「創発システム研究の現状そして今後の展開総合特集号」

論　文

言語測度に基づいた最適スーパバイザの強化学習*

谷口　和隆†・山崎　達志‡・潮　俊光†

Reinforcement Learning of Optimal Supervisor Based on Language Measure*

Kazutaka TANIGUCHI†, Tatsushi YAMASAKI‡ and Toshimitsu USHIO†

Recently, Wang and Ray introduced a signed measure for formal languages, called a language measure, to evaluate performance of strings generated by discrete event systems and a synthesis method of an optimal supervisor based on the language measure has been studied. In order to apply the method, exact information about a language measure of a controlled discrete event system is needed. From the practical point of view, it is not always possible to know the information *a priori*. In such a situation, a learning-based approach is useful to obtain an optimal supervisor with respect to the language measure.

This paper considers a synthesis method of an optimal supervisor with respect to a language measure. First, we clarify the relationship between the Bellman equation in reinforcement learning and performance of the language generated by the controlled discrete event systems. Next, using the relationship, we propose a learning method of the optimal supervisor where costs of disabling events are taken into consideration. Finally, by computer simulation, we illustrate an efficiency of the proposed method.

1. はじめに

離散事象システム [1] 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 [2,3]．スーパバイザ制御では，制御仕様を満足するという制約のもとに，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．これは，できる限りシステム本来の挙動を制限せずに多くの事象の生起を許容することが望ましいという考え方に基づいている．このとき，あらかじめ制御対象および制御仕様が，厳密に形式言語あるいはオートマトンで記述されている必要がある．しかし，多くの場合，要求される仕様を正確な形式言語で表現することは簡単ではない．また，スーパバイザ制御は論理的な制御の枠組みであるため，望ましくない状態に至らないように制御することはできるが，事象の生起や禁止のためのコストは考慮していない．これに関して，コストを考慮した最適スーパバイザ制御についての研究もなされている．Braveらは，望ましい状態にシステムを維持するためのコストに関しての最適スーパバイザ制御について研究している [8]．Kumarらは，事象の禁止コストと到達した状態に関する報酬に関する最適スーパバイザ制御について考察している [9]．Sengupta らは，事象の生起と禁止のコストを考慮し，最悪ケースでのコストを最小とする最適スーパバイザの設計法を提案している [11]．さらに，Marchandらは，部分観測の場合に拡張している [10]．

最近，Rayらによって言語測度とよばれる形式言語に対する符号付の実測度の概念が導入された [4,5]．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できる．既知の環境に対する言語測度に基づく最適スーパバイザの設計法が提案されており [6]，状態の価値が既知の場合に，ロボットの行動選択の学習を言語測度に基づき行う手法も提案されている [7]．

いっぽう，近年注目されている学習手法に強化学習が

* 原稿受付　2005年3月14日

† 大阪大学大学院　基礎工学研究科　Graduate School of Engineering Science, Osaka University; 1-3 Machikaneyama-cho, Toyonaka city, Osaka 560-8531, JAPAN

‡ 関西学院大学　理工学部　School of Science and Technology, Kwansei Gakuin University; 2-1 Gakuen, Sanda city, Hyogo 669-1337, JAPAN

Key Words: language measure, supervisory control, reinforcement learning, discrete event system.

あるが，多くの強化学習では，学習者の受け取る割引報酬の総和が最大となるという意味で最適な行動政策を，試行錯誤を通じて学習していく．

山崎らは，スーパバイザの設計に強化学習を用いて，最適な制御パターンを求めている [12–14]．しかし，最適性の扱いについて，スーパバイザ制御の基礎となる形式言語理論との関連が明らかではなかった．本論文では，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ベクトル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設計する方法を提案する．また提案手法では，従来の最適スーパバイザ制御と異なり，事象のコストが正確にわからない場合や，制御仕様の形式言語による厳密な記述ができない場合にも適用可能である．また，それらが変化する場合にも学習を続けることにより適応できる．提案手法を食事をする哲学者の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ることを示す．

以下，**2.** ではスーパバイザ制御と言語測度について述べる．**3.** ではスーパバイザ制御の数理モデルを示し，言語測度との関係を明らかにする．**4.** では本論文で提案するスーパバイザ学習のアルゴリズムを示す．**5.** では計算機実験により提案アルゴリズムの有効性を示す．最後に **6.** でまとめと今後の課題を述べる．

