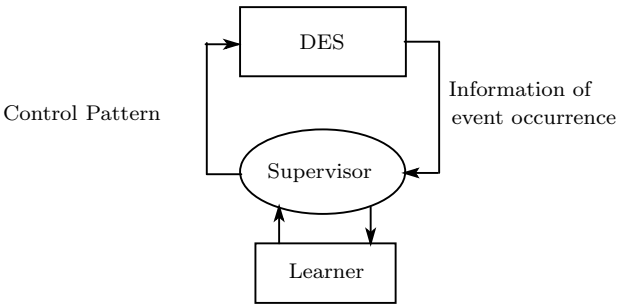

図 1　スーパバイザで制御される離散事象システム

$$\mu_i = \sum_{k=1}^{n} \pi_{ik}\mu_k + y_i \tag{3}$$

式 (3) を行列表現すると，

$$\mu = \Pi\mu + Y$$

となる．これより，

$$\mu = (I - \Pi)^{-1}Y \tag{4}$$

となる．

4. スーパバイザ制御

スーパバイザ制御では，制御対象となる離散事象システムに対し，制御仕様を満たすように，スーパバイザがシステムの可制御な事象の生起を許容または禁止する．

スーパバイザ学習の枠組みは，図 1 で表される．

制御の基本的な流れは，

（1） スーパバイザが，生起を禁止する事象の集合 (生起禁止パターン) を離散事象システムに提示する．

（2） 離散事象システムは生起禁止パターン以外から事象を選択し，新たな状態に遷移する．

（3） スーパバイザは，離散事象システムの生起事象を観測する．

というサイクルによって進められる．

状態 x_i において可制御事象 σ_j が生起して x_k に遷移することを禁止することにより生じるコストを c_{ij}^k とおく．x_k が文脈から明らかなときは c_{ij} と略記する．$n \times m$ 行列 $C = [\, c_{ij} \,]$ を生起禁止コスト行列という

スーパバイザ S によって状態 x_i で可制御事象 σ_j の生起を禁止するアクションを

$$d_{ij}^S = \begin{cases} 1 & \text{if } \sigma_j\text{の生起を禁止} \\ 0 & \text{otherwise} \end{cases}$$

$n \times m$ 行列 $D^S(x_i) = [\, d_{ij}^S \,]$ はスーパバイザ S によって可制御事象の生起を禁止するアクション行列である．スーパバイザ S によって制御されたシステムにおいて状態 x_i で生起が禁止された事象の番号の集合を d_i^S とおく．すなわち，

$$d_i^S = \{\, j \mid d_{ij} = 1 \,\}$$

とおく．制御されたシステムの状態遷移コスト行列 $\Pi^S = [\, \pi_{ik}^S \,]$ は以下のように計算される．

$$\pi_{ik}^S = \begin{cases} \sum_{j \in \sigma_i^k - d_i^S} \tilde{\pi}[\sigma_j \,|\, x_i] & \text{if } \sigma_i^k - d_i^S \neq \emptyset \\ 0 & \text{otherwise} \end{cases}$$

状態 x_i で事象の生起を禁止することによるスーパバイザ S の生起禁止コスト特性を

$$\xi_i^S = \sum_{j \in d_i^S} c_{ij}$$

とおく．スーパバイザ S の生起禁止コスト特性ベクトルを $\xi^S = [\, \xi_1^S,\ \xi_2^S,\ \ldots,\ \xi_n^S \,]^T$ と定義する．状態 $x_i \in X$ の修正特性を

$$y_i^S = y_i - \xi_i^S$$

と定義する．スーパバイザ S のもとでの修正特性ベクトルを

$$Y^S = [\, y_1^S,\ y_2^S,\ \ldots,\ y_n^S \,]^T = Y - \xi^S$$

と定義する．このとき，スーパバイザ S のパフォーマンス測度ベクトルは

$$\eta^S = [\, I - \Pi^S \,]^{-1} Y^S \tag{5}$$

となる．

5. システムの数理モデル

以下，制御対象である離散事象システム $G = (X, \Sigma, \delta, x_1, X_m)$ およびスーパバイザの全体をシステムと略記する．

ここでは，Bellman 方程式における価値関数が Ray らの提案する言語の符号付実測度に一致することを示す．

まず，以下の Bellman 方程式が成り立つ．

$$V^d(x_i) = \sum_{x_k \in X} P(x_i, d_i^S, x_k) \left[r(x_i, d_i^S, x_k) + \gamma V^d(x_k) \right] \tag{6}$$

