実例に基づく強化学習法の頑健性向上に関する一考察 —マルチロボットシステムによる検証—

保田 俊行*・大倉 和博**

Improving the Robustness of Instance-Based Reinforcement Learning for Multi-Robot Systems

Toshiyuki YASUDA* and Kazuhiro OHKURA**

A behavior acquisition mechanism based on reinforcement learning (RL) for an autonomous multi-robot system (MRS) is extended to improve the robustness in this paper. An important fundamental topic for a MRS is how to design an on-line behavior acquisition mechanism that is capable of developing a set of roles for coordinated behavior as well as assigning them to each robot in a timely manner in a physical environment. From this point of view, a robot controller is constructed with RL that has the mechanism of segmenting continuous state space and continuous action space simultaneously and autonomously to extend the adaptability to non-Markovian environment. This RL developed in our research group is called Bayesian-discrimination-function-based Reinforcement Learning (BRL). Although BRL is proved to be effective to develop a globally stable behavior to achieve a task for a MRS, it has the tendency of being brittle against disturbance after successful learning. This problem is generally called over-fitting. This paper shows an approach to extending BRL for overcoming this problem. This extended BRL is examined by three arm-type autonomous robots, the task of which is lifting an object without tilting it. The experiments conducted illustrate the robust performance of the proposed method by developing various types of cooperative behavior as a result of autonomous specialization.

Key Words: multi-robot systems, autonomous specialization, reinforcement learning, over-fitting, robustness

1. はじめに

近年、ロボティクスにおける従来手法のモデルベースコントロールとは異なる枠組として、「移動知」に関する研究が注目を集めている¹⁾。この枠組では、行動主体の能動的な動きを前提とすることで、非構造的で予測不能的に変動する環境に柔軟に適応するための知能の発現・設計原理に関する非経験的なモデルを構築することが目的である。そのようなモデル構築には、多数の運動自由度を統合して適切に運動したり、多数の運動パターンから目的に適合したものを選択・実現する能力が必要であるとしており、それらを実現するための構成論的な研究が進められている。このような行動主体の動きによる身体、制御系と環境と相互作用に着目する手法に、強化学習²⁾がある。強化学習ロボットは試行錯誤を行ない、環境との相互作用を通じた適切な入出力関係を構築・修

正する。また、学習には教師データを必要としないといった特徴がある。これまでに、障害物回避・壁伝い・ゴール到達行動³⁾、歩行動作⁴⁾や起立運動⁵⁾などの獲得が実環境において実現されている。

複数の自律ロボットで構成されるマルチロボットシステム (Multi-Robot System: MRS) では、ロボット間の協調によってタスク達成能力や作業効率が大幅に向上することが期待される。しかし、協調作業を実現するには、ダイナミクスが複雑であることや、故障などのシステム内における変化が発生しやすいといった問題に対処することが必要となる。強化学習を適用した場合、先に述べた特徴に起因して、この問題を陽に意識することなく、ロボット間の自律的な相互作用によって大域的な振舞いを生成する可能性がある。

しかし、強化学習を MRS に適用するには、つぎに挙げる二つが原因となりロボット単体の場合と比べて困難性が増大する。一点目は、学習の成否は状態・行動空間の離散化に大きく依存するが通常の強化学習はあらかじめ離散化された状態・行動空間が求められることである。この離散化に関する一般的な設計方針は現在のところ存在しない。二点目は、同時に学習するほかロボットの存在に起因する動的な環境である MRS では強化学習の有効性が明示的に示せないというこ

* 津山工業高等専門学校

** 広島大学大学院工学研究科

* Tsuyama National College of Technology

** Graduate School of Engineering, Hiroshima University
(Received December 27, 2005)

(Revised July 10, 2006)

とである。これは、ある程度の環境中のダイナミクスを許容するために付加情報を学習器に与えることでMRSへの適用に成功した研究例⁶⁾や Profit Sharing の有効性を示した研究⁷⁾があるものの本来、強化学習はマルコフ環境を前提としていることに起因する。

