移動ロボットにおける状態空間の再構成を可能とする報酬分配法

A Reward Distribution Method Which Enables a Mobile Robot to Reconstruct a State Space

河村竜幸 †, 上野敦志 †, 武田英明 † ††
Tatsuyuki Kawamura, Atsushi Ueno, Hideaki Takeda

奈良先端科学技術大学院大学 情報科学研究科 †
Graduate School of Information Science, Nara Institute of Science and Technology

国立情報学研究所 ††
The National Institute of Informatics

Abstract: In this paper, we propose a new learning method for a real mobile robot. Reinforcement learning can solve problems without human advice. Reinforcement learning must have a discrete state-space. However, it is difficult to calculate the optimum state-space before learning in the continuous real space. One of solving methods for optimizing state-space is to reconstruct a more fitting it in the act of learning. Traditional reinforcement learning can not available to reconstruct a state-space, because different particle size of state-spaces are irreconcilable spaces each other. To facilitate reconstruction, we propose a new algorithm, which can recycle past Q-values after reconstructing to a new state-space. A discount reward distribution with transition time can derive an evaluation function, which dose not depend on particle size of state-spaces. We show that a mobile robot can learn effectively using a state-space reconstruction.

1 はじめに

近年, 移動ロボットに強化学習を適用する研究が増加している [?]. 強化学習に分類されるQ学習 [?] もまた移動ロボットの研究に利用されている.

Q学習は観測可能な環境の状態の組 (状態空間) とロボットの選択可能な行動の組 (行動空間) を利用して学習を行う. 状態空間は有限離散化された空間である. 学習する環境が離散的である場合, 環境と状態空間の最適な関係が決定する. しかし, 学習する環境が連続した場合には, 環境と状態空間の最適な関係を求めることが困難となる.

解決法としては, 学習を行いながら環境に対して最適な状態空間を推測し, 再構成していく方法がある. 通常のQ学習では粒度の異なる状態空間は互いに関係付されていないため, 状態空間の再構成を行っても学習データは引き継がれない. これまでは, 状態空間が適切に分割されていない場合に, 人間が何度も状態空間を再構成して実験をやり直なければならないという問題があった.

本研究ではこのような問題を解決するため, 状態遷移にかかる時間 (遷移時間) を考慮した報酬分配法を提案する. また, 実験によって学習途中に状態空間を再構成しても学習がなめらかに継続されることを示す.

[0]連絡先：河村 竜幸
〒 630-0101 生駒市高山町 8916-5
Tel (0743)72-5265 Fax (0743)72-5269
mail: tatsu-k@is.aist-nara.ac.jp
http://ai-www.aist-nara.ac.jp/

2 Q学習

強化学習は, ある知覚入力とそのときに選択した行動の結果から得られる報酬に基づいて学習を行う手法である.

Q学習は状態 s と行動 a の組（s,a）に対して行動価値関数 Q（s,a）が定義される. a は選択可能な行動の組 A に属する. 行動 a の選択により状態 s から状態 s' へと遷移する. 以下の式で状態が遷移するごとにQ値を更新する。

$$Q(s,a) \leftarrow Q(s,a) + \alpha(r + \gamma \max_{a' \in \mathbf{A}} Q(s',a') - Q(s,a)) \quad (1)$$

ここで、 r は報酬,α（$0 < \alpha < 1$）は学習率であり、γ（$0 < \gamma < 1$）は減衰係数である.

3 遷移時間を考慮した報酬分配法

一般的なQ学習で, 報酬分配は状態を遷移するごとに一定の割合ずつ減衰するように行われる. 異なる状態空間の間で同じ知覚入力に対するQ値を比較した場合, 目標状態までの状態遷移回数に違いが生じるため,2 つのQ値は同じにならない. つまり, 学習途中に状態空間を再構成すると, 再構成前後の状態空間に対し期待されるQ値の差が生じるため, 状態空間の再構成以前の学習データが引き継がれないことになる.

本手法では, 目標状態からの時間的距離によって割り引かれる値を決定する, すなわち, 状態が変化するまでにかかる時間によって分配する報酬の割合を決定する手

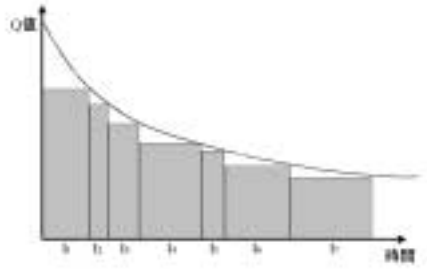

図 1: 遷移時間による報酬分配

法を採用する (図**??**). この手法によって, 同じ知覚入力に対して期待される報酬は状態空間の再構成前後においても変化がなくなる.

