岡山理科大学紀要 第45号 A pp. 71-77(2009)

# モジュール型強化学習で複数センサの二次相関を学習するための検討

中間 隼人・山田 訓*

岡山理科大学大学院工学研究科修士課程知能機械工学専攻

* 岡山理科大学大学工学部知能機械工学科

(2009年9月8日受付、2009年11月5日受理)

## 1. はじめに

強化学習 [1] は，エージェントが環境との試行錯誤により行動の目標に応じた報酬の総和を最大にするような行動則を獲得するための枠組みである．制御結果に対する評価だけを用いて学習し，制御対象に対する事前知識を必要としないため，幅広い制御対象に適用できる可能性がある．しかし，強化学習は基本的に試行錯誤で学習を行うため，制御が複雑になると学習効率が低下するという問題と，入力次元が増えると学習に時間がかかるという問題がある．

このような問題の解決法として，モジュール型強化学習 [2] が検討されている．モジュール型強化学習は，単純な制御を学習する制御モジュールと，制御モジュールを状況に合わせて選択するよう学習する選択モジュールで構成される．モジュール型強化学習を用いることで，従来の強化学習のままでは困難な制御の学習が可能になることが示されている [2][3]．これまでは，単一のセンサ入力から適切な制御を決定できる課題が学習されていた．しかし，さらに複雑な制御になると，複数のセンサ入力の組み合わせから適切な制御を決定することが必要になると考えられる．

本研究では，3種類のセンサを持つロボット制御にモジュール型強化学習を適用し，このうち2種類のセンサ入力の組み合わせ（センサのAND条件）から適切な制御を決定する必要のある課題を学習させた．その結果モジュール型強化学習のみでは効率が悪く学習できなかった．学習効率を改善するため，この課題に対して経験強化と初期学習を導入した．経験強化は成功時の経験を繰り返し強化する方法で，初期学習は移動制御の学習の前に成功の状態を学習する方法である．経験強化，初期学習を組合わせる事で学習効率を改善することができたので，結果を報告する．

## 2. モジュール型強化学習

### 2.1 学習システム

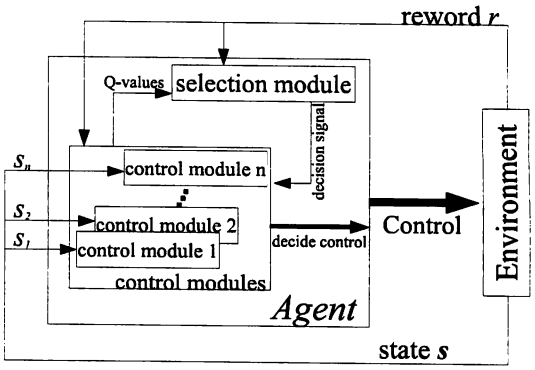


図1 モジュール型強化学習システムの構成

モジュール型強化学習は，図1に示すように複数の制御モジュールと選択モジュールから構成されるシステムである．与えられた制御課題から，制御できるよう状態空間を分割しモジュールに割り当てる．各制御モジュールは，状態空間の一部を入力とし，各状態に対する適切な制御を学習する．選択モジュールは，各制御モジュールの制御結果の予測値（行動価値関数）を入力とし，状態に対する適切な制御モジュールを選択するよう学習する．選択モジュールにより決定された制御モジュールの，決定した制御がシステム全体の制御出力となる．制御の結果に対する報酬はすべてのモジュールに与えられる．

### 2.2 学習アルゴリズム

各モジュールはSarsa($\lambda$)[1] で学習し，各モジュールの状態表現に関数近似の1つであるタイルコーディング [1] を用いる．関数近似を用いるとき，行動価値関数は $\vec{w}$ をパラメータベクトルとするパラメータ関数として表される．

