マルチロボットシステムにおける頑健性向上のための知識の獲得・利用法の構築

Preservation and Application of Acquired Knowledge for Improving a Robustness of Multi-Robot Systems

○ 坂ノ上 隼紀 正 保田 俊行 正 大倉 和博 (広島大)

Junki SAKANOU, Hiroshima University, sakanoue@ohk.hiroshima-u.ac.jp

Toshiyuki YASUDA, Hiroshima University, yasu@hiroshima-u.ac.jp

Kazuhiro OHKURA, Hiroshima University, kohkura@hiroshima-u.ac.jp

A multi-robot systems (MRS) is composed of many robots. So far MRS has been studied in various fields. For a robust MRS, we focus on Reinforcement Learning to control a MRS. Reinforcement Learning works well in a static environment. Since, it is difficult for Reinforcement Learning to adopt a dynamic environment. In addition to that, in general, a designer needs to appropriately decide how the discretization level of the state and action spaces is represented. However, there is no general design guideline. To overcome these problems, we have been developing BRL as a type of Reinforcement Learning techniques. In this paper, a technique is proposed for effectively using messy knowledge acquired using BRL for improving the robustness of MRS. The technique reconstructs the state space by using Support Vector Machine based on the input-output data acquired by BRL. To investigate our proposed technique, We conduct computer simulations of a cooperative carrying task with three autonomous mobile robots.

Key Words: Multi-Robot-Systems, Distributed Coordination, Reinforcement Learning

1 はじめに

複数のロボットにより構成されるシステムとしてマルチロボットシステム (Multi-Robot Systems: MRS) という研究が進められている。多くの MRS の設計方針は、設計者が予め行動機構を組み込んでおくトップダウンアプローチをとっている [1]。このアプローチは、事前に非均質な MRS を設計し、タスク処理能力の効率化に重点を置いている。しかし、システムの大規模化などの要因により制御が困難になるということも考えられる。

この設計方針と異なる手法として、ロボット自身が適応的な振る舞いを創発するボトムアップアプローチがある [2][3]。このアプローチは、均質なロボットが状況に応じた振る舞いを発現することにより異なる役割を担うことで、システムの頑健性が期待される。このボトムアップアプローチの一つとして、オンラインで入出力関係を構築する強化学習 [4] が用いられる。

強化学習を用いた学習の成否は状態・行動空間の離散具合に大きく依存する。しかし、この設計に対する一般的な指針はないのが現状である。そこで、我々はこの状態・行動空間の設計問題に対する一手法として、強化学習を機能拡張した Bayesian-discrimination-function-based Reinforcement Learning(BRL)[5] を研究・開発してきた。

一般的な強化学習では静的な環境を前提としている。一方、BRL では連続空間における行動獲得が可能となっている。過去にこれまで、自律移動ロボット群を用いた協調搬送問題 [5] やアーム型ロボット群による協調荷上げ問題 [6] を取り上げ、MRS における BRL の有効性を示してきた。

近年、我々は、より頑健性の高い MRS の構築を目指し、BRL の獲得した知識に着目した手法を提案している [7]。この提案手法には、パターン認識手法の一つである Support Vector Machine(SVM)[8] を応用した。具体的には、BRL により獲得した入出力データを教師データとして、SVM で学習させ、状態空間の再構成を行う。

実験結果により、再構成を行った状態空間を使用する事で、行動選択の精度が従来型 BRL よりも提案手法の方が精度の高い行動選択を行う事が可能であるという結果を示している。

本稿では、過去の提案手法では手動により BRL と SVM との切り替えを行っていた機構を変更し、BRL による知識獲得と SVM による知識利用を並列に行うことが可能な手法への改良を試みる。提案手法の検証には、三台の自律移動ロボットを用いた協調搬送問題を取り上げる。そして、提案手法を従来型 BRL と比較し、性能評価を行う。