各記号の意味は以下の通りである．

- $\sigma_i^k : \sigma_i^k = \{ j \mid \delta(x_i, \sigma_j) = x_k \}$ で定義される事象のインデックスの集合．
- $\tilde{d}_{ij}^S : d_{ij}^S = 1$ となる確率を表すパラメータ．$(\tilde{d}_{ij}^S \in [0, 1])$
- $P(x_i, d_i^S, x_k)$：状態 $x_i \in X$ でスーパバイザ S が生起禁止パターン $d_i^S \in D^S(x_i)$ を選択したときに，状態 $x_k \in X$ に遷移する確率．
- $V^d(x_i)$：状態 $x_i \in X$ での期待収益．
- $r(x_i, d_i^S, x_k)$：状態 $x_i \in X$ において，生起禁止パターン $d_i^S \in D^S(x_i)$ が選択されたときに $x_k \in X$ への遷移が起こった場合に受け取る報酬の期待値．
- γ：報酬の割引率．

ここで，制御対象はスーパバイザによって与えられた生起禁止パターン以外から生起事象を選択することから，

$$P(x_i.d_i^S, x_k) = \sum_{j \in \sigma_i^k - d_i^S} P_1(x_i, d_i^S, \sigma_j) P_2(x_i, \sigma_j, x_k)$$

が成り立つ．ただし，

- $P_1(x_i, d_i^S, \sigma_j)$：状態 $x_i \in X$ で，スーパバイザが生起禁止パターン $d_i^S \in D^S(x_i)$ を選択したとき，事象 $\sigma_j (j \in \sigma_j^k - d_i^S)$ が制御対象によって選択される確率．
- $P_2(x_i, \sigma_j, x_k)$：状態 $x_i \in X$ で，事象 $\sigma_j \in F(x_i)$ が生起したとき，状態 $x_k \in X$ に遷移する確率．

6. 仮　　定

提案手法において，制御対象に対して以下の仮定を設ける．

(1)　各状態 $x_i \in X$ について，G の事象の選ばれやすさを表すパラメータとして，事象コスト $\tilde{\pi}^*(x_i, \sigma_j)$ を導入する．

このとき，

$$P_1(x_i, d_i^S, \sigma_j) = \frac{\tilde{\pi}^*(x_i, \sigma_j)}{\sum_{l \notin d_i^S} \tilde{\pi}^*(x_i, \sigma_l) + \tilde{\pi}^*(x_i, \epsilon)} \tag{7}$$

とする．ここで，

$$\tilde{\pi}^*(x_i, \sigma_j) = \tilde{\pi}_{ij}^* \in [0, 1)$$
$$\sum_{\sigma_j \in F(x_i)} \tilde{\pi}^*(x_i, \sigma_j) + \tilde{\pi}^*(x_i, \epsilon) = 1$$

という関係が成り立っているとする．ただし，$F(x_i)$ は，状態 $x_i \in X$ において生起可能な事象の集合とする．

また，空事象を ϵ とする．空事象は，次の時間に何の事象も生起せずにその状態にとどまり，時間のみが 1 ステップ進むことを意味している．$\tilde{\pi}^*(x_i, \epsilon)$ が大きいと，その状態に長くとどまる可能性が高くなることを意味する．(各 $x_i \in X$ で $\tilde{\pi}^*(x_i, \epsilon) > 0$ と仮定する．)

