過去情報を用いたProfit Sharingによる学習

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．環境の状態遷移が決定的であるような問題での部分観測可能マルコフ決定過程 (POMDPs) 環境における強化学習の手法としてEpisode-based Profit Sharing (EPS)[1] が提案されている．ここで，部分観測可能マルコフ決定過程 (POMDPs) 環境とは，エージェントの知覚能力が制限されているために，実際には異なる状態が同じ状態であるとエージェントに認識されたり，同じ環境であるにも関わらず別の状態であると認識されてしまう場合が存在するような環境をさす．EPSでは同じ観測に対して複数の行動をとる必要があるような場合には，すべての行動が同じ確率で選択されるように学習が行われる．

本研究では，EPSを拡張し，過去の観測と行動の系列を用いて学習を行うことで，同じ観測に対して複数の行動をとる必要があるような場合にもより適切な確率で行動の選択が行えるような手法を提案する．

2 Episode-based Profit Sharing

Episode-based Profit Sharing (EPS)[1] は，Profit Sharing[2] に基づいた手法であり，POMDPs環境においてエピソードの長さを考慮しつつ，ループ系列に含まれるルールの価値の強化の抑制を行うように報酬を分配する方法として提案されている．ここで，エピソードとはエージェントが行動を開始してから報酬が得られるまでの間をさす．また，ルールは知覚した観測 o とそのときにとる行動 a の対で (o,a) のように表される．

EPSでは，エージェントが報酬を獲得したとき，以下のようにルールの価値を更新する．

$$\omega(o_x,a_x) \leftarrow \omega(o_x,a_x) + r \cdot f_x \qquad (x = 1,\cdots,W) \quad (1)$$

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

ここで，$\omega(o_x,a_x)$ は報酬を獲得した時刻から x ステップ前に用いられたルール (o_x,a_x) に対する価値，r は報酬の値，W はエピソード (行動系列) の長さを表す．また，f_x は報酬からルール (o_x,a_x) に対する報酬の分配量を決める強化関数であり，

$$f_x = \begin{cases} \sum_{k=x}^{W} \frac{1}{L^k} d_k^{o_x}, & x = \min_i\{i | o_i = o_x \text{かつ } a_i = a_o\} \\ 0, & \text{それ以外} \end{cases} \quad (2)$$

で与えられる．ここで，L はとりうる行動の数を表す．また，$d_k^{o_x}$ は，

$$d_k^{o_x} = \begin{cases} 0, & t_{o_x}^{last} \le k < t_{o_x}^{first} \\ 1, & \text{それ以外} \end{cases} \quad (3)$$

で与えられる．ここで，$t_{o_x}^{first}$, $t_{o_x}^{last}$ は観測 o_x がエピソード中で知覚された最初と最後の時刻を表す．

EPSでは，このようにルールの価値の更新を行うことで，報酬の獲得に必要なルールの価値が不要なルールの価値を上回るように更新され，最終的にエージェントが報酬獲得に必要なルールのみを選択することが可能になる．

3 過去情報を用いたEpisode-based Profit Sharing

従来のEPSや，その他のPS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．提案する過去情報を用いたEpisode-based Profit Sharing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 o_p とそのときにとった行動 a_p の組み合わせ (o_p,a_p) 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．

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

提案手法では，時刻tにおいて観測o_tが知覚されたとき，どの行動をとるかは過去の情報を考慮したルールの価値$\omega'(o_t,a)$に基づき，ルーレット選択により決定する．過去の情報を考慮したルールの価値$\omega'(o_t,a)$は，

$$\omega'(o_t,a)=\begin{cases}\omega(o_t,a), & t=1\\ \sum_{i=t_0}^{t-1}\gamma(o_i,a_i\rightarrow o_t,a)+\omega(o_t,a), & \text{それ以外}\end{cases} \tag{4}$$