制御モジュール $m$ の状態 $s^m$，行動 $a$ に対する行

動価値関数 $Q_m(s^m,a)$ 及び，選択モジュールの状態 $s^s$，モジュール $m$ に対する行動価値関数 $Q_s(s^s,m)$ は次式で計算される．

$$Q_m(s^m,a)=\vec{w}_m^T\vec{\phi}_{s^m a}=\sum_{i=1}^{n_m}w_m(i)\phi_{s^m a}(i)$$

$$Q_s(s^s,m)=\vec{w}_s^T\vec{\phi}_{s^s m}=\sum_{i=1}^{n_s}w_s(i)\phi_{s^s m}(i) \quad (1)$$

ここで，$\vec{w}_m$ は制御モジュール $m$ のパラメータベクトル，$\vec{w}_s$ は選択モジュールのパラメータベクトル，$\vec{\phi}_{s^m a}$ は制御モジュール $m$ の行動 $a$ に対する状態を表す特徴列ベクトル，$\vec{\phi}_{s^s m}$ は選択モジュールのモジュール $m$ に対する状態を表す特徴列ベクトルである．$n_m$，$n_s$ はそれぞれ制御モジュール $m$ の状態・行動の要素数，選択モジュールの要素数である．

選択モジュールは $Q_s(s^s,m)$ に基づき，制御モジュールを決定する．決定された制御モジュールは $Q_m(s^m,a)$ に基づき，行動を決定する．モジュール，行動は，$Q$ 値が等しい場合はその中からランダムに選択し，それ以外は最大の Q 値を与えるモジュールや行動を選択する．

制御モジュールのパラメータの更新は，次式で行う．

$$\hat{r}_t^m=\begin{cases}0 & m_t\neq m\\ r_{t+1}+\gamma_m Q_m(s_{t+1}^m,a_{t+1}) & \\ -Q_m(s_t^m,a_t) & m_t=m,m_{t+1}=m\\ r_{t+1}+\gamma_m Q_s(s_{t+1}^s,m_{t+1}) & \\ -Q_m(s_t^m,a_t) & m_t=m,m_{t+1}\neq m\end{cases}$$

$$w_m(i)=w_m(i)+\frac{\alpha_m\hat{r}_t^m e_m(i)}{tile_m}$$
$$(i=1,2,\ldots,n_m) \quad (2)$$

ここで，$m_t$ は時刻 $t$ で選択された制御モジュール，$a_t$ は時刻 $t$ で選択された制御モジュールの出力，$r(t)$ は時刻 $t$ での報酬である．$\gamma_m$，$\alpha_m$ はそれぞれ制御モジュールに対する減衰率，学習率，$tile_m$ はタイリング数である．$\vec{e}_m$ は制御モジュール $m$ の適格度トレースである．

選択モジュールのパラメータの更新は次式で計算する．

$$\hat{r}_t^s=r_{t+1}+\gamma_s Q_s(s_{t+1}^s,m_{t+1})-Q_s(s_t^s,m_t)-\hat{r}_t^{m_t}$$

$$w_s(i)=w_s(i)+\frac{\alpha_s\hat{r}_t^s e_s(i)}{tile_s}$$
$$(i=1,2,\ldots,n_s) \quad (3)$$

ここで，$\gamma_s$，$\alpha_s$ はそれぞれ選択モジュールに対する減衰率，学習率，$tile_s$ はタイリング数である．$\vec{e}_s$ は選択モジュールの適格度トレースである．$\hat{r}_t^{m_t}$ は，選択された制御モジュールで計算された TD 誤差である．選択モジュールで計算される TD 誤差には，選択された行動に起因する誤差と，モジュール選択に起因する誤差が含まれると考えられる．前者の誤差を選択モジュールの TD 誤差の計算に考慮するため，この項を加える．これにより，制御モジュールがある程度学習されてから選択モジュールを学習するようになる．