2. スーパバイザ制御と言語測度

本論文では事象は，可制御な事象と不可制御な事象にわけることができると仮定する．制御対象となる離散事象システムに対して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 [2,3]．スーパバイザ制御の枠組みは，Fig. 1 で表される．

制御の基本的な流れは，以下の通りである．

(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．

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 $G=(X, \Sigma, \delta, x_1, X_m)$ を用いる．ただし，X は状態の集合，Σ は事象の集合，$\delta: \Sigma \times X \to X$ は状態遷移関数，$x_1 \in X$ は初期状態，$X_m \subseteq X$ 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^* とおき，δ を $\delta: X \times \Sigma^* \to X$ に拡張する．Σ^c を可制御な事象の集合，Σ^{uc} を不可制御な事象の集合とすると，$\Sigma = \Sigma^c \cup \Sigma^{uc}$, $\Sigma^c \cap \Sigma^{uc} = \emptyset$ である．集合の要素数を $|\cdot|$ で表し，G では $|X|=n$, $|\Sigma|=m$ とする．$\mathcal{I}=\{1, 2, \ldots, n\}$ はインデックス集合である．状態 x_i から状態 x_k に遷移する事象のインデックス集合を $\sigma_i^k = \{j | \delta(x_i, \sigma_j) = x_k\}$ で定義する．状態 x_i で遷移が定義されている事象のインデックス集合を $\hat{\sigma}_i = \{j | \delta(x_i, \sigma_j) \text{ is defined}\}$ で定義する．また，状態 x_i から始まり，G によって生成される言語を，$L(G, x_i) = \{s \in \Sigma^* | \delta(x_i, s) \in X\}$ によって定義する．

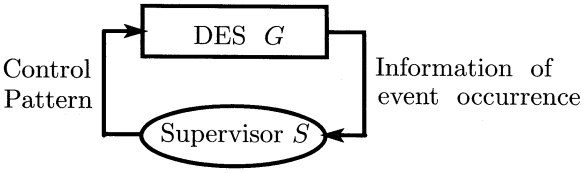

Fig. 1 Discrete event system controlled by the supervisor

Ramadge と Wonham による元々のスーパバイザ制御の枠組みは論理的なものであり，最大可制御部分言語の設計が重要な関心であった．本論文では，Ray らによって拡張されたスーパバイザ制御の枠組みを考える．ここでは事象や状態に定量的な指標を割り当て，それらを用いてシステムの生成言語に対し，符号付の実測度を与えることでスーパバイザを評価する [4,5]．X_m を可到達が望ましい状態の集合 X_m^+ と，可到達が望ましくない状態の集合 X_m^- に分割する．すなわち，$X_m = X_m^+ \cup X_m^-$, $X_m^+ \cap X_m^- = \emptyset$ である．状態に対する評価値として，任意の $i \in \mathcal{I}$ に対し特性関数 $y: X \to \Re$ を以下のように定義する．

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

$Y = [y_1, y_2, \ldots, y_n]^{\mathrm{T}}$ を特性ベクトルとよぶ．

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 の事象コストと定義する．

(1) $\tilde{\pi}[\sigma_j | x_i] = \tilde{\pi}_{ij} \in [\ 0,\ 1\ )$, $\forall i \in \mathcal{I}$, $\sum_j \tilde{\pi}_{ij} < 1$
(2) $\tilde{\pi}[\sigma_j | x_i] = 0$ if $\delta(x_i, \sigma_j)$ is undefined
(3) $\tilde{\pi}[\sigma_j s | x_i] = \tilde{\pi}[\sigma_j | x_i]\ \tilde{\pi}[s | \delta(x_i, \sigma_j)]$

スーパバイザ S によって状態 x_i で可制御事象 $\sigma_j \in \Sigma^c$ の生起を指定するアクションを次のようにおく．

$$d_{ij}^S = \begin{cases} 1 & \text{if } \sigma_j \text{の生起を禁止} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

スーパバイザ S によって制御されたシステム S/G の状態遷移コスト $\pi^S: X \times X \to [0,1)$ を以下のように定義する．

$$\pi^S[x_k | x_i] = \pi_{ik}^S = \begin{cases} \displaystyle\sum_{j \in \sigma_i^k - d_i^S} \tilde{\pi}[\sigma_j | x_i] & \text{if } \sigma_i^k - d_i^S \neq \emptyset \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