さらに，報酬の割引率 γ が次のような構造を持つとする．

$$\gamma = \gamma(x_i, d_i^S) \tag{8}$$
$$= \sum_{l \notin d_i^S} \tilde{\pi}^*(x_i, \sigma_l) + \tilde{\pi}^*(x_i, \epsilon) \tag{9}$$

これより，$\gamma \in (0, 1]$ である．

また，G の制御された状態遷移コスト $\pi^S : X \times X \to [0, 1)$ を以下のように定義する．

$$\pi^S(x_k \mid x_i) = \pi_{ik}^S = \begin{cases} \sum_{j \in \sigma_i^k - d_i^S} P_1(x_i, d_i^S, \sigma_j)\gamma(x_i, \sigma_l) & \text{if} \quad \sigma_i^k - d_i^S \neq \emptyset \\ 0 & \text{otherwise} \end{cases}$$

ここで，(7),(9) 式より，次式が得られる．

$$\pi_{ik}^S = \begin{cases} \sum_{j \in \sigma_i^k - d_i^S} \tilde{\pi}^*(x_i, \sigma_j) & \text{if} \quad \sigma_i^k - d_i^S \neq \emptyset \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

(2)　(10) 式を用いて，状態遷移コスト行列 Π^S（以下，Π 行列と呼ぶ.）を以下のようにおく．

$$\Pi^S = \begin{bmatrix} \pi_{11}^S & \pi_{12}^S & \dots & \pi_{1n}^S \\ \pi_{21}^S & \pi_{22}^S & \dots & \pi_{2n}^S \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{n1}^S & \pi_{n2}^S & \dots & \pi_{nn}^S \end{bmatrix}$$

このとき，明らかに，

$$\sum_{k=1}^{n} \pi_{ik}^S < 1 \qquad \forall i \in \mathcal{I}$$

である．

(3)　報酬 $r(x_i, d_i^S, x_k)$ について，

$$r(x_i, d_i^S, x_k) = r_1(x_i, d_i^S) + r_2(x_i, \sigma_j, x_k)$$

という構造を持つとする．

ここで，r_1, r_2 の意味は以下の通りである．

- $r_1(x_i, d_i^S)$：状態 x_i で生起禁止パターン d_i^S を選択したときの報酬の期待値．
- $r_2(x_i, \sigma_j, x_k)$：状態 x_i で事象 σ_j が生起され，状態 x_k に遷移したときの報酬の期待値．

ここで，Asok Ray らの定義 [2] に従うと，事象の生起に伴う報酬を考慮していないので，$r_2(x_i, \sigma_j, x_k) = 0$ であると仮定する．

また，$r_1(x_i, d_i^S)$ は次のような構造を持つとする．

$$r_1(x_i, d_i^S) = y(x_i) - \xi(x_i, d_i^S)$$

ここで，$y(x_i)$ と$\xi(x_i, d_i^S)$ の意味は以下の通りである．

- $y(x_i)$：状態 $x_i \in X$ における特性関数 $y(x_i) : X \to \Re$ を以下のように定義する．

$$y(x_i) \in \begin{cases} \{\ 0\ \} & \text{if } x_i \notin X_m \\ (\ 0,\ 1\ ] & \text{if } x_i \in X_m^+ \\ [\ -1,\ 0\ ) & \text{if } x_i \in X_m^- \end{cases}$$

- $\xi(x_i, d_i^S)$：状態 x_i で生起禁止パターン d_i^S によって事象の生起を禁止することによるスーパバイザの生起禁止コスト特性．これは，次式のように表される．

$$\xi_i(x_i, d_i^S) = \sum_{j \in d_i^S} c_{ij}$$

- c_{ij}：状態 x_i において，可制御事象 σ_j が生起して遷移することを禁止するときに必要なコスト．

また，R を以下のように定義する．

$$R = \left[r_1(x_1, d_1^S), r_1(x_2, d_2^S), \dots, r_1(x_n, d_n^S)\right]^T \tag{11}$$