適格度トレースは，入れ替え更新トレース [1] を用いる．次式により計算する．

$$e_m(i)=\begin{cases}1 & a=a_t,\vec{\phi}_{s^m a}=\vec{\phi}_{s_t^m a}\\ 0 & a\neq a_t,\vec{\phi}_{s^m a}=\vec{\phi}_{s_t^m a}\\ \gamma_m\lambda_m e_m(i) & otherwise\end{cases}$$

$$(i=1,2,\ldots,n_m)$$

$$e_s(i)=\begin{cases}1 & m=m_t,\vec{\phi}_{s^s m}=\vec{\phi}_{s_t^s m}\\ 0 & m\neq m_t,\vec{\phi}_{s^s m}=\vec{\phi}_{s_t^s m}\\ \gamma_s\lambda_s e_s(i) & otherwise\end{cases}$$

$$(i=1,2,\ldots,n_s) \quad (4)$$

ここで，$\lambda_m$ は制御モジュール $m$ に対するトレース減衰率，$\lambda_s$ は選択モジュールに対するトレース減衰率である．

## 3. 課題設定

### 3.1 制御対象

本研究では，e-puck を使用する．e-puck は Khepera ロボットと同系統の小型ロボットであり，8 個の赤外線センサとカメラを持つ．Khepera simulator を基本とし，これら各センサの実測値を元に修正したモデルを用いて，次節で述べる課題の制御学習を行う．

ロボットの行動は，左旋回，前進，右旋回の 3 つとする．ロボットが行動を実行し，次の状態を観測するまでを 1 ステップとし，成功・失敗までを 1 エピソードとする．

### 3.2 制御課題

フィールド（1000 × 1000mm）に，障害物となる壁を配置する．また，ターゲットとなるオブジェク

トと，ダミーオブジェクトを配置する（図 2(b)）．この環境において，障害物やダミーオブジェクトを回避しながら，ターゲットに近づくことを課題とする．ターゲットは，ランプが点いている正しい絵を貼り付けたオブジェクトとする．ダミーオブジェクトは，異なる絵を貼り付けたオブジェクト (dummy1)，正しい絵を貼り付けたランプが点いていないオブジェクト (dummy2)，ランプのみ (dummy3) の 3 種類とする．障害物は距離センサ，ランプは光センサ，絵の識別はカメラを使用する．この課題は，光センサ値とカメラ入力の組み合わせから制御を選択する必要がある．

ターゲットに近づいた場合は制御成功とし，+1 の報酬を与える．ダミーに近づいた場合は制御失敗とし −1 の報酬を与える．また，壁にぶつかる，その場で回転する，一定 Step 以内に成功しない場合も −1 の報酬を与える．

### 3.3 モジュール構成

モジュール型強化学習システムのモジュール構成は，障害モジュール，光モジュール，カメラモジュール，選択モジュールとする．障害モジュールは，赤外線センサ距離センサモードの計測値，計 4 個を入力とする．光モジュールは，赤外線センサ光センサモードの計測値，計 4 個を入力とする．カメラモジュールは，20×20 ピクセル，グレースケールモードで取得した画像に対して，位置や大きさなどを調整した入力画像ベクトルと，テンプレート画像ベクトルの相関係数（2 個），黒いピクセルの中央のピクセル番号，幅と高さの計 5 個を入力とする．選択モジュールは，各制御モジュールの現在の状態で計算される行動価値関数（Q 値）の最大値，計 3 個を入力とする．

各モジュールのタイルコーディングにおけるタイルリング数を 7 とし，各タイリングのオフセットは，$offset$ = 入力最大値/(タイリング数 × 分割数) とする．障害モジュールと光モジュールは各センサの最大値 (2048) を 6 分割する．カメラモジュールは，相関係数 (0 から 1) を 6 分割，各ピクセル値は最大値 20 を 6 分割する．選択モジュールは-1 から 1 の間を 10 分割する．

## 4. 学習効率の改善法

学習効率を改善するための方法として，初期学習 (Initial Learning : InL) と経験強化 (Iterative Learning : ItL) を検討する．

### 4.1 初期学習

初期学習は，移動制御を学習する前に，成功しやすい環境（ターゲットの近く）で成功の状態を学習するという方法である．

本研究では，初期学習の環境を次のように設定した．ターゲットの中心から x 方向に +150mm，y 方向 ±20mm の領域内で，ロボットの向きがターゲットに対面で水平の場合を 0° とすると，±30° の間であらかじめランダムに 20 個の座標を設定する．この設定した座標からランダムに選択した位置を開始点とし初期学習を行った．