ただし，$d_i^S = \{j \mid d_{ij}^S = 1\}$ は状態 x_i で生起が禁止された事象のインデックスの集合を表し，生起禁止パターン

とよぶ．$D^S(x_i)$ により，状態 x_i の生起禁止パターンの集合を表す．この (i,k) 要素が π^S_{ik} となる行列を Π^S とおく．以下，この行列を Π^S 行列とよぶ．

状態 x_i において可制御事象 $\sigma_j \in \Sigma^c$ が生起して状態 x_k に遷移することを禁止することで生じるコストを c^k_{ij} 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よぶ．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

$$\xi^S_i = \xi_i(x_i,d^S_i) = \sum_{j \in d^S_i} c_{ij} \tag{4}$$

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 $\xi^S=[\xi^S_1, \xi^S_2, \ldots, \xi^S_n]^{\mathrm{T}}$ と定義する．生起禁止パターンの与え方によって，Y は変化し，スーパバイザ S のもとでの修正特性ベクトルを

$$Y^S=[y^S_1, y^S_2, \ldots, y^S_n]^{\mathrm{T}} = Y - \xi^S \tag{5}$$

と定義する．ただし，$y^S_i = y_i - \xi^S_i$ である．

このとき，制御されたシステム S/G の言語測度ベクトル μ^S は，

$$\mu^S=[\mu^S_1, \mu^S_2, \ldots, \mu^S_n]^{\mathrm{T}} = [I-\Pi^S]^{-1}Y^S \tag{6}$$

で与えられ，パフォーマンスベクトルとよばれる．ただし，μ^S_i は，スーパバイザ S によって制御されたときの $L(G,x_i)$ の言語測度であり，状態 x_i におけるパフォーマンスを表す．言語測度を用いることにより，スーパバイザについて，従来の最大可制御言語に基づく定性的な評価でなく，定量的な評価が可能となる．

3. 言語測度と Bellman 方程式

制御されたシステム S/G に対し，以下の Bellman 方程式が成り立つ．

$$V^d(x_i) = \sum_{x_k \in X} P(x_i,d^S_i,x_k) \times \left[r^*(x_i,d^S_i,x_k) + \gamma V^d(x_k)\right] \tag{7}$$

ここで，$P(x_i,d^S_i,x_k)$ は状態 $x_i \in X$ でスーパバイザ S が生起禁止パターン $d^S_i \in D^S(x_i)$ を選択したときに状態 $x_k \in X$ に遷移する確率，$V^d(x_i)$ は状態 x_i での期待収益(以後に獲得する割引報酬の総和の期待値)，$r^*(x_i,d^S_i,x_k)$ は状態 x_i において生起禁止パターン d^S_i が選択され，x_k への遷移が起こった場合に受け取る報酬の期待値，γ は報酬の割引率を表す．$V^d(x_i)$ は状態価値関数とよばれる．G は S 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，

$$P(x_i,d^S_i,x_k) = \sum_{j \notin d^S_i} P_1(x_i,d^S_i,\sigma_j) P_2(x_i,\sigma_j,x_k) \tag{8}$$

が成り立つ．ただし，$P_1(x_i,d^S_i,\sigma_j)$ は状態 x_i で，スーパバイザが生起禁止パターン d^S_i を選択したとき，事象 $\sigma_j (j \notin d^S_i)$ が制御対象 G によって選択される確率，$P_2(x_i,\sigma_j,x_k)$ は状態 x_i で事象 $\sigma_j (j \in \sigma^k_i)$ が生起したとき，状態 x_k に遷移する確率を表す．

提案手法においては，S/G に対して以下の仮定を設ける．

仮定 1： 各状態 $x_i \in X$ と事象 $\sigma \in \Sigma \cup \{\epsilon\}$ 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，$\tilde{\pi}^*(x_i,\sigma)$ を導入する．

このとき，

$$P_1(x_i,d^S_i,\sigma) = \frac{\tilde{\pi}^*(x_i,\sigma)}{\sum_{l \notin d^S_i} \tilde{\pi}^*(x_i,\sigma_l) + \tilde{\pi}^*(x_i,\epsilon)} \tag{9}$$

