2R-8

ゲート-モジュール型強化学習アルゴリズム

吉田　裕昭†　　松本　友実‡　　中村　真吾¶　　橋本　周司‡

早稲田大学大学院 先進理工学研究科 物理及応用物理学専攻†

芝浦工業大学¶　早稲田大学 理工学術院‡

1. はじめに

強化学習[1]は，エージェントが置かれた環境において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，入力数が多く複雑なシステムの最適な制御器を獲得しようとすると，状態空間が指数関数的に拡大し，膨大な学習時間が必要となってしまう．この問題の解決策として複数の単純な制御器を用意し，系の制御方法を学習するモジュール型強化学習が提案されている[2][3]．しかし，単純な制御器のいずれかを選択するだけの従来手法では，結局のところ，単純な制御しか行うことができない．

そこで，本研究では複数の制御器に対して更にゲートを設け，制御器ごとに報酬を与える手法を提案する．これにより，状態空間の爆発を抑えつつ複雑な系の制御則を獲得することが期待できる．著者らは既にこの手法を考察している[4]が，ゲートでの行動選択において予め人の手によって優先順位をつけなければならないという問題点があった．本稿では行動価値関数を用いることによってゲート選択を自動化し，以前に提案したゲート選択則を含む従来手法との比較を行った．実験では，テレビゲームのキャラクタの操作制御に提案手法を適用し，その有効性を確認した．

2. 提案手法

2.1 アルゴリズム概要

制御目的に応じて複数の制御モジュールを用意する．モジュール m は対応した状態 $\boldsymbol{s}_m$ を観測し，行動 $\boldsymbol{a}_m$ を出力する．行動 $\boldsymbol{a}_m$ の k 番目の要素 $a_{m,k}$ をゲート G_k に渡し，ゲートの選択則に従ってどのモジュールの行動要素を選択するか決定する．このようにして各モジュールの行動要素

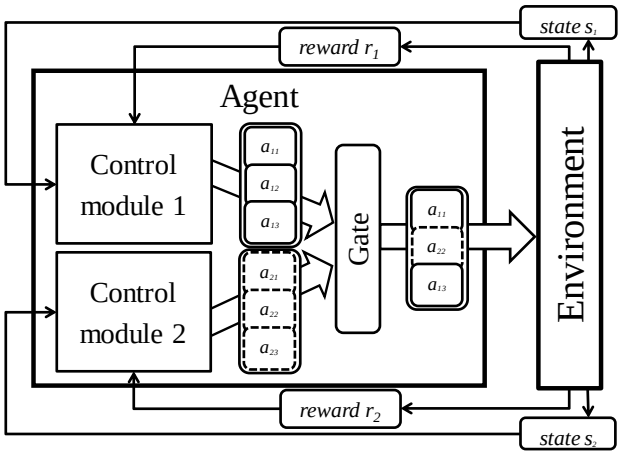

図 1　提案学習アルゴリズム概要例

を組み合わせた行動 $\boldsymbol{a}$ がシステム全体の制御出力となり，より複雑な制御を可能とする．各モジュールの報酬 r_m は制御目的に応じて決定される．図 1 にモジュール数 2，行動要素数 3 の場合のアルゴリズムの概要図を示す．

2.2 ゲート選択則

各モジュール m から得られた行動 $\boldsymbol{a}_m$ の要素の選択方法として，優先選択と多数決選択，重み付き多数決選択の 3 つの方法で実験を行い，評価を行う．

優先選択では各モジュールに予め優先順位を付け，優先度の高いモジュールの行動を優先的に選択するようにする．ただし，優先度の高いモジュールの行動価値 Q_m が他の行動を取る時に比べて低い場合には，次に優先度の高いモジュールについて同様の処理を行う．一方，多数決選択では最も多く出力された行動を選択する．最後の重み付き多数決選択では，各モジュールの出力に重み $w_{m,k}$ をかけたもので多数決をとる．重み $w_{m,k}$ は以下の(2)(3)式によって行動 $a_{m,k}$ を出力する度に更新される．

$$w_{m,k} = Q_m(\boldsymbol{s}_m, \boldsymbol{a}_m) - Q_m(\boldsymbol{s}_m, \boldsymbol{b}_m) + \beta \qquad (2)$$

$$b_{m,n} = \begin{Bmatrix} n = k : a_{m,n} \\ n = k : a^c_{m,n} \end{Bmatrix} \qquad (3)$$

ここで，β は定数であり，$a^c_{m,k}$ は $a_{m,k}$ を除いた k 番目の要素の中で行動価値 Q_m を最大にするものである．

Gates-Modular reinforcement learning algorithm.