初期学習の過程は，始めに光モジュールと選択モジュールのみで学習を行う．一定回数成功すると（本研究では 20 回），カメラモジュールと選択モジュールのみで学習する．両方の成功回数が一定回数以上 (150 回) で初期学習を終了し，移動制御の学習を開始する．

### 4.2 経験強化

経験強化は，成功したエピソードに含まれる，各モジュールの状態・行動を繰り返し強化することで，学習効率を向上させようとする方法である．経験強化では，制御成功の場合に，そのエピソードの成功時点から遡り，一定ステップ分 (500) の状態・行動を保存する．そして，成功したエピソードの後，過去数回分 (5 回) の成功エピソードを強化する．これにより，1 回の成功で複数のエピソードを強化でき，同じエピソードを複数回強化することが出来る．ここで強化とは，各モジュールの各状態・行動に対するパラメータ関数に強化の方法によって計算した値を上乗せすることを意味している．

本研究では強化の方法に Profit Sharing (PS) [4] を使用した．PS によるパラメータの更新は次式を用いる．

$$\vec{w}_m = \vec{w}_m + \alpha_m f_m(t, r_T, T)\vec{\phi}_{s_t^m a}$$

$$\vec{w}_s = \vec{w}_s + \alpha_s f_s(t, r_T, T)\vec{\phi}_{s_t^s m} \qquad (5)$$

ここで，$\vec{\phi}_{s_t^m a}$ は，制御モジュール $m$ のステップ $t$ での行動 $a$ に対する状態を表す特徴ベクトル，$\vec{\phi}_{s_t^s m}$ は選択モジュールのステップ $t$ でのモジュール $m$ に対する状態を表す特徴ベクトルである．$f_m$ は制御モジュールに対する信用割当関数で，$f_s$ は選択モジュールに対する信用割当関数である．両者とも次式を用いる．制御モジュールの信用割当関数は全て共通とする．

$$f_m(t, r_T, T) = \gamma_{mps}^{T-t-1} r_T$$
$$f_s(t, r_T, T) = \gamma_{sps}^{T-t-1} r_T \qquad (6)$$

ここで，$\gamma_{mps}$ は PS で用いる制御モジュールに対する減衰率，$\gamma_{sps}$ は PS で用いる選択モジュールに対する減衰率である．$T$ は保存したステップ数，$t$ は現在のステップ数であり，過去に対して信用割当する．

各モジュールの強化は，次のようにする．選択モジュールは常に強化する．制御モジュールについては，そのステップで選択されたモジュールのみ強化する．

## 5. 結果と考察

### 5.1 結果

3.2 節で示した課題の学習を行い，また各手法の組み合わせによる性能を比較した．5000 エピソードを 1 試行とし，50 試行行った．学習パラメータは，$\alpha_m = 0.02, \gamma_m = 0.95, \lambda_m = 0.9, \alpha_s = 0.001, \gamma_s = 0.7, \lambda_s = 0.9, \gamma_{mps} = \gamma_{sps} = 0.25$ とした．

図 2(a) は，この場合の学習曲線である．この図から，モジュール型強化学習のみでは学習出来ておらず性能が悪いが，初期学習と強化学習を組み合わせることで，性能が改善されている．また，初期学習と経験強化は，両方を組み合わせた方法が，それぞれを単独で組み合わせた場合より良い性能であった．初期学習の際に経験強化を行った場合と，常に経験強化を行った場合では，結果にあまり違いがなかった．このことから，経験強化は初期学習の際に行うことが重要で，移動制御の学習の際に行っても性能はあまり変わらないことがわかった．