とする．ここで，

$$\tilde{\pi}^*(x_i,\sigma_j) \in [0,1),\ \tilde{\pi}^*(x_i,\epsilon) > 0 \tag{10}$$

$$\tilde{\pi}^*(x_i,\sigma_j) = 0 \quad \text{if } \delta(x_i,\sigma_j) \text{ is undefined} \tag{11}$$

$$\sum_{j \in \hat{\sigma}_i} \tilde{\pi}^*(x_i,\sigma_j) + \tilde{\pi}^*(x_i,\epsilon) = 1 \tag{12}$$

という関係が成り立っているとする．本論文では，空事象 ϵ は，どの事象も生起せずに現在の状態にとどまり，報酬を得ることなしに時間のみが 1 ステップ進むことを意味するとする．そのため，$\tilde{\pi}^*(x_i,\epsilon)$ が大きいと，その状態に長くとどまる可能性が高くなる．さらに，報酬の割引率 γ が次のような構造を持つとする．

$$\gamma = \gamma(x_i,d^S_i) = \sum_{l \notin d^S_i} \tilde{\pi}^*(x_i,\sigma_l) + \tilde{\pi}^*(x_i,\epsilon) \tag{13}$$

これより，$\gamma \in (0,1]$ である．

また，S/G に対して，$\pi^{*S}: X \times X \to [0,1)$ を以下のように定義する．

$$\pi^{*S}(x_k \mid x_i) = \pi^{*S}_{ik} = \begin{cases} \displaystyle\sum_{j \in \sigma^k_i - d^S_i} P_1(x_i,d^S_i,\sigma_j)\gamma(x_i,d^S_i) & \\ \qquad \text{if} \quad \sigma^k_i - d^S_i \neq \emptyset & \\ 0 \qquad \text{otherwise} & \end{cases} \tag{14}$$

ここで，(9), (13) 式より，次式が得られる．

$$\pi^{*S}_{ik} = \begin{cases} \displaystyle\sum_{j \in \sigma^k_i - d^S_i} \tilde{\pi}^*(x_i,\sigma_j) \ \text{ if } \ \sigma^k_i - d^S_i \neq \emptyset & \\ 0 \qquad \text{otherwise} & \end{cases} \tag{15}$$

(i,k) 要素が π^{*S}_{ik} となる行列を Π^{*S} とおき，以下，Π^{*S} 行列とよぶ．

仮定 2： 報酬 $r^*(x_i,d^S_i,x_k)$ について，

$$r^*(x_i,d^S_i,x_k) = r^*_1(x_i,d^S_i) + r^*_2(x_i,\sigma_j,x_k) \tag{16}$$

という構造を持つとする．ここで，$r_1^*(x_i,d_i^S)$ は状態 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，$r_2^*(x_i,\sigma_j,x_k)$ は状態 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を表す．

つぎに，Bellman方程式における状態価値関数がRayらの提案する言語測度に一致することを示す．

Rayらの定義[4]に従うとき，受け取る報酬は生起禁止パターンの与え方と現在の状態の評価値に依存し，事象の生起に伴う報酬は考慮しないので，$r_2^*(x_i,\sigma_j,x_k)=0$ とおける．

また，$r_1^*(x_i,d_i^S)$ は次のような構造を持つとする．

$$r_1^*(x_i,d_i^S)=y(x_i)-\xi(x_i,d_i^S) \tag{17}$$

上式より報酬 r_1^* は状態の特性関数と生起禁止コストの差より決まるため，修正特性ベクトルを表す(5)式の y_i^S に対応するものととらえることができる．また本論文では，学習においては y と ξ は陽には現れず，報酬としてまとめて扱われる．

ベクトル R を以下のように定義する．

$$R=\left[r_1^*(x_1,d_1^S),r_1^*(x_2,d_2^S),\ldots,r_1^*(x_n,d_n^S)\right]^{\mathrm{T}} \tag{18}$$

以上より，状態の遷移が決定的の場合，(7)式は，

$$V^d(x_i)=r_1^*(x_i,d_i^S)+\sum_{x_k\in X}\pi_{ik}^{*S}V^d(x_k) \tag{19}$$

となる．ここで，

$$V=\left[V^d(x_1),V^d(x_2),\ldots,V^d(x_n)\right]^{\mathrm{T}} \tag{20}$$

とすると，以下の式が成り立つ．

$$V=R+\Pi^{*S}V \tag{21}$$

これより，

$$V=(I-\Pi^{*S})^{-1}R \tag{22}$$

という関係が得られる．ここで $\Pi^{*S}=\Pi^S$, $R=Y$ であるから，Bellman方程式における状態価値関数 V^d のベクトル V は(6)式で定義したパフォーマンスベクトル μ^S と一致することが示された．

