言語測度を用いたスーパバイザの強化学習

A Reinforcement Learning of Supervisor Based on Language Measure

大阪大学 大学院基礎工学研究科　○ 谷口　和隆，潮　俊光
関西学院大学 理工学部　山崎　達志
○ Kazutaka Taniguchi and Toshimitsu Ushio
Graduate School of Engineering Science, Osaka University
Tatsushi Yamasaki
School of Science and Technology, Kwansei Gakuin University

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.

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 らによって形式言語に対する符号付の測度の概念が導入された [1,2]．その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，本報告では，この言語測度を用いて最適な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．

2　強化学習によるスーパバイザ設計

離散事象システム G のモデルは多く提案されているが，ここでは，オートマトン $G = (X,\ \Sigma,\ \delta,\ x_1,\ X_m)$ でモデル化する．

スーパバイザ学習の枠組みは，図 1 で表される．

学習の基本的な流れは，以下のようなサイクルで進められる．

1. スーパバイザが，DES の生起事象，状態の遷移と獲得した報酬を観測．

2. 学習器によるスーパバイザの学習．

i. 観測した情報より，DES が獲得する報酬と事象選択の確率の推定値を更新．

ii. 更新した推定値を用いて Q 値を更新．

iii. 最大の Q 値を与える生起禁止パターンを用いて，生起禁止確率を更新．

3. 更新した生起禁止確率より，スーパバイザが DES に生起禁止パターンを提示．

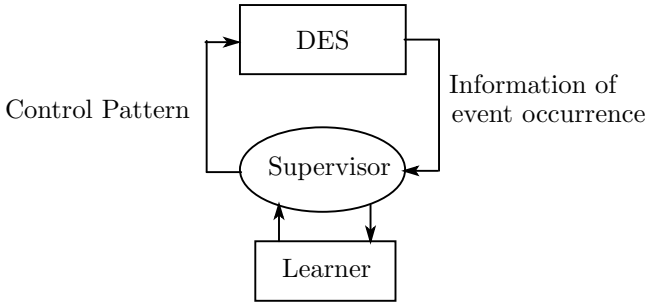

図 1: スーパバイザで制御される離散事象システム

4. DES は可制御事象のうち，生起禁止されていない生起事象を選択，新たな状態に遷移する．

スーパバイザの性能評価を言語測度を用いて行うとき，強化学習によりスーパバイザを求める方法を提案する．強化学習の基となる Bellman 方程式は次式となる．

$$V^d(x_i) = r_1(x_i, d_i^S) + \sum_{x_k \in X} \pi_{ik}^S V^d(x_k) \qquad (1)$$

但し，

- $V^d : X \to \Re$ は生起禁止パターン d に対する価値関数．
- $V = \left[V^d(x_1), V^d(x_2), \ldots, V^d(x_n)\right]^T$
- $r_1(x_i, d_i^S)$ は状態 x_i で生起禁止パターン d_i^S を選択したことによる報酬．
- π_{ik}^S は状態遷移 $x_i \to x_k$ に対する制御された状態遷移コストである．ここで，$\pi^S : X \times X \to [0, 1)$．
- Π^S は制御されたシステムの状態遷移コスト行列．$\Pi^S = [\pi_{ik}^S]$
- $R = \left[r_1(x_1, d_1^S), r_1(x_2, d_2^S), \ldots, r_1(x_n, d_n^S)\right]^T$

とおくと，(1) 式より以下の関係が得られる．

$$V = (I - \Pi^S)^{-1} R \tag{2}$$

この V は言語の符号付実測度の定義におけるスーパバイザのパフォーマンス測度ベクトルとなっている．したがって，(1) 式より，スーパバイザのパフォーマンスを求めることができる．

Bellman 最適方程式は，次のように変形できる [3]．

$$Q^*(x_i, d_i^S) = r_1(x_i, d_i^S) + \sum_{j \notin d_i^S} \tilde{\pi}(x_i, \sigma_j) \max_{d_k^S \in D(x_k)} Q^*(x_k, d_k^S) \tag{3}$$

(3) 式より，Q 値は $r_1, \tilde{\pi}$ 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 Q-learning の更新式に基づいて $r_1, \tilde{\pi}$ を推定する．これらの推定値を用いて Q 値の更新を行うことができる．

さらに，最大の Q 値を与えるスーパバイザの生起禁止パターンを用いて，スーパバイザが状態 x_i において，可制御事象 σ_j の生起を禁止する確率 $\tilde{d}_{ij}$ を更新する [4]．

3　シミュレーション

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 $Y = [1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ \ -0.5\ \ -0.5\ 1\ 1]^T$ 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 はシミュレーション結果の Q 値とエピソード数との関係を表したグラフである．

[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ことが分かる．

次に，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．

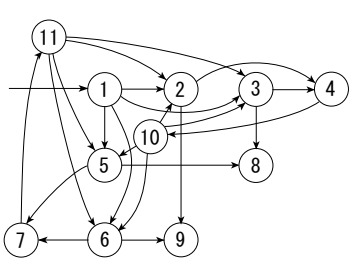

図 2: 食事をする哲学者の問題

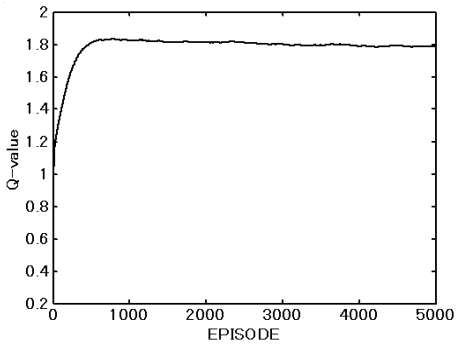

図 3: Q 値とエピソード数のグラフ

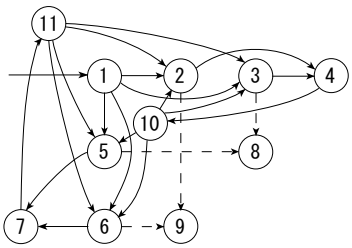

図 4: 閉ループシステム

図 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．

4　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学習するアルゴリズムを提案し，食事をする哲学者の問題に適用した．

参考文献

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

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

[3] 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)

[4] 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)