### 5.2 考察

ターゲットとダミーの識別をどのように学習しているかを検討するため，ターゲットや各ダミー付近

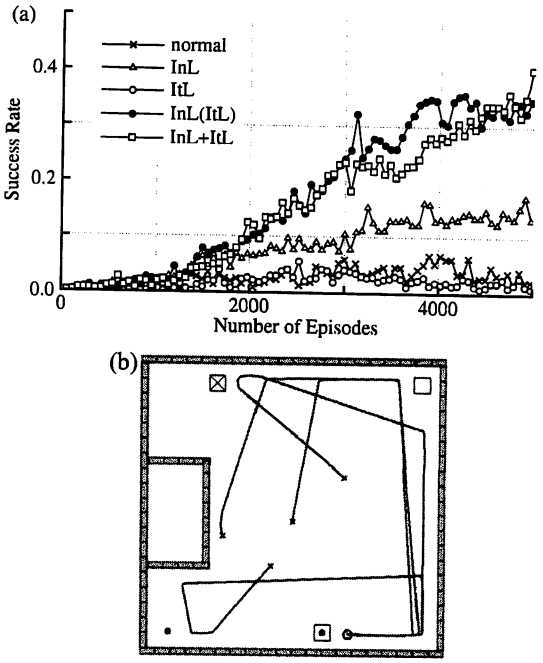


図 2 (a) 学習曲線．× はモジュール型強化学習のみの場合，△ は初期学習を行った場合．○ は経験強化を行った場合．□ は初期学習中に経験強化を行った場合．● は初期学習を行い，常に経験強化を行った場合．(b) 学習された制御によるロボットの軌跡．× はロボットの初期位置，○ はロボットの終点位置，⊡ はターゲット，⊠ は dummy1，□ は dummy2，● は dummy3 を示す．

での選択モジュールへの入力とモジュール選択を比較した．図 3 は，モジュール型強化学習に初期学習と経験強化を組み合わせた方法で学習した場合の，各オブジェクト正面の各距離における選択モジュールの入力（制御モジュールで計算される Q 値の最大値）と，その状態で選択された制御モジュールと，決定された行動を示す．(a) は dummy3，(b) は dummy1，(c) は dummy2，(d) はターゲットである．

図 3 を見ると，ランプを持つオブジェクト (dummy3(a) とターゲット (d)) の前では，光モジュールの Q 値が正の値になっている．一方，正しい絵を貼り付けたオブジェクト (dummy2(c) とターゲット (d)) の前では，カメラモジュールの Q 値が正

の値になっている．選択モジュールは，光モジュールとカメラモジュールのQ値が両方正の場合，カメラモジュールを選択するように学習している（ターゲット (d)）．dummy2(c) の 85mm のところでは，カメラモジュールのQ値だけが大きいのにカメラモジュールを選択している．しかし，他の距離では，障害モジュールが選択され dummy2 を回避している．

以上のように，制御モジュールが，該当する特徴を持ったオブジェクトの前で，正のQ値を出力し，選択モジュールはその組み合わせで，適切なモジュールを選択するように学習していることが分かった．

以上の結果から，初期学習の際に経験強化を行うことの重要性について次のように考えられる．モジュール型強化学習のみで学習を行うと，ランプに近づいた時に成功する場合（ターゲット）と失敗する場合（ランプのみ，dummy3）がある．成功の場合にプラスの報酬，失敗の場合にはマイナスの報酬を与えるという設定なので，ランプの前で光モジュールのQ値が正の大きな値にならない可能性がある．カメラモジュールの場合も同様である．初期学習を行うことで，ランプや正しい絵を貼ったオブジェクトの前で計算される光モジュールやカメラモジュールのQ値が，正の大きな値を持つように学習できる．

また，ターゲットに貼り付けた絵をカメラによって認識しているが，e-puck のカメラの視野角は ±30° 程度であり，この範囲に来ないと成功しないため，成功確率が低い．成功回数が少ないとQ値が大きくならず，識別に必要な差異を獲得できない．初期学習を行うことで，成功の状態を学習し成功確率の低さを改善できると考えられる．

さらに，初期学習中に経験強化を行うことで，成功の状態に対するQ値が大きくなり，ターゲットの識別に必要になる制御モジュールのQ値の差異を，早い段階で学習できると考えられる．その結果，モジュール型強化学習のみでは学習効率が悪かったが，初期学習と経験強化（初期学習中に経験強化）を行うことで学習効率を改善できたと考えられる．