Rayらの提案する言語測度は，システムの生成言語に基づく定量的な性能指標であるが，これをBellman方程式における状態価値関数として解釈できることがわかる．そのため，直観的には言語測度は，事象の制御コストと状態の評価値に関する期待報酬を与えるものといえる．Bellman方程式の状態価値関数として言語測度を求めることができ，こうして得られた言語測度に基づく最適スーパバイザのもとでは，期待報酬に関して最適なシステムの振る舞いが得られる．

4. スーパバイザ学習のアルゴリズム

3. の結果より，Bellman方程式に基づいてパフォーマンスベクトルを計算する手法を提案する．提案手法では，言語測度における，状態の特性関数，事象コスト，生起禁止コストが未知の場合に，パフォーマンスベクトルを最大にする生起禁止パターンを求めるために，Q-learningに基づく更新式による学習を行う．これらの値は，実環境においては，学習者側からは未知であったり，ノイズが含まれていたりして，実際に観測することのみを通じて得られる場合がある．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 S 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 G 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．

(7)式に対応する，Q 値に関するBellman最適方程式は，

$$Q^*(x_i,d_i^S)=\sum_{x_k\in X}P(x_i,d_i^S,x_k)\times\left[r^*(x_i,d_i^S,x_k)+\gamma\max_{d_k^S\in D^S(x_k)}Q^*(x_k,d_k^S)\right] \tag{23}$$

となる[12]．ただし，$Q^*(x_i,d_i^S)$ は，状態 $x_i\in X$ で生起禁止パターン $d_i^S\in D^S(x_i)$ を選択し，以後は各状態で最大の Q 値を持つ生起禁止パターンを選択するときの期待収益である．ここで，仮定1, 2と状態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．

$$Q^*(x_i,d_i^S)=r_1^*(x_i,d_i^S)+\sum_{j\notin d_i^S}\tilde{\pi}^*(x_i,\sigma_j)V^*(\delta(x_i,\sigma_j)) \tag{24}$$

ただし，状態 $x_k=\delta(x_i,\sigma_j)\in X$ に対し，

$$V^*(x_k)=\max_{d_k^S\in D^S(x_k)}Q^*(x_k,d_k^S) \tag{25}$$

である．このとき，制御対象 G では，生起が禁止されていない事象の中から(9)式の確率で事象 $\sigma=\sigma_j(j\in\hat{\sigma}_i-d_i^S)$ が生起する．G での事象の生起により，状態 x_i が x_k に遷移し，報酬 r を獲得する．ただし，空事象 $\sigma=\epsilon$ 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ステップ進むとする．(24)式より，Q^* は r_1^*, $\tilde{\pi}^*$ を用いて間接的に求めることができる．そこで，本論文では，

$$r_1'(x_i,d_i^S)\leftarrow r_1'(x_i,d_i^S)+\alpha[r-r_1'(x_i,d_i^S)] \tag{26}$$
For all $\sigma'=\sigma_l(l\in\hat{\sigma}_i-d_i^S)$ and $\sigma'=\epsilon$

$$\tilde{\pi}'(x_i,\sigma') \leftarrow \begin{cases} (1-\beta)\tilde{\pi}'(x_i,\sigma') \\ \qquad \text{if} \quad \sigma' \neq \sigma \\ \tilde{\pi}'(x_i,\sigma') + \beta\Bigg[\displaystyle\sum_{m\in\hat{\sigma}_i - d_i^S} \tilde{\pi}'(x_i,\sigma_m) \\ \quad +\tilde{\pi}'(x_i,\epsilon) - \tilde{\pi}'(x_i,\sigma')\Bigg] \\ \qquad \text{if} \quad \sigma' = \sigma \end{cases} \tag{27}$$