以上の仮定より，(6) 式は，

$$V^d(x_i) = r_1(x_i, d_i^S) + \sum_{x_k \in X} \pi_{ik}^S V^d(x_k) \tag{12}$$

となる．ここで，

$$V = \left[V^d(x_1), V^d(x_2), \dots, V^d(x_n)\right]^T \tag{13}$$

とすると，以下の式が成立する．

$$V = R + \Pi^S V$$

これより，

$$V = (I - \Pi^S)^{-1} R \tag{14}$$

という関係が得られる．

以上より，この価値関数 V は前章で定義したスーパバイザ S のパフォーマンス測度ベクトル (5) 式と一致することが示された．

7. スーパバイザ学習のアルゴリズム

ある時刻において状態が $x_i \in X$ であるとする．ここで，学習者たるスーパバイザは生起禁止パターン $d_i^S \in D^S(x_i)$ として，生起を禁止する事象の集合を制御対象に提示する．これにより，離散事象システムの状態に対して制御パターンを決定するという，状態フィードバック制御となる．

本報告では，パフォーマンス測度ベクトル V を最大にするような生起禁止パターンを求めるために，Q-learning を用いて学習させる．

Bellman 最適方程式は，

$$Q^*(x_i, d_i^S) = \sum_{x_k \in X} P(x_i, d_i^S, x_k) \left[r(x_i, d_i^S, x_k) + \gamma \max_{d_k^S \in D^S(x_k)} Q^*(x_k, d_k^S) \right]$$

である．

ここで，前節仮定と遷移が決定的であると仮定することにより，Bellman 最適方程式は次のように変形できる．

$$Q^*(x_i, d_i^S) = r_1(x_i, d_i^S) + \sum_{j \in \sigma_i^k - d_i^S} \tilde{\pi}(x_i, \sigma_j) \max_{d_k^S \in D^S(x_k)} Q^*(x_k, d_k^S) \tag{15}$$

また，制御対象側では，状態フィードバックによって生起が許容されている事象の中から (7) 式の確率で事象 σ_j $(j \in \sigma_i^k - d_i^S)$ が生起する．制御対象での事象の生起により，状態 x_i が x_k に遷移し，報酬 r_1 を獲得する．

ただし，空事象 ϵ が生じた場合は何の事象も生起せずにその状態にとどまり，報酬を受け取らずに時間のみが 1 ステップ進む．

(15) 式より，Q^* は $Q^*, r_1, \tilde{\pi}$ を用いて求めることができる．そこで，提案アルゴリズムでは，

$$r_1(x_i, d_i^S) \leftarrow r_1(x_i, d_i^S) + \beta[r - r_1(x_i, d_i^S)] \tag{16}$$

For all $l \notin D^S(x_i)$

$$\tilde{\pi}_{ij} \leftarrow \begin{cases} (1-\eta)\tilde{\pi}_{ij} & (\text{if} \quad \sigma_l \neq \sigma_j) \\ \tilde{\pi}_{ij} + \eta \left[\sum_{\sigma_m \notin D(x_i)} \tilde{\pi}_{im} - \tilde{\pi}_{il} \right] & (\text{if} \quad \sigma_l = \sigma_j) \end{cases} \tag{17}$$

として，$r_1, \tilde{\pi}$ を推定する．ここで，β, η は学習率である．これらを用いて，実際に選択した生起禁止パターン d_i^S で禁止されていない事象を含む全パターンに対して同時に Q 値の更新を行うことができる．すなわち，

$d_k^S \cap d_i^S \neq \emptyset$ を満たす，すべての $d_k^S \in D^S(x_i)$ に対して，

$$Q(x_i, d_k^S) \leftarrow r_1(x_i, d_k^S) + \sum_{j \in \sigma_i^k - d_i^S} \tilde{\pi}(x_i, \sigma_j) \max_{d_k^S \in D^S(x_k)} Q^*(x_k, d_k^S) \tag{18}$$

