1M-04

自己組織化特徴マップに基づいた時系列パターンのための確率的連想メモリによる強化学習を用いたロボットの行動学習

水野暁翔　長名優子

東京工科大学 コンピュータサイエンス学部

1 はじめに

部分観測マルコフ決定過程 (POMDPs : Partially Observable Markov Decision Processes) 環境においても決定的な政策を学習することのできる手法として，POMDPs 環境のための決定的政策を学習する Profit Sharing[1] が提案されている．この手法では，過去の観測の系列を用いることで，同じ観測に対して複数の行動をとる必要がある場合にも適切な行動を選択することができる．また，ニューラルネットワークを用いた強化学習に関する様々な研究も行われており，そのような手法のひとつとして自己組織化マップに基づいた時系列パターンのための確率的連想メモリによる強化学習 [2] が提案されている．

本研究では，自己組織化マップに基づいた時系列パターンのための確率的連想メモリによる強化学習 [2] を用い，ロボットの行動学習を実現する．これにより，POMDPs 環境下におけるロボットの行動規則の獲得を目的としている．

2 POMDPs 環境のための決定的政策を学習する Profit Sharing

従来の POMDPs 環境のための報酬獲得効率に基づく強化学習法 [3] やその他の Profit Sharing[1] をベースとした手法では，現在の観測 o に関するルール (o, a) の価値のみを行動選択の基準として用いている．本研究で用いる POMDPs 環境のための決定的政策を学習する Profit Sharing[1] では，現在の観測 o に対応するルールの価値 (o, a) だけでなく，エージェントがそのエピソード中で観測した過去の観測の系列を考慮して行動の決定を行うことで，より適切な行動の選択を実現する．

この手法では，ルール (o, a) に対して過去の観測 o_p がどれだけ影響を及ぼすかを表す値 $\gamma(o_p \to o, a)$ を定義し，ルールの価値を求める際に利用する．

時刻 x において観測 o_x が知覚されたときにどの行動をとるかは過去の行動系列を考慮したルールの価値 $\omega'(o_x, a)$ に基づき，ボルツマン選択により決定する．過去の行動系列を考慮したルールの価値 $\omega'(o_x, a)$ は，

$$\omega'(o_x,a)=\begin{cases}\omega(o_x,a), & x=1\\ \beta\dfrac{\sum_{i=h(x)}^{x-1}\gamma(o_i\to o_x,a)}{x-h(x)}+(1-\beta)\omega(o_x,a), & \text{それ以外}\end{cases} \quad (1)$$

のように計算する．ここで，$\gamma(o_i \to o_x, a)$ は過去の観測 o_i がルール (o_x, a) に及ぼす影響力，$\beta(0 < \beta < 1)$ は過去の観測の影響の度合いを決める係数を表している．$h(x)$ は時刻 x より前で最後に観測 o_x が知覚された時刻であり，

$$h(x)=\begin{cases}1, & \min_i\{i|o_i=o_x\}=x\\ \max_i\{i|o_i=o_x, i<x\}, & \text{それ以外}\end{cases} \quad (2)$$

で与えられる．

この手法では，報酬を獲得したときにルールの価値 $\omega(o, a)$ と過去の観測がルールに与える影響力 $\gamma(o_y \to o_x, a_x)$ を更新する．ルールの価値 $\omega(o, a)$ の更新は POMDPs 環境のための報酬獲得効率に基づく強化学習法 [3] と同様に行う．

1. $(o_y \to o_x)$ という観測の組み合わせがエピソード中にはじめて現れたとき，$(o_y \to o_x)$ でのルールのための報酬分配 $F(o_y \to o_x)$ を以下のように決定する．

$$F(o_y \to o_x) = 1/(W - x_{o_x}) \quad (3)$$

ここで，y は $h(x) \le y < x$ を満たす時刻である．観測 o_y が時刻 $h(x) \sim x$ の間に複数回観測される場合には，y は条件を満たすもののうち最大の値をとるものとする．また，x_{o_x} は $(o_y \to o_x)$ がはじめて現れたときの時刻を表す．

2. $(o_y \to o_x, a_x)$ がエピソード中にはじめて現れたとき，強化回数 $I(o_y \to o_x, a_x)$ と影響力 $\gamma(o_y \to$

Action Learning of Robot by Reinforcement Learning using Self-Organizing Map-based Probabilistic Associative Memory for Sequential Patterns
Akito Mizuno and Yuko Osana (Tokyo University of Technology, osana@stf.teu.ac.jp)

Copyright ©2017 Information Processing Society of Japan.
All Rights Reserved.