これらの問題に対処するため、筆者らは学習能力の高い強化学習手法として連続な状態・行動空間を自律的に分割する機能をもつ Bayesian-discrimination-function-based Reinforcement Learning (BRL) を開発し、MRSにおける協調行動獲得を実現してきた⁸⁾。しかし、その後のBRLの追加実験では、機械学習における典型的な課題である過学習の影響が観測されることもあった。すなわち、安定な協調行動獲得後もさらに学習を継続すると頑健性が次第に落ち、システム内の変化や外乱に対して振舞いが不安定になった。

本論文では、協調行動をするMRSにおいてBRLを適用し、その過学習に対処することで環境変化に対する頑健性を向上させる手法を提案する。以下、まず、実験で取り扱う協調タスクと実験に用いるロボットについて述べる。つぎに、学習過程において有効であったルールを保護して多様性のあるルール集合を構成することによる過学習への対処法、およびそのルール保護のための基準を示す。その後、実機実験を行ない、提案手法の有効性を検証する。

2. アーム型マルチロボットシステムの協調荷上げ問題

2.1 問題設定

協調タスクとして三台のアーム型ロボットによる協調荷上げ問題を扱う (Fig. 1)。各ロボットのアーム先端は正三角形の荷の各頂点に回転自由に連結されている。各ロボットは自身の各関節の角度と荷の傾きを知覚するのみであり、直接的に他ロボットの状態を知覚することはできない。また、他ロボットとの通信や報酬伝播などは行わない。

問題の特徴はつぎのように記述できる。

- ロボットごとに独立した学習器をもち、互いに非同期に制御される。
- 実空間での行動学習であるため、センサ・ノイズや決定行動と実行された行動のズレが生じる。
- 荷が所定の高さに到達したかどうかは、荷の三つの頂点付近に取り付けられたIRセンサにより判定される。そのため、ノイズによってタスク達成の判定が正しく行なえない

Fig. 1 Three arm-type autonomous robots

い場合があり、評価系にも不確定な要素が含まれる。

- ロボットは受動車輪をもち、他ロボットに押されるまたは引っ張られることで前後に移動する。

三台のロボットは機能や構造的な初期状態は均質であるが、非同期で動作するためタスクの達成にはそれぞれが異なる働きをしながらシステム全体として安定した協調的振舞いをしなければならない。そのため、各ロボットは自身の動作だけでなく、集団の一員としての役割を獲得し、適切に振舞う必要がある。本論文では、このような協調過程を自律的機能分化と呼ぶ。

2.2 MRSの自律的機能分化

一般的なMRSの研究では効率的な協調行動の実現が目標であるため、それに必要なロボットの役割や行動をあらかじめ与えることがほとんどである^{9), 10)}。このアプローチにおけるMRSは機能・形態的に非均質であり、機能分化はすでにされているといえる。一方、本論文では均質なMRSを取り扱う。そして、機能分化も相互作用を通して自律的に獲得させることを目指す。

同様の視点による研究は、われわれの知る限りでは、制御器として進化型人工神経回路網を適用した例があるのみである。Baldassarre *et al.*¹¹⁾は四台の光源到達問題の計算機実験、Quinn *et al.*¹²⁾は三台のフォーメーション形成問題の実機実験^(注1)を行なっている。しかし、進化型人工神経回路網によるアプローチは解の探索能力が高いものの、大きな計算量が必要になるため、行動獲得に時間がかかる傾向にある。

3. 強化学習による頑健な協調行動の獲得

3.1 ベイズ判別法を用いた強化学習法：BRL

MRSの自律的機能分化を短時間で実現することを目指し、強化学習を適用する。なかでも、われわれの研究グループで開発したBRLを用いる。BRLは実例に基づく強化学習の一種であり、連続な状態・行動空間を自律的に分割することができる。状態空間はルールクラスタを表わす多変量正規分布によって分割される。各ルールは状態ベクトル、分散共分散行列、事前確率、有効度、センサ入力サンプル集合、および動作で構成される。

BRLでは統計的にパターン分類を行なうベイズ判別法を用いて入力がどのクラスに分類されるかを識別する。ルールがもつ動作はただ一つであるため、入力が属するクラスが識別されるとそのときの行動出力は一意に決定する。行動空間の探索は後述するルールの追加時に行なう。