として，間接的に Q 値を推定する．

さらに，最大の Q 値を与える生起禁止パターン $d_k^S \in D^S(x_i)$ を用いて，$\tilde{d}_{ij}$ の値を次のように更新する．

$$\tilde{d}_{ij} \leftarrow \begin{cases} \tilde{d}_{ij} + \lambda(1 - \tilde{d}_{ij}) & (\text{if} \quad j \in d_k^S) \\ \tilde{d}_{ij} + \lambda(0 - \tilde{d}_{ij}) & (\text{if} \quad j \notin d_k^S) \end{cases} \tag{19}$$

ただし，λ は学習率である．

8. シミュレーション

8.1 食事をする哲学者の問題

文献 [1] で例題として取り上げられた食事をする哲学者の問題を考える．この問題の離散事象システムは，図 2 のオートマトンで表される．[1] では，状態重みベクトル Y は $Y = [1\ 0\ 0\ 0\ 0\ 0\ 0\ -0.5\ -0.5\ 1\ 1]^T$ となる．

この問題の制御目標は，以下の二つである．

1. 哲学者が状態 S_{10}, S_{11} に到達する可能性を増やす．
2. 哲学者が状態 S_8, S_9 に到達する可能性を減らす．

図 3 は縦軸が Q 値の最大値，横軸がエピソード数を表している．つまり，縦軸はスーパバイザのパフォーマンス測度 $\mu(L_m(S_1/G))$ を表している．[1] より，全ての条件が既知である場合の理論値は，$\mu(L_m(S_1/G)) = 1.7933$ である．図 3 から，この値に収束しているのが分かる．

図 4 はシミュレーションにより得られたスーパバイザで制御

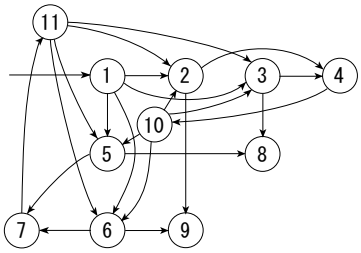

図 2　食事をする哲学者の問題

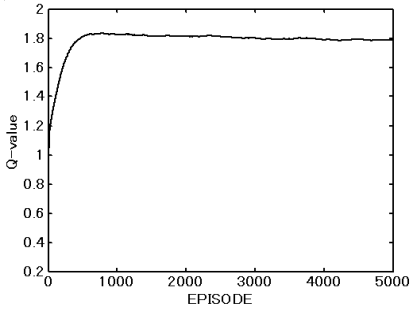

図 3　Q 値とエピソード数のグラフ

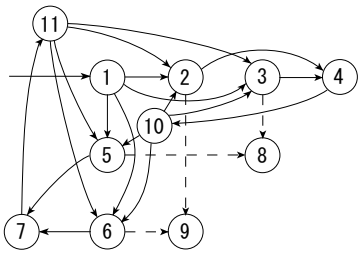

図 4　閉ループシステム

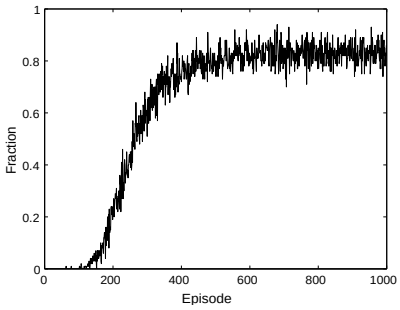

図 5　学習率の推移

されたときの閉ループシステムであり，点線は遷移が禁止されていることを表している．

図 4 より，スーパバイザは，状態 S_8，S_9 への遷移を禁止していることが分かる．これは，[1] で示されている全ての条件が既知である場合の閉ループシステムと一致しており，スーパバイザが正しい生起禁止パターンを提示していることが示された．

図 5 は，最適な生起禁止パターンをスーパバイザが学習した割合を示している．学習率はいずれも 0.01 に設定し，生起禁止パターンを学習する際に greedy な選択を行っている．このため学習率が 1 に収束していないが，すべての選択を Q 値に基づいて選択すれば学習率は 1 に収束することが確かめられる．