として，r_1^*, $\tilde{\pi}^*$ を推定する．ここで，r_1', $\tilde{\pi}'$ はそれぞれ r_1^*, $\tilde{\pi}^*$ の推定値，α, β は学習率である．なお，空事象 ϵ が生起した場合は，報酬を受け取らないため，(26) 式による更新は行わない．r_1', $\tilde{\pi}'$ を用いて，実際に選択した生起禁止パターン d_i^S で許可された事象をすべて禁止している生起禁止パターン以外に対して同時に Q 値の更新を行うことができる．すなわち，$(\hat{\sigma}_i - d_i^{S'}) \cap (\hat{\sigma}_i - d_i^S) \neq \emptyset$ を満たす，すべての $d_i^{S'} \in D^S(x_i)$ に対して，

$$Q(x_i,d_i^{S'}) \leftarrow r_1'(x_i,d_i^{S'}) + \sum_{j\in\hat{\sigma}_i - d_i^{S'}} \tilde{\pi}'(x_i,\sigma_j) V'(\delta(x_i,\sigma_j)) \tag{28}$$

として，間接的に Q 値を推定する．ただし，状態 $x_k = \delta(x_i,\sigma_j) \in X$ に対し，

$$V'(x_k) = \max_{d_k^S \in D^S(x_k)} Q^*(x_k,d_k^S) \tag{29}$$

である．

状態 x_i で事象 σ_j を禁止する確率を $\tilde{d}_{ij} \in [0,1]$ で表す．$\tilde{d}_{ij}$ から生起禁止パターンを構成する．状態 x_i で最大の Q 値を与える生起禁止パターンを $\hat{d}_i^S$ とおく．

$$\hat{d}_i^S = \arg \max_{d_i^S \in D^S(x_i)} Q(x_i,d_i^S) \in D^S(x_i) \tag{30}$$

$\hat{d}_i^S$ を用いて $\tilde{d}_{ij}$ を以下のように更新する．

$$\tilde{d}_{ij} \leftarrow \begin{cases} \tilde{d}_{ij} + \lambda(1-\tilde{d}_{ij}) & \text{if} \quad j \in \hat{d}_i^S \\ \tilde{d}_{ij} + \lambda(0-\tilde{d}_{ij}) & \text{if} \quad j \notin \hat{d}_i^S \end{cases} \tag{31}$$

ただし，λ は学習率である．(31) 式は Q 値を最大にする生起禁止パターンに属する事象は禁止する確率を増やし，そうでない事象は減らすことで Q 値が最大な生起禁止パターンを学習する．また，生起禁止パターンを $\tilde{d}_{ij}$ に基づき決定することにより，最適と考えられる生起禁止パターンおよびその周辺をより高い確率で選択しつつ学習を進める．

以上より，スーパバイザの学習アルゴリズムは Fig. 2 のようになる．

5. シミュレーション

文献 [4] で例題として取り上げられた 2 人の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，Fig. 3 のオートマトンで表される．ここで，各哲学者を P_1, P_2，各フォークを F_1, F_2 とすると，各事象 $\sigma_1,\sigma_2,\ldots,\sigma_6$ の説明は Table 1 の通りである．ただし，$\sigma_1,\ldots,\sigma_4$ は可制御事象，σ_5,σ_6 は不可制御事象である．各状態の意味は，初期状態 1 が P_1, P_2 が思索，目標状態

(1) Initialize $r_1'(x_i,d_i^S)$ and $\tilde{\pi}'(x_i,\sigma)$ at each state.
(2) Calculate the initial Q value at each state by Eq. (24).
(3) Repeat (for each episode):
  (a) $x_i \leftarrow$ *initial state* x_1.
  (b) Repeat until x_i is a terminal state (for each step of episode) :
    i. Assign a control pattern $d_i^S \in D^S(x_i)$ based on $\tilde{d}_{ij}$.
    ii. Observe the occurrence of event σ and state transition $x_i \xrightarrow{\sigma} x_k$ in the DES G.
    iii. Acquire a reward r.
    iv. Update $r_1'(x_i,d_i^S)$ and $\tilde{\pi}'(x_i,\sigma)$ by Eqs. (26) and (27), respectively.
    v. Update the Q values for all $d_i^{S'} \in D^S(x_i)$ s.t. $(\hat{\sigma}_i - d_i^{S'}) \cap (\hat{\sigma}_i - d_i^S) \neq \emptyset$ by Eq. (28).
    vi. Calculate $\hat{d}_i^S$ by Eq. (30) and update the probability $\tilde{d}_{ij}$ by Eq. (31).
    vii. $x_i \leftarrow x_k$.