環境の確率モデルが事前にわかっていることは稀であり、自律ロボットの行動学習では入出力の完全なデータセットを用意することは困難である。そこで、観測データから環境の確率モデルを推定する必要があるため、BRLでは、(1) ルールの追加と削除、(2) 観測データに基づく環境の確率モデルを区間推定法を用いてリアルタイムに更新することで、状態空

(注1) 計算機実験で得られた結果を実機で利用

間の分割を行なう。なお、バイズ判別法での事後確率の最大値が事前に設定したしきい値以下のときを未経験の状態であるとみなして、ルールを追加を行なう。そのとき、状態ベクトルはそのときの入力データそのもの、動作はランダムな実数値、そのほかは固定の初期値として決定する。このように、実際に経験した状態の集合によって定義される領域の中でタスク達成可能な領域に絞りで学習が進行することで、安定な行動を獲得することができる⁸⁾。このような特徴に起因してBRLは汎用性を持ち、本論文のシステムだけでなく、異なる台数での協調荷上げ問題⁸⁾や移動ロボットによる協調搬送問題¹³⁾において、協調行動の獲得に成功している。

3.2 環境変化に対する頑健性の低下

3.2.1 BRLにおける過学習の発生メカニズム

ロボット単体のシステムと比べて、MRSではシステム内部の変化が生じやすくなる。BRLを適用した場合、パラメータの更新やルールの生成をすることで、そのような変化に対してある程度の頑健性を示すことを確認できている。しかし、行動獲得後に実験を続けると、徐々にその頑健性が損なわれる。これは、タスク達成に寄与しないルールは削除され、寄与するルールのみが強化されてルール集合に残ることが原因である。具体的には、(i) 変化前に頻繁に用いられていたルールは、有効であるか否かにかかわらず変化後も頻繁に用いられること、および(ii) ルールが少なくなるとともにルールで覆われた状態空間が狭まるに従い未経験なセンサ入力を得ることが多くなってルール追加時のランダムな行動探索が振舞いに影響を与えることによって頑健性が損なわれる。すなわち、BRLでは環境に特化したルール集合となる結果、過学習状態となるためにシステムが不安定になる。このような従来型BRLでは環境変化への頑健性の低下に対して、対処法が必要となる。

3.2.2 環境変化を考慮した強化学習法

変化する環境での適応を目指した強化学習に関する研究が行なわれている。大きく、二つのカテゴリーに分けることができる。

まずひとつめは、追加学習をするアプローチである。石井¹⁷⁾は逆温度パラメータを制御することにより exploration と exploitation のバランスを制御することで学習が収束しても常に新たな探索を行なう手法を提案している。Katoら¹⁸⁾は Profit Sharing に忘却の概念を導入したアプローチを提案している。福田ら¹⁹⁾は異なる学習率の学習器を用意して環境変化の認識を行ない、制御器の切り替えと学習知識の上書き修正をする手法を提案している。この研究では、実際の意思決定を行なう学習器と強化値の更新のみを行なう学習器を用い、ふたつの学習器の強化値の相関がなくなった場合に環境が変化したとみなしている。Sutton¹⁶⁾の Dyna-Q では学習と同時にプランニングによって環境のモデルを構築することで環境に適応している。

ふたつめは、再学習をするアプローチである。港ら¹⁴⁾はナビゲーション問題においてタスク成功率の低下によりロボッ

トに環境の変動を認識させ、問題が起こった状態付近のみ行動政策を修正する手法を提案している。松井ら¹⁵⁾はロボットが過去の環境で獲得した行動政策を新たな環境で適用するための政策事前条件を概念学習を用いて獲得し、政策を部分的に学習し直す手法を提案している。

3.3 過学習の抑制のためのBRLの拡張

3.3.1 ルール集合の多様性維持によるアプローチ

本論文で用いるBRLは実例に基づく強化学習法であることやバイズ判別法を用いるという特徴をもつ。そのため、保持ルール数が可変であるため、過学習状態では保持ルール数が過小となって状態空間が狭くなっている。

