5Q-4

POMDPs環境下での過去の観測系列を用いた学習

高森洋平　長名優子

東京工科大学大学院 バイオ・情報メディア研究科コンピュータサイエンス専攻

1　はじめに

教師信号を用いずに環境との相互作用により適切な行動系列を獲得するための学習手法として，強化学習に関する様々な研究が行われている．部分観測可能マルコフ決定過程 (Partially Observable Markov Decision Processes：POMDPs) 環境における強化学習に関する研究も盛んに行われており，そのような手法の1つとしてPOMDPs環境のための報酬獲得効率に基づく強化学習法 [1] が提案されている．ここで，POMDPs環境とはエージェントの知覚の制限のために実際の状態とは異なるような観測がされてしまう可能性があるような環境をさす．POMDPs環境のための報酬獲得効率に基づく強化学習法では同一観測からの報酬獲得に必要な行動の選択確率が等しくなるように学習が行われる．

本研究では，POMDPs環境のための報酬獲得効率に基づく強化学習法を拡張し，POMDPs環境において決定的な政策を獲得できるような手法を提案する．提案手法では決定的な行動選択をするために過去の観測の履歴を利用する．ある時刻において過去の観測の履歴が行動の決定に必要かどうかの判断は不完全知覚判定法を導入したProfit Sharing[2] の不完全知覚判定方法に基づいた方法を利用して行う．

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

提案手法は，POMDPs環境のための報酬獲得効率に基づく強化学習法 [1] に基づいた手法であり，不完全知覚の生じている観測上において，報酬獲得に必要な行動の選択確率が等しくなるように学習が行われる．学習がある程度進行した段階で，行動の選択が決定的に行われているかどうかを行動の決定度 [2] を用いて判断し，行動の選択が確率的に行われているようであれば，それ以降のエピソードにおいてより過去の観測の情報を考慮した行動の選択を行えるようにする．

2.1　行動選択

提案手法ではエージェントが時刻 x の観測 o_x において行動 a を選択する確率 $P(o_x, a, x)$ は以下のように与えられる．

$$P(o_x,a,x)=\begin{cases}\dfrac{\exp(q_n(o_x,a)/T(o_x))}{\sum_{b\in C^A}\exp(q_n(o_x,b)/T(o_x))}, \\ \qquad o_x \notin C^{PO} \text{のとき} \\ \dfrac{\exp(q_n(o_x,a)/T(o_x))+Q(o_x,a,x)}{\sum_{b\in C^A}(\exp(q_n(o_x,b)/T(o_x))+Q(o_x,b,x))}, \\ \qquad o_x \in C^{PO} \text{のとき}\end{cases} \quad (1)$$

ここで，$q_n(o_x, a)$ は観測ごとのルールの価値の和が1になるように正規化したルール (o_x, a) の価値，$T(o_x)$ は観測 o_x における温度パラメータ，C^A はエージェントの取り得る行動の集合，C^{PO} は不完全知覚状態であると判断されている観測または観測の系列の集合である．また $Q(o_x, a, x)$ は時刻 x からさかのぼって観測の履歴を考慮した観測 o_x において行動 a をとるというルールの価値であり，以下の式で与えられる．

$$Q(o_x,a,x)=\sum_{o_i\to O\in C^{ref}(x)}\exp(q_n(o_i\to O,a)/T(O)) \quad (2)$$

ここで，O は $(o_3 \to o_5 \to o_6)$ のような観測の系列である．また，$C^{ref}(x)$ は時刻 x での行動選択の際に考慮される観測または観測の系列の集合，$q_n(o_i \to O, a)$ は正規化したルール $(o_i \to O, a)$ の価値，$T(O)$ は観測の系列 O における温度パラメータである．提案手法では時刻 x での観測 o_x が不完全知覚状態であると判断されている場合には過去の履歴を含めた観測の系列に関するルールの価値を考慮して行動を決定する．

2.2　不完全知覚状態の判定

提案手法では時刻 x の観測 o_x において行動を決定的に選択するために十分に過去の観測の履歴が考慮されているかを，行動の決定度 $d(o_x, x)$ と学習の進行度

Learning by Profit Sharing using History in POMDPs
Yohei Takamori and Yuko Osana (Tokyo University of Technology takamori@osn.cs.teu.ac.jp, osana@cs.teu.ac.jp)