この手法では, 実際の経過時間を用いるので, ロボットが最適経路を通らずに寄り道をした場合などには, 報酬が適切に分配されない. しかし, 学習が進み適切な行動が獲得されるにつれて, 次第に適切なQ値に収束することが期待できる.

報酬を遷移時間によって分配する手法は通常のQ学習の減衰係数 γ を式 (2) へと時間の関数とすることで実現できる.

$$\gamma(t) = exp(-t/T) \tag{2}$$

ここで,T は遷移時間定数とする.

4 状態空間の再構成

学習時における状態空間の再構成は以下の要領で行う.

1) 再構成する状態空間の領域を選択する.

2) 選択された領域内の空間を全ての軸に対して 2 等分する.

3) 分割前の状態のQ値を分割してできた新たな状態に継承する.

分割前に粗く近似されたQ値は状態空間を分割することで, より適切な値へと修正が可能となる.

5 実験

学習途中に状態空間を再構成しても, 学習がなめらかに継続することを示す実験を行った.

タスクとして、ボールをゴールにシュートするサッカーロボットのコンピュータシミュレーションを行う. ロボットが得られる情報は、ロボットに搭載された全方位視覚カメラからの画像情報のみである. ロボットはボールに対して距離と角度、ゴールに対して幅、距離と角度を知覚する.

実験では 2 つの状態空間を用いた. 粗い状態空間 (以後,4 分割) ではボールに対して距離方向に 4 等分、角度方向に 4 等分する. ゴールに対しては距離方向に 4 等分、角度方向に 4 等分, 幅方向に 4 等分を等分する. 細かい状態空間 (以後,16 分割) ではそれぞれ 16 等分する. ロボットとボールの初期配置は図 **??**に示す領域内にランダムに配置する.

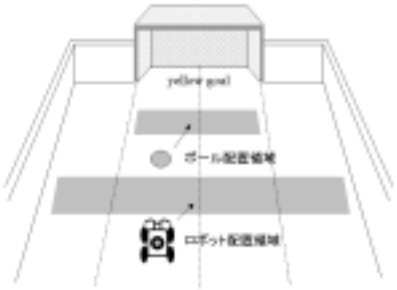

図 2: ロボットとボールの初期配置

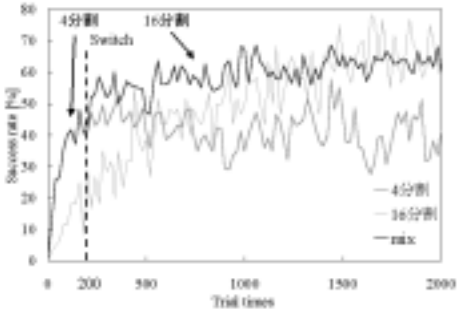

図 3: 状態空間の再構成によるなめらかな学習の継続

実験は 3 種類行った. 初めに 4 分割のみの学習, 次に 16 分割のみの学習, 最後は 4 分割の状態空間から学習を始め, 試行回数が 200 回に達したところで状態空間を 16 分割に再構成. それぞれについて 2000 試行の実験を各 10 回ずつ行った.

結果は図 **??**に示す. 4 分割のみの学習では学習曲線の立上りは早いが, シュートの成功率の収束値が最も低くなった. 16 分割のみの学習では最終的なシュート成功率は高くなったが, 学習の立上りが最も遅かった. 状態空間を途中で再構成した学習は 4 分割から 16 分割の学習へとなめらかに移行していることがわかる.

6 まとめ

今回は移動ロボットの学習において状態の遷移時間を考慮することで, 状態空間の再構成が可能な報酬分配手法を提案した.

実験ではQ学習アルゴリズムを非常に単純な方法で拡張することで, 状態空間の再構成時に対しても連続して効率的に学習が行えることが示せた. 本手法は局所的な状態空間の再構成に対しても適応可能である. 今後の展望として, 状態空間中で分割すべき部分を特定する手法と組み合わせることで, 行動の学習を行いながら, 最適な状態空間の構成を行うことが可能であると期待できる.

参考文献

[1] J. H. Connel, S. Mahadevan, editors..: “Robot Learning.”, Kluwer Academic Publishers, 1993.

[2] Leslie Pack Kaelbling, Michael L. Littman, Andrew W.Moore: “Reinforcement Learning: A Survey”, Journal of Artificial Intelligence Research, Vol.4,pp.237-285,Mar. 1996.