のように計算する．ここで，$\gamma(o_p,a_p\rightarrow o_t,a)$は過去の観測と行動の対$(o_p,a_p)$がルール$(o_t,a)$に及ぼす影響の大きさを表している．ここで，$t_0$は時刻$t$より前で最後に観測$o_t$が知覚された時刻であり，

$$t_0=\begin{cases}1, & t=\min_i\{i|o_i=o_t\}\\ \max_i\{i|o_i=o_t,i<t\}, & \text{それ以外}\end{cases} \tag{5}$$

で与えられる．

提案手法では，報酬を獲得したときにルールの価値$\omega(o,a)$と過去の系列が与える影響の大きさ$\gamma(o_y,a_y\rightarrow o_x,a_x)$を更新する．ルールの価値$\omega(o,a)$の更新はEPSと同様、式(1)により行う．また，過去の系列が与える影響の大きさは以下のように更新する．

$$\gamma(o_y,a_y\rightarrow o_x,a_x)\leftarrow\begin{cases}\gamma(o_y,a_y\rightarrow o_x,a_x)+r\cdot f_x, & \\ \quad x<y\leq x' \text{かつ} & \\ \quad x=\min_i\{i|o_i=o_x,a_i=a_x\} & \\ 0, \quad \text{それ以外} & \end{cases} \tag{6}$$

ここで，$\gamma(o_y,a_y\rightarrow o_x,a_x)$は報酬を得る$y$ステップ前に用いられたルール$(o_y,a_y)$が報酬を得る$x$ステップ前に用いられたルール$(o_x,a_x)$に与える影響を表す．また，$x'$は$x$ステップより前で最後に観測$o_x$が知覚された時刻であり，

$$x'=\begin{cases}W, & x=\max_i\{i|o_i=o_x\}\\ \min_i\{i|o_i=o_x,i>x\}, & \text{それ以外}\end{cases} \tag{7}$$

で与えられる．ここで，Wはエピソード(行動系列)の長さを表す．

提案手法では，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,a)が適切な行動であれば報酬の分配量が多くなり，$\gamma(o_p,a_p\rightarrow o,a)$の更新量も多くなることが期待でき，過去の情報を考慮したルールの価値$\omega'(o,a)$を用いることでより適切な行動選択が行えるようになる．

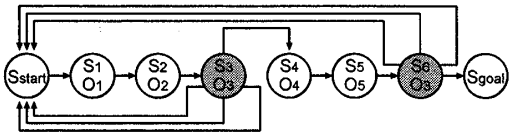

図1: 実験に用いたPOMDPs環境

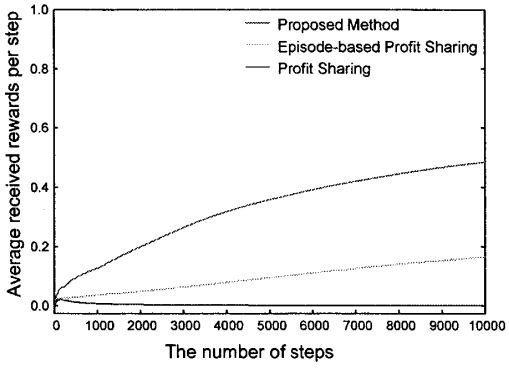

図2: 報酬の獲得効率

4 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左右のそれぞれ4方向に動くことができ，図1で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_6で右へ移動したときのみにゴール状態に到達し，報酬が得られる．また，報酬の値は7とした．このPOMDPs環境では状態の一部が不完全知覚によりs_3とs_6が同一の観測o_3であると認識される．エージェントは観測o_3で，状況によって上か右かを適切に選択する必要がある．

図2に提案手法，EPS[1]，PS[2]における報酬の獲得効率を示す．図2において，横軸は全エピソードを通しての総ステップ数，縦軸は全エピソードを通しての獲得報酬量の合計を総ステップ数で割ったものを表している．図2より，提案手法が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ることが分かる．

参考文献

[1] 植村渉, 上野敦志, 辰巳昭治：“POMDPs環境のためのエピソード強化型強化学習法,” 電子情報通信学会論文誌, A Vol.J88-A, No.6, pp.761–774, 2005.

[2] J. J. Grefenstette : “Credit assignment in rule discovery systems based on genetic algorithms,” Machine Learning, Vol.3, pp225–245, 1998.