実環境ではノイズなどの外乱やロボットの故障などの不確定要素により、なんらかの環境変化が常に生じているといえる。本研究では、環境の変化を陽に意識することなく、さまざまな環境に対応し得るルール集合を構成することで頑健性の向上をはかることを狙う。すなわち、上述の過学習状態の問題に対処するために学習過程で有効であったルールを削除することなく保持しておき、将来の不確定要素に対応し得るように多様性のあるルール群により状態空間を幅広く覆うことで頑健な大域的秩序形成を試みる。ルールの保持により、(i) 特定ルールが発火し続ける確率の上昇を抑制する、(ii) ランダムな行動探索であるルール生成の頻度を抑制するという効果がある。また、(iii) 保持していたルールがはじめに獲得した行動とは異なる安定な行動に寄与することも期待できる。

しかし、ルールの削除を行わずに単に学習過程において生成したルールをすべて保持すれば、ルール数が増加するのみである。この場合、明らかに不要なルールであっても残り続けることで、学習効率が低下する可能性が生じる。そこで、学習過程において削除から保護すべき信頼度の高い有効ルールを選定する基準が必要となる。

3.3.2 有効ルール保護のための指標

強化学習は試行錯誤を通して徐々に適切な入出力関係を構築する手法であるため、学習進度を基にルールの信頼度を判断できると考える。この学習進度は学習効率の向上を主な目的とした研究で用いられている。石井ら¹⁷⁾は状態価値関数の逆分散を行動の確信度として各状態における学習の進行度を定義している。また、尾川ら²⁰⁾は予測誤差に基づき更新される内部モデルの信頼度²¹⁾を学習進行度の指標として用いている。

本論文ではこれらの研究と同様に、学習を通して更新されるパラメータを基に学習進度を計算する。すなわち、パラメータが更新されたルールほど学習が進んだものとする。BRLのルールパラメータのうち、有効度と正規分布の特徴量である分散共分散行列が学習の進行に伴って変化する。有効度は、Profit Sharing と Bucket Brigade 戦略により報酬を過去に遡って伝播させることで更新される。そのほかの有効度の更新法には、ループ行動を防ぐために選択されたルールに課すコスト、報酬獲得に寄与しないルールを削除してメモリ消費量を抑えるためにタスク達成時に全ルールに作用させる消散が

ある。この四つの更新方法のうち、Profit Sharing と Bucket Brigade 的戦略によって有効度は向上するが、このときの更新量はゴール付近のルールほど大きい。また、ゴール判定にノイズによる不確定な要素が含まれる。そのため、有効度の大小のみで、エピソードを通してのルールの構成度合を判断することはできない。そのため、BRLにおける分散共分散行列に注目する。

分散共分散行列の各成分は、そのルールが発火した場合に次式により更新される。

$$s_j^2 \leftarrow s_j^2 + \alpha^2 [s_j^2 - s_j^2]. \quad (1)$$

ここで、 $s_j^2 = \sum_{k=1}^{n_s} (\phi_k^j - \bar{x}_j)^2 / (n_s - 1)$, $\bar{x}_j = \sum_{k=1}^{n_s} \phi_k^j / n_s$ であり、 ϕ_k^j は k 番目 ($k \in [1, n_s]$) のサンプルデータの j 次元成分 ($j \in [1, n_d]$) を表わす (n_s : ルールが記憶しているサンプル数, n_d : 入力次元数, α : 定数)。分散共分散行列の各成分の更新量はルールの用いられる回数が増えるほど大きくなる。つまり、学習が進行して行動は安定して特定のルールの発火する回数が増えるほど各成分の値は小さくなる。

3.3.3 BRL の拡張

これから、分散共分散行列のすべての成分に関してどれだけ更新されたかを考慮するためにその行列式を用いることとし、構成状況の指標 $\mathcal{R}$ を次式のように定義する。

$$\mathcal{R}_i = \frac{|\Sigma_i|}{|\Sigma_0|} \quad i = 1, 2, \dots, n. \quad (2)$$

ただし、 Σ_0 は分散共分散行列の初期値、 n はエピソード終了時におけるルール数である。 $\mathcal{R}_i$ はルール i の分散共分散行列がルールが生成された時点と比較してどれだけ変化したかを表わし、多く発火しているルールほど小さくなる。