$l(o_x, x)$ から以下の式で判断する．

$$\phi(l(o_x, x))(1 - d(o_x, x)) > \theta^{PO} \tag{3}$$

ここで $\phi(\cdot)$ はシグモイド関数，θ^{PO} は判定のためのしきい値である．時刻 x の観測 o_x において行動を決定的に選択するのに十分な情報が得られていないと判断された場合には，不完全知覚状態であると判断されている観測または観測の系列の集合 C^{PO} に新たな観測の系列を追加し，以降のエピソードでより過去の観測を考慮した行動の選択を行えるようにする．

2.3 学習

提案手法ではエージェントが報酬を獲得したときにルールの価値の更新を行う．観測に関するルール (o, a) については，POMDPs 環境のための報酬獲得効率に基づく強化学習法と同様にルールの価値を更新する．また，時刻 x における観測 o_x が不完全知覚状態であると判断されているとき，時刻 x に観測 o_x から過去にさかのぼって考慮される観測の系列に関するルールの強化回数 $I(O \to o_x, a_x)$ と価値 $q(O \to o_x, a_x)$ を以下のように更新する．

$$I(O \to o_x, a_x) \leftarrow I(O \to o_x, a_x) + 1 \tag{4}$$

$$q(O \to o_x, a_x) \leftarrow \left(1 - \frac{1}{I(O \to o_x, a_x)}\right) q(O \to o_x, a_x) + \frac{r \cdot F(o_x)}{I(O \to o_x, a_x)} \tag{5}$$

ここで，r は報酬，$F(o_x)$ は観測 o_x におけるルールへの報酬分配である．ただし，観測の系列 $O \to o_x$ が複数の時刻において行動の決定に関わっている可能性があると考えられる場合には時刻 x における行動 a_x に関するルールの価値だけでなく，以下の条件を満たす行動 a_y に関するルールをすべて更新する．

$$\underset{x=1,\cdots,W}{\operatorname{argmin}}\{(O \to o_x) \in (C^{ref}(x) \cap C^{ref}(x'))|x \neq x'\} \leq y \quad \text{かつ } o_x = o_y \tag{6}$$

観測の系列 $O \to o_x$ がエピソード中に複数回出現している場合にはエピソード中に $O \to o_x$ という系列が出現した以降に存在する複数の o_x という観測においてとった行動のうちどれが報酬獲得に必要であるかを判断できないため，提案手法では複数の o_x での行動に関するルールの価値を等しく更新するという方法をとっている．

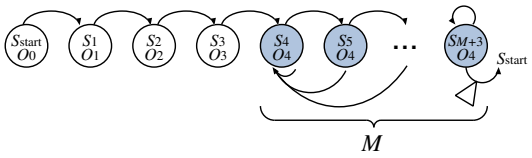

図 1: 実験に用いた POMDPs 環境

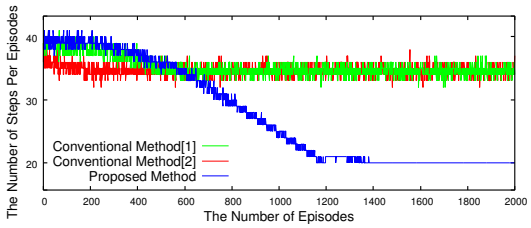

図 2: ステップ数の遷移

3 計算機実験

提案手法の有効性を確認するために図 1 のような POMDPs 環境において実験を行った．エージェントは各状態において上下どちらかの行動を選択することができ，図 1 の矢印で示された方向に行動することで状態が遷移する．なお，矢印で示されていない方向への行動をとった場合には同じ状態にとどまることになる．この例では $s_4 \sim s_{M+3}$ までが不完全知覚状態となっており，状態 s_{M+3} でとるべき行動のみが他の観測でとるべき行動と異なっている．そのため，エージェントが最短経路を選択するためには観測 s_{M+3} において過去に $M-1$ 回 o_4 を観測したという情報を利用しなければならない．ここでは，$M = 16$ として実験を行った．

図 2 に提案手法，POMDPs 環境のための報酬獲得効率に基づく強化学習法 [1]，不完全知覚判定法を導入した Profit Sharing[2] におけるステップ数の遷移を示す．図 2 において，横軸はエピソード数，縦軸は報酬を獲得するまでのステップ数の 100 回の試行の平均である．図 2 より，提案手法が最終的に決定的な政策を学習することで最短ステップ数で報酬を獲得し，従来の手法に比べて効率よく報酬を獲得できるように学習が行えていることが分かる．

参考文献

[1] 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.

[2] 斉藤健, 増田士郎：“不完全知覚判定法を導入した Profit Sharing,” 人工知能学会論文誌, Vol.19, No.5, pp.379–388, 2004.