2 BRL

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 x がどの状態 C_k に分類されるかを識別する。ベイズ判別法は、識別対象のクラス $C = \{C_k\}_{k=1}^K$ および各クラスの事前確率 $P(C_k)$ と確率分布 $p(x|C_k)$ が既知の場合、入力 x が観測されたときの各クラス C_k の事後確率 $P(C_k|x)$ をベイズの公式から求め、事後確率最大となるクラスに入力を識別する方法である。

$$P(C_k|x) = \frac{P(C_k)p(x|C_k)}{\sum_{k=1}^K P(C_k)p(x|C_k)} \quad (1)$$

ただし、 $\sum_{k=1}^K P(C_k|x) = 1$, $\sum_{k=1}^K P(C_k) = 1$ である。BRL では、(1) クラスの追加と削除、(2) 確率分布モデルのパラメータ更新によって観測データから環境の確率モデルをリアルタイムに更新し、状態空間の分割を行なう。

2.1 ルール構成

各クラスをガウス分布によって表現し、各クラスの確率分布を表すパラメータとそのときの出力を if-then 形式で記述したルールとして学習器に記憶する。これ以降、クラスとルールを同義として扱う。ルール集合 R はルール $rl \in R$ により構成され、各ルールは次式で記述される。

$$rl := \langle v, \Sigma, f, u, \Phi, a \rangle \quad (2)$$

各ルール rl は状態ベクトル $v = \{v_1, \dots, v_{n_d}\}^T$, 共分散行列 Σ , クラスの事前確率 f , クラスの信頼性を表す有効度 u , 各クラスで観測されたセンサ入力集合 $\Phi = \{\phi, \dots, \phi_{n_s}\}^T$, そして、動作 $a = \{a_1, \dots, a_{n_a}\}^T$ より構成されている。ただし, n_d は入力空間の次元数, n_a は出力空間の次元数, n_s は各クラスが記憶しているサンプルデータを表す。学習初期, 状態空間にはクラスは存在せず, ロボットが実際に観測した入出力を基に状態空間にクラスを追加し, 状態空間をガウス分布で覆っていく。

2.2 動作選択

以下に BRL の行動選択の概要を示す。

- 知覚した入力を, バイズ判別法によりどの状態に属するかの判別を行う。
- 既存のルールに属さない場合, ランダム行動を出力し, 罰を受けなければ新たなルールを作成する。
- 既存のルールに属する場合, そのルール行動を出力する。

入力に対する各ルールの事後確率をバイズの公式から求め, 事後確率最大のルールに記述されている出力を実行する。ここでは, まず事後確率の負の対数を取り, 誤って識別する確率 g_i が最小となるルールを勝者ルール rl_w とする。

$$g_w = \min_i \{g_i\} \quad (3)$$

$$g_i = -\log\{f_i \cdot p(x|C_i)\} \quad (4)$$

このとき, 事後確率が非常に小さいルールが選択されるのは適切でないと考え, 事後確率にしきい値 P_{th} を設ける。そして, それを基に計算されるしきい値 $g_{th} = -\log\{f_0 \cdot P_{th}\}$ によって rl_w の動作を実行するかどうかを以下のように判断する。なお, f_0 および P_{th} は定数である。

- $g_w < g_{th}$ の場合, rl_w の動作 a_w を実行する。
- $g_w \geq g_{th}$ の場合, ランダム動作を実行する。

2.3 有効度の更新

Profit Sharing と Bucket Brigade 的戦略により報酬を過去に遡って伝播させる。その他, 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。

2.4 パラメータの更新

各ルールは入力をもとに確率分布のパラメータをオンラインで更新していく。このように, リアルタイムに更新することで環境やシステム変動に対する迅速な対応が期待できる。その反面, ノイズや一時的な入力の偏りに影響を受けやすいため何らかの対処が必要となる。BRL では, 区間推定法を用いたパラメータ更新によりこの問題を解決する。区間推定法は, 確率分布のパラメータがある区間に入る確率を設定した確率以上になるように保証する手法であり, サンプルデータが増大するにつれて推定精度が上がる。そのため, 観測データの増加に伴ってより信頼性の高いパラメータ推定が期待できる。

$$v_j \leftarrow v_j + \alpha(\bar{x}_j - v_j) \quad (5)$$

$$\sigma_j^2 \leftarrow \sigma_j^2 + \alpha^2(s_j^2 - \sigma_j^2) \quad (6)$$

$$f \leftarrow \begin{cases} f + \beta(1-f) & (\text{if } P \geq 0) \\ (1-\beta)f & (\text{otherwise}) \end{cases} \quad (7)$$