$o_x, a_x)$ を以下のように更新する．

$$I(o_y \to o_x, a_x) \leftarrow I(o_y \to o_x, a_x) + 1 \quad (4)$$

$$\gamma(o_y \to o_x, a_x) \leftarrow \left(1 - \frac{1}{I(o_y \to o_x, a_x)}\right) \times \gamma(o_y \to o_x, a_x) + \frac{R \times F(o_y \to o_x)}{I(o_y \to o_x, a_x)} \quad (5)$$

3. 2 度目以降に同じ組み合わせ $(o_y \to o_x, a_x)$ が現れたときは，更新を行わない．よってエピソード中の各影響力は，それぞれ 1 度だけ更新される．

この手法ではエピソード中に同じ観測が複数回知覚されるような場合には最後に同じ観測が知覚された時刻からのエピソードの部分系列をルールに影響する系列として学習を行うことになる．あるルール (o_x, a_x) が適切であれば，影響力 $\gamma(o_y \to o_x, a_x)$ への報酬の分配量も多くなることが期待できる．

3 自己組織化特徴マップに基づいた時系列パターンのための確率的連想メモリによる強化学習

自己組織化特徴マップに基づいた時系列パターンのための確率的連想メモリによる強化学習 [2] では，POMDPs 環境のための決定的政策を学習する Profit Sharing[1] におけるルールの価値の学習を，自己組織化特徴マップに基づいた時系列パターンのための確率的連想メモリによって実現している．この手法では行動の決定に利用する観測の系列と行動の対 (ルール) を学習する．また，信頼度はルールの価値に応じて変更し，確率的な行動の選択を可能としている．

4 自己組織化特徴マップに基づいた時系列パターンのための確率的連想メモリによる強化学習を用いたロボットの行動学習

本研究では，ロボットの行動学習に自己組織化特徴マップに基づいた時系列パターンのための確率的連想メモリによる強化学習 [2] を用いることで，POMDPs 環境における行動規則の獲得を行う．

ロボットは，スタート地点から移動して，鍵を入手した上でゴール地点に到達することを目的として学習を行う．ロボットは観測として壁までの距離を受け取り，その観測に基づいて，(1) 前進，(2) 右を向く，(3) 左を向くの 3 種類の行動の中から，ルールの価値に基づいて行動を選択し，学習を進めるものとする．また，ロボットに装着された小型カメラを用いて撮影した画像の RGB 値を測定することで鍵の獲得判定を行う．各ステップにおいて一つの行動を行うものとし，

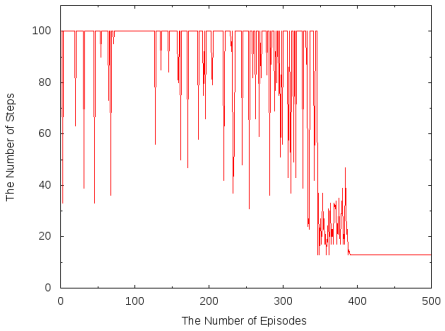

図 1: ステップ数の推移

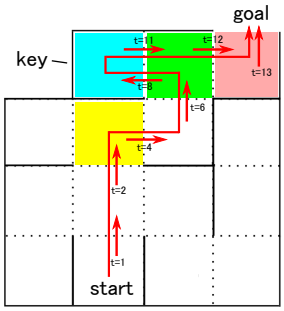

図 2: 500 回目の試行における移動経路

スタート地点からゴール地点到達までを 1 エピソードとする．各エピソードにおいて，ゴール地点到達までにかかったステップ数が少ないほど，大きな報酬を与える．また，総ステップ数が 100 以上になった場合，報酬を与えずに環境を初期化するものとする．

5 実験

ロボットを用いた実験，およびシミュレーションによる実験を行った．図 1 にシミュレーションにおける試行ごとのステップ数の推移を示す．また，500 試行目におけるロボットの移動経路を示す．学習が行われ，最短ステップ数でゴールまで到達できていることが分かる．

参考文献

[1] Y. Takamori and Y. Osana : “Profit sharing that can learn deterministic policy for POMDPs environments,” Proceedings of IEEE International Conference on System, Man and Cybernetics, Anchorage, 2011.

[2] 新妻純，長名優子：“自己組織化特徴マップに基づいた時系列パターンのための確率的連想メモリによる強化学習の実現,” 情報処理学会第 77 回全国大会, 2015.

[3] 河合宏和, 上野敦志, 辰巳昭治：“POMDPs 環境のための報酬獲得効率に基づく強化学習法,” 人工知能学会論文誌, Vol.23, No.1, pp.1–12, 2008.

Copyright ©2017 Information Processing Society of Japan.
All Rights Reserved.