†Hiroaki Yoshida, Department of Applied physics, Waseda University

¶Shingo Nakamura, Faculty of Science and Engineering, Shibaura University

‡Tomomi Matsumoto, Shuji Hashimoto, Faculty of Science and Engineering, Waseda University

3. 評価実験

3.1 実験対象

オープンソーステレビゲーム『INFINITE MARIO BROS』[5]に提案手法を適用し，性能を評価した．このゲームはスクロール型アクションゲームで，制御するキャラクタをステージ右端まで進めることを目的とする．ステージには所々に段差や落とし穴，敵等の障害物が設置されており，地形の状態の多さや，時間の経過により状態が刻々と様々に変化するといった点で複雑な制御を必要とする．

ゲームステージは制御キャラクタの幅の 150 倍とし，9 度グラウンドレベルが変化するものとした．更に，落とし穴を 2 つ設置し，踏み倒す以外に接触してはいけない敵キャラクタの数を 1 に設定した．

3.2 モジュール構成

実験では制御目的ごとにモジュールを 3 つ用意した．モジュールの制御目的と状態数を表 1 に示す．各モジュールが扱う状態は，$\boldsymbol{s}_1$ は段差の x, y 座標と高さ，$\boldsymbol{s}_2$ は穴の x, y 座標と幅，$\boldsymbol{s}_3$ は敵の x, y 位置と向きとし，これらにキャラクタの状態（x, y 方向の速度，ジャンプ状態）を加えて，各モジュールに渡す．行動は，ゲームを制御するジャンプ・ダッシュ・方向の 3 種類のボタンの ON/OFF とし，全部で 8 種類とした．各モジュールに与える報酬値を行動結果ごとに表 2 に示す．また，学習率 α=0.3，割引率 γ=0.9，τ=5.0 として実験を行った．

3.3 結果

提案手法を評価するため，優先選択，多数決選択および提案手法である重み付き多数決選択および通常の強化学習の 4 つの手法を用いて実験を行った．なお，優先選択における選択則のモジュール優先順位は，「落とし穴飛び越え」「敵撃破」「段差飛び越え」モジュールの順とした．Q 値の更新は 3 フレームごとに行われ，全モジュールの更新回数の総和が 1000 に達するごとに，ステージクリア率を求めた．その結果を図 2 に示す．

優先選択と重み付き多数決選択は，通常強化学習よりも早い段階で高いステージクリア率を達成していることが確認できるが，重み付き多数決選択は優先順位を予め人の手でつけなくて良いという点でより汎用性の高いアルゴリズムと言える．また，多数決選択ではあまり良い結果が得られなかった．

表 1　モジュールの制御目的と状態数

モジュール	制御目的	全状態数
1	段差を越えて先に進む	3780
2	落とし穴を飛び越える	3528
3	敵を倒す、または避ける	4860

表 2　報酬定義

種類	行動結果	値
r_1	右に進む	1 ステップに進んだ距離
	壁に衝突	-20
r_2	落とし穴を飛び越える	+100
	落とし穴に落ちる	-100
r_3	敵を倒す	+100
	敵と接触	-100
	敵を飛び越える	+10

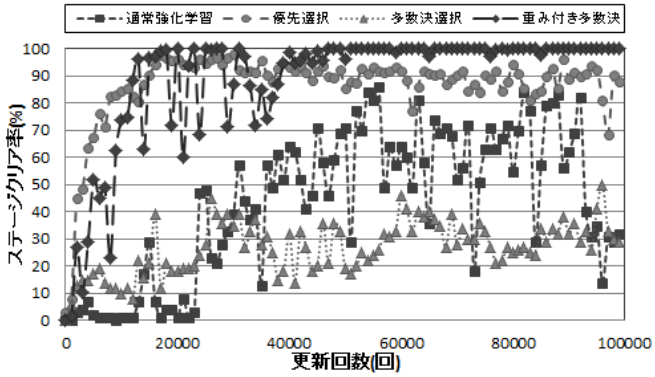

図 2　更新回数とステージクリア率の関係

4. まとめ

複数の報酬とゲート機構を用いた学習アルゴリズムを提案し，テレビゲームのキャラクタ制御に適用することでその有効性を確かめた．予め優先順位をつけなければならない優先選択とほぼ同程度のクリア率を示した重み付き多数決選択はより汎用性の高いアルゴリズムである．今後はさらに遺伝的アルゴリズムや集団学習を組み合わせることによって制御モジュール自体を自動生成するなど，汎用性と柔軟性の高いアルゴリズムに改良していきたい．

謝辞

本研究の一部は，早稲田大学ヒューマノイド研究所，グローバルCOE プログラム「グローバル ロボット アカデミア」の研究助成を受けて行われた．

参考文献

[1]Richard S.Sutton and Andrew G.Barto, “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，NC97(623), pp.139-146, 1998.
[3]中間隼人ら，“3 種類のセンサを持つロボット制御へのモジュール型強化学習の適用”, 電子情報通信学会, NC108(480), pp.301-306, 2009.
[4]吉田裕昭ら，“複数の報酬とゲート機構を用いた モジュール型強化学習アルゴリズム”, 情報処理学会第 74 回全国大会, pp.293-294, 2012.
[5]M.persson, “INFINITE MARIO BROS” Available: http://www.mojang.com/notch/mario/