ここで, v は rl_w の平均, σ^2 は rl_w の分散, α, β は区間 $[0,1]$ の定数, j は入力ベクトルにおける各次元, $\bar{x}$ はサンプル入力の平均, s^2 はサンプル入力の分散, P は報酬である。

2.5 既存ルールのパラメータに基づく行動空間の適応的探索

常に行動空間をランダムに探索するのは非効率であるという観点から, 知識獲得がある程度行われた状況では幅広く行動空間を探索するよりも既存のルールの近傍を探索して行動の調整を行うことが有効であると考え。そこで, ランダム探索をするためのしきい値 P_{th} の他に新たにしきい値 P'_{th} を設定 ($P'_{th} < P_{th}$) し, $g_{th} \leq g_w < g'_{th}$ の場合はその間にあるルールのパラメータを参照して新しいルールパラメータを決定する。つまり, 行動選択を以下のように変更する。

- $g_w < g_{th}$ の場合, rl_w の動作 a_w を実行する。
- $g_{th} \leq g_w < g'_{th}$ の場合, この間にあるルールを基に動作を生成する。
- $g_w \geq g'_{th}$ の場合, ランダム動作を実行する。

* $g_{th} \leq g_w < g'_{th}$ の行動 この範囲には, rl_w 以外にも複数のルールが含まれる場合がある。これらのルールはその状況化での選択確率としては大きな差はないものの, それまでの学習過程におけるタスク達成への貢献度にしたがって有効度が異なる。そのため, 新しいルールの動作 a' は, この範囲に含まれるルールの有効度に基づく荷重平均により求める。

$$a' = \sum_{l=1}^{n_r} \left(\frac{u_l}{\sum_{k=1}^{n_r} u_k} \cdot a_l \right) + N(0, \sigma) \quad (8)$$

n_r はこの範囲に含まれるルール数であり, $N(0, \sigma)$ は平均 0・標準偏差 σ の正規分布を用いたノイズである。ノイズを付加することで, rl_w 以外のルールがない場合であっても, a_w の近傍を探索することができる。

3 提案手法

Fig.1 に提案手法の概要を示す。提案手法では, まず, 従来型 BRL と同様に識別関数値の計算を行う。このとき, BRL による知識獲得と SVM による知識利用を切り替えるための閾値を追加することで, BRL によるルール生成と SVM によるルール選択のいずれかを決定する。

また, ルール判別できる範囲を BRL を選択する場合よりも広範囲に覆う新たな閾値を設定する。閾値を広げる事で, SVM での識別でより多くのサンプルデータを用いる事ができるため, 精度の高いルール判別が可能となることが期待される。

さらに, BRL によりタスク達成に有効なルールを獲得できずサンプルデータが不十分な状態で, SVM による知識利用を行う事は非効率である。そこで, ロボットの行動が安定し始めたタイミングで提案手法を使用し始めるため指標を設ける。この指標を行動の安定度と呼ぶ事とする。

行動の安定度とは, ルール発火の情報エントロピーを利用した, ロボットの行動系列の安定度を表す指標である。行動の安定度は以下の式により計算される。

$$e = |E_{eps} - E_{eps-1}| \quad (9)$$

ここで、 E_{eps} は現在のエピソード終了時のエントロピー、 E_{eps-1} は前回のエピソード終了時のエントロピーを表す。ルール発火の情報エントロピーは、ロボットの行動系列が安定するほど値の変動が小さくなる特徴がある。そのため、 e の値は 0 に収束するように推移していく。