Fig. 2 Proposed algorithm

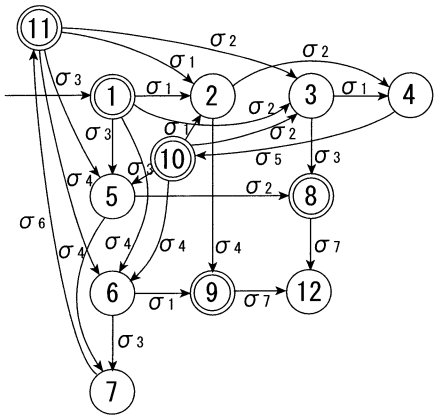

Fig. 3 Automaton of dining philosophers

Table 1 Event definition for the dining philosophers

Event	Description
σ_1	P_1 picks up F_1 from the table
σ_2	P_1 picks up F_2 from the table
σ_3	P_2 picks up F_1 from the table
σ_4	P_2 picks up F_2 from the table
σ_5	P_1 places F_1, F_2 on the table
σ_6	P_2 places F_1, F_2 on the table

10, 11はP_1またはP_2が食事後に思索，デッドロック状態の8, 9はP_1, P_2が1本ずつフォークを持つという状態を意味している．また，提案アルゴリズムでは，事象コストを遷移後に受け取るとしているためデッドロック状態では受け取ることができない．これを回避するためにシミュレーションの際に状態12というdummy stateを追加した．これは，不可制御事象σ_7により状態8, 9から遷移する．

特性ベクトルは$Y=[1\ 0\ 0\ 0\ 0\ 0\ 0\ -0.5\ -0.5\ 1\ 1]^{\mathrm{T}}$，また$\tilde{\pi}^*$は各状態$x_i$において$\tilde{\pi}^*(x_i,\epsilon)=0.04$とし，残りの事象はすべて等確率で生起するとした．ただし，σ_7のみ確率1で生起する．これらの情報はスーパバイザにとっては未知であり，学習を通じて獲得する．本例題では，生起禁止コストは考慮していないので，スーパバイザの受け取る報酬rは$y(x_i)$で決定される．各学習率の値は，$\alpha=0.7$, $\beta=0.01$, $\lambda=0.2$, $\tilde{d}_{ij}$の初期値はすべて0.5に設定した．各Q値の初期値は0とし，各$\tilde{\pi}'(x_i,\sigma)$は，事象$\sigma=\sigma_j(j\in\hat{\sigma}_i)$と空事象$\sigma=\epsilon$に対し，$1/(|\hat{\sigma}_i|+1)$で初期化した．本論文のシミュレーションでは，生起禁止パターンを選択する際に$\epsilon=0.1$としたϵ-greedy選択を行っている．これは，確率ϵでランダムに選択した生起禁止パターンを用い，残りの確率で$\tilde{d}_{ij}$に基づいた生起禁止パターンの決定を行う．これにより，様々な生起禁止パターンを探す余地をもったアルゴリズムとなっている．学習は，エピソードを繰り返し経験することにより進められる．1エピソードは，初期状態から始まり，20ステップ経過するかデッドロック状態になることで終了とする．

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10, 11に到達する可能性を増やす．

(2) 哲学者が状態8, 9に到達する可能性を減らす．

Fig. 4は，学習により得られたスーパバイザで制御された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．Fig. 4より，スーパバイザは，状態8, 9への遷移を禁止していることがわかる．これは，文献[4]で示されているすべ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，スーパバイザが最適な生起禁止パターンを学習していることを示している．

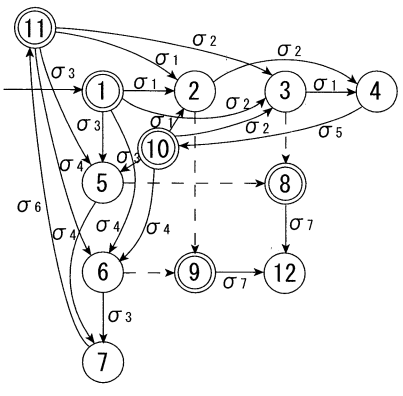

Fig. 4 Controlled plant model

つぎに，100回のシミュレーションを行ったとき，各状態でスーパバイザが学習した最大のQ値を持つ生起禁止パターンが，Fig. 4で表されている最適な生起禁止パターンとすべて合致した割合をFig. 5に示す．エピソードが進むにしたがって，最適な生起禁止パターンを学習していることがわかる．