従来のBRLではタスクを達成した場合、すべてのルールの有効度を η 倍 ($0.0 < \eta < 1.0$) して減じさせる (消散)。これを変更し、各ルールの有効度の更新法のうち、消散は $\mathcal{R}_i$ が事前に設定した定数のしきい値よりも大きいルールについてのみ適用することとする。すなわち、しきい値 $\mathcal{R}_{th}$ ($0.0 \leq \mathcal{R}_{th} \leq 1.0$) よりも $\mathcal{R}_i$ が小さいものを有効ルールとして消散を適用せず、削除から保護することでルール集合の多様性を維持し、過学習に陥るのを防ぐ。

$$u_i \leftarrow \eta u_i \quad \text{if } \mathcal{R}_i > \mathcal{R}_{th} \quad (3)$$

$\mathcal{R}_{th}$ が大きいほど、消散の影響で削除されるルールは少なくなる。0.0 のときは従来型 BRL と等価であり、1.0 のときは消散はどのルールにも適用されない。

4. 実験

4.1 実験設定

Fig. 2 にアーム型ロボットの制御システムを示す。アームは三関節をもち、それぞれに取り付けられたサーボモータの動作範囲は $0 \sim 90$ 度で 0.35 度の精度で制御できる。各関節角度 (p_1, p_2, p_3)、およびアーム先端における荷のピッチ角 (Pit) とロール角 (Rol) を測定するためにポテンシオメータ

(角度センサ) がそれぞれ取り付けられている。各ロボットの制御システムはロボットのサーボモータとセンサを管理する。各ロボットの意思決定はそれぞれの PC に実装された学習システムによって行なわれる。

ロボットは床から約 130mm の高さから荷を持ち上げ始める。荷の各頂点付近に取り付けた IR センサのすべての値がしきい値 θ_{IR} を越えた時点でゴールとする。そのときの床から荷までの高さは約 270mm である。入力 $x = \{Pit, Rol, p_1, p_2, p_3\}$ は正規化した連続値を用いる。出力 $y = \{\theta_1, \theta_2, \theta_3\}$ は関節の角度変化量であり、 θ_i は $|\theta_i| < \theta_{max}$ の範囲の実数として出力される ($\theta_{max} = 9^\circ$)。入出力の1サイクルを1ステップとし、ゴールに到達するか、あるいは120ステップまでにゴールできない場合にエピソードを終了・更新する。エピソードの開始は手動で行ない、終了後は手動で初期状態に戻す。強化信号 (正: 報酬, 負: 罰) は、それぞれ以下の場合に与える。

● 報酬: ゴール到達時

● 罰: 荷の傾きがしきい値 ($T_1 \approx 27^\circ$) 以上になったとき
実験で用いた学習パラメータを Table 1 に示す。なお、これらのBRLのパラメータは予備実験から得た推奨値である^{8), 13)}。

4.2 大域的秩序獲得実験: 実験 1

本論文の目的である環境変化に対する頑健性の検証に先立つ予備実験として、拡張型 BRL を用いた協調荷上げ行動獲得実験を行なう。なお、BRL の学習は経験に基づく実行可能解、すなわち局所解に収束するために、試行によって獲得する行動は異なる。十数回行なった実験のすべてでタスク達成に成功している。ここで示す結果はそのうちの一例である。

■ 学習履歴

Fig. 3 に学習履歴を示す。Fig. 3(a) は各エピソードにお

Fig. 2 Control system

Table 1 Parameters of learning system

n_{max}	maximum size of the rules	100
P	payoff(reward)	30.0
P	payoff(penalty)	$-0.01u$
u_0	initial utility	10.0
u_{min}	threshold for extinction	$0.92u_0$
c_f	cost for an action	0.007
γ	distribution rate of utility	0.5
κ	utility spread rate	0.15
η	evaporation rate	0.98
f_0	initial prior probability	0.001
σ_0	initial variance	0.04
$1 - \phi$	confidence limit	0.99
α	constant in eq.(1)	0.1
$\mathcal{R}_{th}$	threshold for evaporation	0.90