行動の安定度を利用することで、ロボットの行動が安定的であるかどうかを判断する。本稿では、 e の値に閾値を設けることで、行動選択を BRL により行うか、SVM により行うかを判定する。

以下に、提案手法の行動選択の概要を示す。

- 知覚した入力に基づき、BRL により知識探索を行うか SVM により知識利用を行うかを、 e の値と閾値との比較により決定する。
- BRL により知識探索を行う場合、ランダム行動を出力し、罰を受けなければ新たなルールを作成する。
- SVM により知識利用を行う場合、SVM により状態空間の再分割を行い、再分割した状態空間により判別されたルールの行動を出力する。

4 計算機実験

4.1 実験設定

タスクとして、三台の連結された自律移動ロボットが目標まで到達する協調搬送問題を取り扱う。実験環境を Fig.2(a) に示す。フィールドは四方を壁で囲まれている。ロボットは差動駆動型を用い、二輪の駆動輪を持つ。各ロボットは全方位カメラにより他のロボットの状態を知覚し、ロボット間で通信を行わない。ロボットの一度の意思決定を 1 ステップとし、タスクを達成するか 300 ステップ経過した時点でエピソードを更新し、ロボットをスタート地点に戻す。学習成功の定義は、規定連続エピソード (50, 100, 150 エピソードの三パターン) でタスクを達成したときとする。規定エピソードが大きいくほど、学習が進んでいる事を意味する。

また、環境変化に対するシステムの頑健性を検証するため、学習成功後、目標が左右に振幅運動を繰り返す設定となっている (Fig.2(b))。目標の移動幅が Fig.2(a) の基準位置から 1.0m, 2.0m, 3.0m の三パターン設定し、再学習までに要するエピソード数を検証する。なお、再学習の定義は、移動する目標に 10 エピソード連続で到達したときとする。タスクの成否に関わらず、

500 エピソードまで繰り返す。500 エピソード経過までを 1 試行とし、30 試行繰り返す。

比較のため、一般的な強化学習の一つである Q-Learning を連続空間へ機能拡張した手法と従来型 BRL、提案手法の三つの制御器で比較を行う。なお、シミュレーション環境はオープンソース三次元物理エンジン ODE[9] により作成している。

4.2 学習器の設定

入力は、 $I = \{r_0, \cos \theta_0, \sin \theta_0, r_1, \cos \theta_1, \sin \theta_1, \cos \theta_2, \sin \theta_2, \cos \theta_3, \sin \theta_3\}$ の 10 次元である。 r 、 θ は対象物までの距離とその角度を、添字 0, 1 は対象物がそれぞれ目標、最近傍の壁を、添字 2, 3 は左右の連結棒を示している。ただし、両端のロボットの連結されていない方の入力は 0.0 となっている。

Q-Learning の出力はロボットの左右のモータ回転速度を予め設定し、前進、右旋回、左旋回、右回転、左回転の 5 パターン用意する。BRL の出力はロボットの左右のモータ回転速度 $O = \{m_0, m_1\}$ の 2 次元である。SVM の設定は、多クラス分類が可能なライブラリ LIBSVM[10] を使用し、カーネル関数は RBF 関数を用いる。カーネル関数の γ の値は 0.1 ($=1/\text{入力の次元数}$)、コストパラメータ C の値は 10000 とする。

4.3 実験結果・考察

Q-Learning では、30 試行中一度も学習に成功する事はなかった。一方、従来型 BRL、提案手法ともに 30 試行中 30 試行で学習に成功した。

従来型 BRL と提案手法のそれぞれの制御器を用いて学習に成功した後、目標が振幅移動を行う環境において再学習を行った。30 試行における各制御器が再学習までに要したエピソード数の平均と分散を Fig.4 に示す。

■規定エピソード (50 エピソード) の場合 従来型 BRL と提案手法において再学習に要したエピソード数に大きな差は見られず、少ないエピソードで再学習ができていた。また、目標の移動が大きくなった場合においても少ないエピソードで再学習を行っている事がわかる。