8.2　n 本腕バンディット問題

次に，n 本腕バンディット問題を考える．DES は n 種類の行動から 1 つを選択し報酬を受け取る．今回は全ての行動を可制御とし，$2^n - 1$ 個の生起禁止パターンの中から学習をする．この例題では 1 エピソードは一回の生起禁止パターンの選択で終了する． 図 6，7，8 はそれぞれ n=2, n=5, n=8 の場合の学習率の推移を表している．これらの生起禁止パターンの数は，それぞれ 3 個，32 個，255 個となっている．食事をする哲学者の

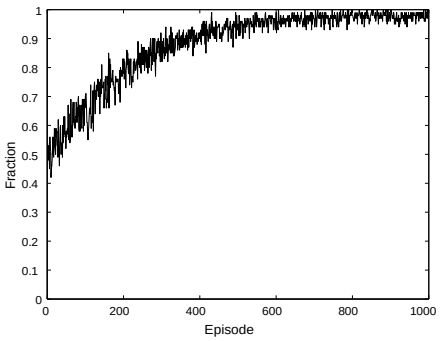

図 6　n=2 の学習率の推移

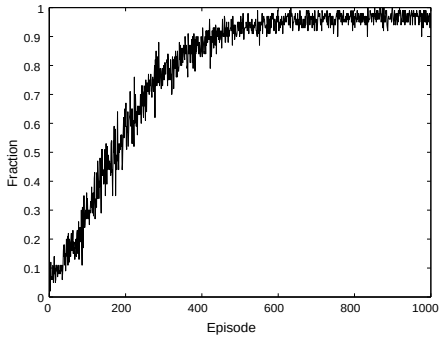

図 7　n=5 の学習率の推移

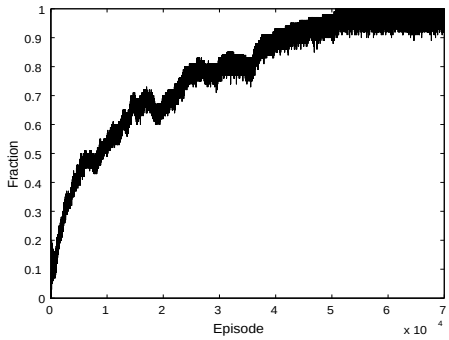

図 8　n=8 の学習率の推移

問題と同様に，学習率はいずれも 0.01 に設定し，生起禁止パターンを学習する際に greedy な選択を行っている．腕の本数が増えるにしたがって，最適な生起禁止パターンを学習するまでに必要なエピソード数が増えていることが分かる．

9.　あ と が き

Bellman 方程式における状態価値関数が，Ray らの提案する言語の符号付実測度に一致することを示した．さらに，言語測度に関して最適となるスーパバイザを獲得するための手法として，強化学習を用いた生起禁止パターンの学習法を提案した．食事をする哲学者の問題，n 本腕バンディット問題に対してこのアルゴリズムを適用し，言語測度に関して最適となるスーパバイザが得られることを確かめた．

文　　献

[1] Asok Ray, Xi Wang : “Signed Real Measure of Regular Languages”, American Control Conference, Anchorage, pp. 3937-3942 (2002)

[2] Asok Ray, Xi Wang : ”A language measure for performance evaluation of discrete-event supervisory control systems”, Applied Mathematical Modelling,(2004)

[3] Jimbo Fu, Asok Ray, Constantino M.Lagoa : “Unconstrained Optimal Control of Regular Languages”, Automatica, vol. 40, pp. 639-646 (2004)

[4] 山崎 達志, 潮 俊光 : “強化学習を用いた離散事象システムのスーパバイザ制御”, システム制御情報学会論文誌 vol. 47, no. 3, pp. 118-124 (2003)

[5] 谷口 和隆 : “言語測度を用いたスーパバイザの強化学習”, 大阪大学 特別研究報告 (2004)