いて終了までに要したステップ数^(注2)と各ロボットが罰を受けた回数である。実験開始後、タスクを達成できず、多くの罰を受けるエピソードが続く。その後、荷を傾けずにゴールする行動を獲得している。Fig. 3(b) は各エピソード終了時にロボットが保持しているルール数である。実験初期は未経験の領域が多いために多くのルールを生成し、保持するルールが徐々に増えている。学習収束後、保持しているルール数はほぼ変化していない。これは、発火しているルール以外は徐々に削除される従来型 BRL と異なる傾向である。すなわち、ルール構成の進行度に基づく学習過程における有効ルールの保護により、ルール集合の多様性が維持できている。

■状態空間の構成

各ロボットは自身の経験に基づいて状態・行動空間を自律的に分割し、その分割具合の違いにより異なる役割を果たすようになる。ここでは、状態空間を可視化するために、各学習器が記憶しているルールの特徴ベクトルに関して相関行列による主成分分析を行ない、第一・第二主成分からなる二次元状態空間を生成する。各ルールの特徴ベクトルと分散共分散行列を新たに生成した状態空間に射影し、状態空間を観測する。

Fig. 4 に 100 エピソードにおける各ロボットの状態空間を示す。図中の四角は獲得されたルール、矢印はルールの発火系列を表わす。白四角は提案手法によって保護されたルール、すなわち従来の BRL では消散の作用により削除されるルールである。ロボットごとに異なる数・粒度のルールで状態空間を分割していることがわかる。さらに、発火ルール以外にもルールを削除せずに保護することで、状態空間を広く覆っていることがわかる。

Fig. 3 Learning history for Experiment-1

(注2) 非同期の制御のために各ロボットで若干異なるが、ここではロボット A のものを示す。

■獲得した振る舞いの観測

前節に記述した状態空間の構成に基づいて機能分化が生じる。その結果、各ロボットは状況に応じて動作を切り替え、荷をしきい値以上に傾けずに持ち上げる協調行動を実現する。

Fig. 5 は、100 エピソードにおける各ステップで観測したセンサ値を用いてロボットの姿勢を再現したもの、および発火ルールの番号である。ただし、センサ値はノイズを含むために、厳密には実際の姿勢と一致しない。また、非同期であるために、横軸はロボットによって異なる。各ロボットの振る舞いは、以下のとおりである。

ロボット A: まず、ロボット B に次ぐ早さで荷を持ち上げる。その後、いったん荷を下ろす。最終的に、ロボット C が高く持ち上げた後に持ち上げを再開してタスク達成に至る。どちらかのロボットの傾きに応じて行動を変更する調整役を果たしているといえる。

ロボット B: アームを大きく動かしながら持ち上げた後、持ち上げる速度を落とす。最終的には高さを維持する。常に持ち上げを先導し、リーダ役を果たしているといえる。

ロボット C: 後半まで少しずつ持ち上げ、ロボット A, B がある高さまで持ち上げた後に持ち上げる速度を上げる。他のロボットに追従するフォロワ役を果たしているといえる。

4.3 システムの頑健性の検証実験: 実験 2

提案手法による環境変動に対する頑健性の向上を検証するため、安定な行動を獲得後に一台のロボットを初期化する。これは、一台のロボットが故障した場合に新しいロボットを投入することと等価である。初期化の後、残りの二台と新たな荷上げ行動を実現しなければならない。なお、予備実験から調整役(ロボット A)とフォロワ役(ロボット C)を初期化した場合は、従来型 BRL でも安定に行動を獲得できることを確認している。ここでは従来型 BRL では過学習の影響が最も大きかったリーダ役のロボット B を初期化した実験について詳細を述べる。

■学習履歴

Fig. 6 に拡張型 BRL を用いた場合の学習履歴を示す。Fig. 6(a) は各エピソードでエピソード終了までに要したステップ数と各ロボットが罰を受けた回数である。図が示すように、環境変化後にシステムの振る舞いは一時的に不安定になってタスクを達成できていない。しかし、その後は安定してタスクを達成している。また、Fig. 6(b) は各ロボットの BRL におけるエピソード終了時の保持ルール数とエピソードごとの生成ルール数の推移である。ロボット A と C はロボット B のランダムな行動に合わせて新しい協調行動を獲得するためにルールを生成するが、ロボット B はそれまでの知識がないためにより多くのルールを生成する。再び安定な行動を獲得した後もノイズなどの影響によって若干のルール生成があるものの、ここでもルールの保護の影響により、ルール数の増減は小さい。