Fig. 1 Proposed Method

Fig. 2 Experimental Environment

Fig. 3 Robots

■**規定エピソード (100 エピソード) の場合** 従来型 BRL においては、目標の移動が 3.0m になった場合に、再学習に 20 エピソード以上要している事が分かる。つまり、環境変化に対して時間を要している。一方、提案手法においては、目標の移動が 1.0m, 2.0m の場合 10 エピソード以下で再学習を行っている。3.0m の場合においても、従来型 BRL と比較し少ないエピソードで再学習を行っている。

■**規定エピソード (150 エピソード) の場合** 従来型 BRL では、再学習に要するエピソード数の増加量が目標の移動距離が大きくなるにつれて、Fig..4(a), Fig..4(b) と比較し、大きくなっている。提案手法では、Fig..4(b) と同様のエピソード数により再学習を行っている。

再学習に要するエピソード数の比較の結果から、従来型 BRL に比べ提案手法では、環境変化に対し頑健性が高い事が示され、BRL による知識探索と SVM による知識利用が効果的に行われていたと言える。

5 おわりに

頑健な MRS を目指し、連続な状態・行動空間を構成する強化学習法 BRL の獲得した知識に着目した手法を提案した。具体的には、BRL による知識獲得と SVM による知識利用を並列に行い、行動の安定度という指標を基に、制御器の切り替えを行う手法である。BRL により収集した知識である入力データを教師データとして、SVM で学習させ、状態空間の再構成を行う事で、知識を利用した。

三台の自律移動型ロボットを用いた協調搬送問題において、行動獲得後に環境変化があった場合、従来型の BRL では、再学習するまでのエピソード数が大きかった。一方、提案手法での再学習するまでに要するエピソード数は、従来型の BRL に比べ小さい結果となった。このことから、提案手法では、環境変化に対し頑健性が高く、BRL による知識探索と SVM による知識利用が効果的であったと示された。

今後は、提案手法をより大規模なマルチロボットシステムへ適用することや、実機ロボットを用いた実験での検証を行う予定である。

文 献

- [1] C. Kube and H. Zhang: The Use of Perceptual Cues in Multi-Robot Box-Pushing, IEEE International Conference on Robotics and Automation, pp.2085-2090, 1997.
- [2] Y. Takahashi, Y. Tamura, M. Asada, and M. Negrello: Emulation and behavior understanding through shared values, IEEE International Conference on Intelligent Robots and Systems, pp.3950-3955, 2007.
- [3] G. Baldassarre, V. Trianni, M. Bonani, and F. Mondada: Self-Organized Coordinated Motion in Groups of Physically Connected Robots, IEEE Transactions on Systems, Man, and Cybernetics, Part. B, Vol. 37, pp.224-239, 2007.
- [4] R. Sutton and A. Barto: Reinforcement Learning: An Introduction, MIT Press, 1998.
- [5] T. Yasuda and K. Ohkura : Improving Search Efficiency in the Action Space of an Instance-Based Reinforcement Learning, Advances in Artificial Life, the 9th European Conf. on Artificial Life, LNAI, Vol. 4648, pp. 325-334, 2007.
- [6] T. Yasuda, S. Sakimoto, and K. Ohkura: Analyzing the Process of Autonomous Specialization in Reinforcement Learning Robots, Proceedings of the 3rd International Symposium on Mobiligence, pp.143-148, 2009.

(a) Defined Successive Episodes: 50

(b) Defined Successive Episodes: 100

(c) Defined Successive Episodes: 150

Fig. 4 Number of Episodes for Re-Learning

- [7] J. Sakanoue, T. Yasuda, and K. Ohkura: Preservation and Application of Acquired Knowledge Using Instance-Based Reinforcement Learning, Joint 5th International Conference on Soft Computing and Intelligent Systems and 10th International Symposium on advanced Intelligent Systems, pp.576-581, 2010.
- [8] C. Cortes and V. Vapnik: Support vector networks, Machine Learning, Vol.20, 273-297, 1995.
- [9] Open DynamicEngine, <http://www.ode.org/ode.html>
- [10] LIBSVM, <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>