## 6. おわりに

モジュール型強化学習に，初期学習と経験強化を組み合わせることで，複数のセンサ入力の組み合わせから適切な制御の選択（AND 条件の識別）を，学習により獲得できることが分かった．初期学習，経験強化を組み合わせることで，さらに複雑な制御へ適用の可能性があると考えられる．

初期学習と経験強化を組み合わせた手法は，モジュール型強化学習のみの場合に比べ，良い性能であったが，今回報告した結果では，成功率が低く実用的ではない．今後は，さらに学習性能を改善するため，手法の検討が必要である．また本研究では，シミュレーションしか行っていないため，実機への適用について検討する必要がある．これらを検討することで，より実用的な制御課題へ適用していきたいと考えている．

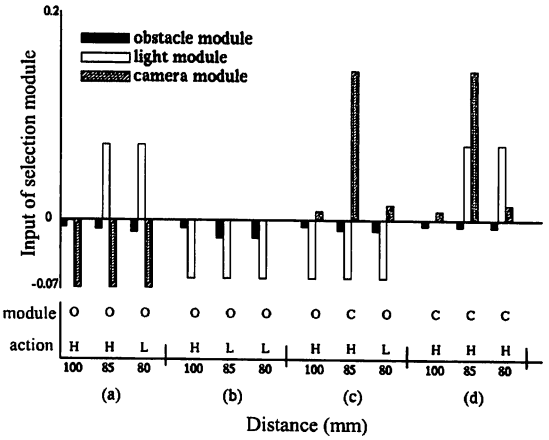


図 3 各オブジェクトまでの各距離における制御モジュールの入力（各制御モジュールで計算されたQ値，■は障害モジュール，□は光モジュール，斜線入り□はカメラモジュール）と，その状態で選択されたモジュール（'O' は障害モジュール，'C' はカメラモジュール），選択された行動（'H' は前進，'L' は左旋回）．(a) は dummy3，(b) は dummy1，(c) は dummy2，(d) はターゲットで，下の数字は各オブジェクトの中心までのそれぞれの距離である．

初期学習と経験強化を組み合わせた手法は，モジュール型強化学習のみの場合に比べ，良い性能であったが，今回報告した結果では，成功率が低く実用的ではない．今後は，さらに学習性能を改善するため，手法の検討が必要である．また本研究では，シミュレーションしか行っていないため，実機への適用について検討する必要がある．これらを検討することで，より実用的な制御課題へ適用していきたいと考えている．

### 参考文献

[1] Sutton, R. S. and Barto, A. G.: *Reinforcement Learning: An Introduction*, The MIT Press (1998).

[2] 山田訓：モジュール型強化学習，信学技報 (1998), NC97-119.

[3] Yamada, S., Watanabe, A. and Nakashima, M.: Hybrid reinforcement learning and its application to biped robot control, *Advances in NIPS*, Vol. 10, pp. 1071–1077 (1998).

[4] 宮崎和光, 山村雅幸, 小林重信：強化学習における報酬割当ての理論的考察, 人工知能学会誌, Vol. 9, No. 4, pp. 580–587 (1994).

# Modular Reinforcement Learning for the Detection of Second Order Correlation of Multi–sensors

Hayato NAKAMA and Satoshi YAMADA*
*Graduate School of Engineering,*
**Department of Inteligent Mechacical Enginnering, Fuculty of Engineering,*
*Okayama University of Science,*
*1–1 Ridai—cho, Kita–ku, Okayama 700–0005, Japan*

(Received September 8, 2009; accepted November 5, 2009)

The modular reinforcement learning system, which is composed of some control modules and a selection module, was developed to apply to the task where several types of sensor information were necessary for the control. In this study, the modular reinforcement learning was applied to the task where the second order correlation of two different sensors must be discriminated. The target (goal) has the correct image and lamp, and other objects have one of them or another image. To discriminate between the target and other objects, the "AND" condition of light sensors and camera must be distinguished. Since the learning efficiency was low, the iterative learning and the initial learning were proposed. As a result, the appropriate module selections and action selections were trained by the modular reinforcement learning.

**Keyworlds:** reinforcement learning; initial learning; iterative learning.