■獲得した振る舞いの観測

Fig. 7 に 150 エピソードにおけるシステムの振る舞いを

Fig. 4 Projected segmented state spaces of the three robots for Experiment-1 in the 100th episode

Fig. 5 Postures of the robots for Experiment-1 in the 100th episode

示す。まずロボット A, B が持ち上げ (Fig. 7(b)), ロボット A はさらに持ち上げ続ける (Fig. 7(c))。ロボット A が持ち上げの速度を落としている間に、ロボット B, C が持ち上げる (Fig. 7(d))。その後、ロボット A と C がゴール位置まで持ち上げ (Fig. 7(e)), ロボット B が最後に持ち上げを完了する (Fig. 7(f))。ロボット A がリーダー役を、ロボット B と C がフォロワー役を果たしている。これは、実験 1 とは異なる機能分化の形態である。

つぎに、Fig. 8 に、このときの各ロボットの姿勢と発火ルールの番号を示す。図中の括弧なし、中括弧付き、および大括弧付きのルールはそれぞれ 100 エピソードで発火していたルール、提案手法によって保護されていたルール、新たに生成したルールを表す。ロボット A と C は、保護ルールの利用と新ルールの生成によって 100 エピソードとは明らかに異なる行動を取っている。このことから、システム変動に

(a) Transitions of time steps and punishments

(b) Transitions of numbers of rules

Fig. 6 Learning history for Experiment-2

合わせて適応的に行動を調整していることがわかる。

4.4 比較実験

従来型 BRL を用いて比較実験を行なう。前節の実験 2 と同様に、101 エピソードにおいてロボット B の学習結果が初期化されるとい環境変化に対して、システムの振舞いを解析する。ここで、比較のために獲得した協調行動を同じものにする必要があるため、従来型 BRL で獲得したルールは拡張型 BRL の実験 1 において獲得されたルール群から提案手法により保護されていたルール (Fig. 4 における白四角) を削除したものとして仮想的に再現している。従来型 BRL と提案型 BRL では、学習途中に関してはほぼ差異はなく、安定してタスク達成が可能になった後 (実験 1 の 71 エピソード以降) のルール集合の構成に関して影響を与えるのみであるため、この操作による有利・不利はない。

■学習履歴

Fig. 9 に学習履歴を示す。Fig. 9(a) は各エピソードでエ

5. おわりに

本論文では、自律的機能分化するMRSのための強化学習の頑健性を向上させる手法を提案した。状態行動空間を自律的に分割する機能をもつ強化学習法・BRLをMRSの制御器として用い、過学習を抑制してシステムの頑健性を向上させるために学習過程で有効であったルールを保護することでルール集合の多様性を維持する機構を付加した。その有効性を検証するために、三台のアーム型自律ロボットによる協調荷上げ問題に適用して実機実験を行なった。その結果、環境変化に対して可塑的な自律的機能分化をすることでシステムの頑健性が向上し、大域的秩序を迅速に再形成することを確認した。

MRSは移動知が対象とする分野の典型例である。本稿では、この枠組において自律的機能分化に基づくアプローチが有効であることを示した。また、機能分化における可塑性の実現例を示した。現状として、提案手法はさまざまに分類されるMRS問題²²⁾の中でも比較的簡単なタスクを取り扱っているだけであるが、このほば未開拓な領域に対する今後の展開の一指針になると考える。