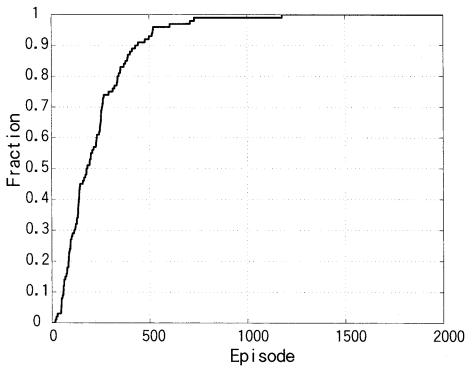

Fig. 5 Relation between the number of episodes and the fraction what the supervisor selects the optimal control pattern

文献[4]より，すべての条件が既知である場合の初期状態におけるパフォーマンスベクトルの理論値について，$\mu_1^S=1.7933$であることが示されている．シミュレーションでは，約8000ステップ後に状態1のQ値がこの値に収束していた．

6. おわりに

スーパバイザ制御の最適制御問題に着目し，Bellman方程式における状態価値関数が，Rayらの提案する言語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なスーパバイザが得られることを確かめた．

言語測度に関する最適スーパバイザがマルコフ決定過程環境におけるBellman方程式の状態価値関数を求めることにより設計できることを明らかにした．これにより，通常の強化学習のアルゴリズムで用いられるさまざまなテクニックによる性能向上を図ることも可能になったと考える．

今後の課題として，事象の数が増えると，生起禁止パターンの数が急激に増加するので，より効率の良いアルゴリズムの検討が必要となる．また，事象の生起が部分観測となる場合への本手法の拡張も検討課題である．

参考文献

[1] C. G. Cassandras and S. Lafortune: *Introduction to Discrete Event Systems*, Kluwer Academic Pub.

(1999)

[2] P. J. Ramadge and W. M. Wonham: Supervisory control of a class of discrete-event processes; *SIAM J. Control Optim.*, Vol. 25, No. 1, pp. 206–230 (1987)

[3] W. M. Wonham and P. J. Ramadge: On the supremal controllable sublanguage of a given language; *SIAM J. Control Optim.*, Vol. 25, No. 3, pp. 637–659 (1987)

[4] X. Wang and A. Ray: Signed real measure of regular languages; *American Control Conference*, Anchorage, pp. 3937–3942 (2002)

[5] X. Wang and A. Ray: A language measure for performance evaluation of discrete-event supervisory control systems; *Applied Mathematical Modelling*, Vol. 28, No. 9, pp. 817–833 (2004)

[6] A. Ray, J. Fu and C. Lagoa: Optimal supervisory control of finite state automata; *Int. J. Control*, Vol. 77, No. 12, pp. 1083–1100 (2004)

[7] X. Wang, J. Fu, P. Lee and A. Ray: Robot behavioral selection using discrete event language measure; *American Control Conference*, pp. 5126–5131 (2004)

[8] Y. Brave and M. Heymann: On optimal attraction of discrete-event processes; *Inform. Sciences*, Vol. 67, pp. 245–276 (1993)

[9] R. Kumar and V. K. Garg: Optimal supervisory control of discrete event dynamical systems; *SIAM J. Control Optim.*, Vol. 33, No. 2, pp. 419–439 (1995)

[10] H. Marchand, O. Boivineau and S. Lafortune: Optimal control of discrete event systems under partial observation; *Proc. of the 40th IEEE Conference on Decision and Control*, pp. 2335–2340 (2001)

[11] R. Sengupta and S. Lafortune: An optimal control theory for discrete event systems; *SIAM J. Control Optim.*, Vol. 36, No. 2, pp. 488–541 (1998)

[12] 山崎, 潮: 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. 16, No. 3, pp. 118–124 (2003)

[13] T. Yamasaki and T. Ushio: Supervisory control of partially observed discrete event systems based on a reinforcement learning; *Proc. of International Conference on Systems, Man, and Cybernetics*, pp. 2956–2961 (2003)

[14] T. Yamasaki and T. Ushio: Decentralized supervisory control of discrete event systems based on reinforcement learning; *Proc. of 10th IFAC/IFORS/IMACS/IFIP Symposium on Large Scale Systems: Theory and Application*, pp. 379–384 (2004)

[15] R. S. Suton and A. G. Barto: *Reinforcement Learning*, MIT Press (1998)