参 考 文 献

- 1) 土屋ら：移動知：能動的な移動機能がもたらす創発的知能，計測と制御，**44**-9, 579/645 (2005)
- 2) R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction, MIT Press (1998)
- 3) 中村, 黒山, 山田, 大倉, 上田: Instance-Based Classifier Generator による自律ロボットの行動獲得, 日本ロボット学会誌, **17**-3, 61/69 (1999)
- 4) 木村, 山下, 小林: 強化学習による4足ロボットの歩行動作獲得, 電気学会電子情報システム部門誌, **37**-12, 1147/1155 (2002)
- 5) J. Morimoto and K. Doya: Acquisition of Stand-up Behavior by a Real Robot using Hierarchical Reinforcement Learning; *Proc. of International Conference on Machine Learning*, 623/630 (2000)
- 6) 大倉, 川上, 上田: 均質な自律ロボット群による協調行動獲得問題: 機能分化に基づくアプローチ, システム制御情報学会論文誌, **15**-9, 451/458 (2002)
- 7) 荒井, 宮崎, 小林: マルチエージェント強化学習の方法論: Q-learning と Profit Sharing による接近, 人工知能学会誌, **13**-6, 609/618 (2001)
- 8) 山田, 大倉, 上田: 強化学習による自律型アームロボットの協調行動獲得, 計測自動制御学会誌, **39**-3, 266/275 (2003)
- 9) L. E. Parker: ALLIANCE: An Architecture for Fault Tolerant Multi-Robot Cooperation; *IEEE Transactions on Robotics and Automation*, **14**-2, 220/240 (1998)
- 10) P. Stone and R. S. Sutton: Scaling Reinforcement Learning toward RoboCup Soccer, *Proc. of 18th International Conference on Machine Learning*, 537/544 (2001)
- 11) G. Baldassarre, S. Nolfi and D. Parisi: Evolution of collective behaviour in a team of physically linked robots; *Applications of Evolutionary Computing*, 581/592 (2003)
- 12) M. Quinn, L. Smith, G. Mayley and P. Husbands: Evolving Team Behavior for Real Robots; *Proc. of EP-SRC/BBSRC International Workshop on Biologically-Inspired Robotics*, 217/224 (2002)

- 13) T. Yasuda and K. Ohkura: Autonomous Role Assignment in Homogeneous Multi-Robot Systems, *Journal of Robotics and Mechatronics*, **17**-5, 596/604 (2005)
- 14) 港, 浅田: 環境の変化に適應する移動ロボットの行動獲得, 日本ロボット学会誌, **18**-5, 706/712 (2000)
- 15) 松井, 犬塚, 世木, 伊藤: 強化学習結果の再構築への概念学習の適用, 人工知能学会論文誌, **17**-2, 135/144 (2002)
- 16) R. S. Sutton: Integrated Architectures for Learning, Planning, and Reacting Based on Approximating Dynamic Programming; *Proc. of the Seventh International Conference on Machine Learning*, 216/224 (1990)
- 17) 石井 信: 強化学習におけるランダムさの自己調節, 日本神経回路学会誌, **6**-4, 254/262 (2002)
- 18) S. Kato and H. Matsuo: A Theory of Profit Sharing in Dynamic Environment; *Proc. of the Sixth Pacific Rim International Conference on Artificial Intelligence*, 136/145 (2000)
- 19) 福田, 船戸, 新井: 多重強化学習法に基づく群ロボットシステムにおける環境変化認識, 日本機械学会論文集 (C 編), **66**-643, 864/869 (2000)
- 20) 尾川, 並木, 石川: 学習進捗を反映した割引率の調整, 電子情報通信学会ニューロコンピューティング研究会, 電子情報通信学会技術研究報告, **NC2002**-129, 73/78 (2003)
- 21) 阪口 豊: 動きの予測を伴う能動的認識のアルゴリズム, 日本ロボット学会誌, **12**-5, 708/714 (1994)
- 22) 太田, 新井: 群知能ロボットシステム, 日本ロボット学会誌, **20**-5, 487/490 (2002)

[著 者 紹 介]

保 田 俊 行 (正会員)

2000年神戸大学工学部機械工学科卒業, 2006年同大学大学院自然科学研究科博士後期課程修了。2006年4月より津山工業高等専門学校電子制御工学科助手, マルチロボットシステム, 強化学習などの研究に従事。博士(工学)。システム制御情報学会, 日本人工知能学会などの会員。

大 倉 和 博 (正会員)

1990年北海道大学大学院工学研究科情報工学専攻修士課程修了。(株)富士通研究所, 神戸大学工学部助手・助教授を経て, 2006年広島大学大学院工学研究科教授。EC, RL, AI, MAS, 生産システムなどの研究に従事。博士(工学)。98年度本学会論文賞, 2003年度システム制御情報学会論文賞砂原賞, 2004年度ファナックFAロボット財団論文賞受賞。日本機械学会, 精密工学会, システム制御情報学会などの会員。