5Q-6

強化学習における政策再利用転移学習

吉田　慎二†　　長谷川　修‡

東京工業大学大学院知能システム科学専攻†　　東京工業大学像情報工学研究所‡

1. はじめに

強化学習は，未知の環境でロボットなどの行動を獲得するために用いられる．しかし，強化学習による最適な政策の学習には，膨大な時間が要求される．この問題に対し，過去に扱ったタスクで獲得した知識を，過去のタスクと類似する新たなタスクで利用する，転移学習の様々な手法が提案され成果を上げている．例として，過去のタスクから学習した行動価値関数で目標タスクの行動価値関数を初期化する手法[1][2]や，学習した政策を目標タスクでの探査戦略に再利用する手法[3]が挙げられる．

本研究では，過去のタスクで学習した政策と行動価値関数の両方を，目標タスクでの探査戦略に利用する手法を提案する．Fernández ら[3]と同様に，過去のタスクから学習した政策を，新たなタスクを探査する際のバイアスとして用いる．エージェントは「現在学習中の政策に従い行動する」「ランダムに探査する」「過去の政策に従い行動する」の３つのいずれかに従って行動を決定し，学習を行う．提案手法ではこの過去の政策の重みを，過去のタスクにおける行動価値関数を利用し制御する．

提案手法の評価のために，Fernández ら[3]との比較実験を行った．実験により，提案手法の学習効率が従来手法[3]よりも上回ることを示した．

2. 強化学習の枠組みと転移学習

強化学習は，基本的なモデルではマルコフ決定過程（MDP）として定式化される．MDP は $< S, A, T, R >$のタプルで表され，S は状態の集合，A は行動の集合，T は確率的な状態遷移関数，R は報酬関数を表す．強化学習の目的は，期待収益を最大化するような政策 π^* を学習することである．

上記の定式化は，ある一つのタスクに対するものである．本研究で扱うすべてのタスクは同一の状態集合，行動集合，状態遷移関数に従い，タスクの違いは報酬関数の違いによって表される．また，本研究が扱うタスクはエピソード的である．

Policy Reuse Transfer Learning in Reinforcement Learning

† Shinji Yoshida, Department of Computational Intelligence and System Science, Tokyo Institute of Technology

‡ Osamu Hasegawa, Imaging Science and Engineering Laboratory, Tokyo Institute of Technology

3. 提案手法

3.1. 探査戦略

本研究では過去のタスクで学習した政策を，新たなタスクの探査戦略に利用する．エージェントは，Q-Learning によって新たなタスクの政策を学習する．探査戦略として，過去の政策に従って新たな環境を探査するか，行動価値関数に従って行動するか，ランダムに探査するかを，毎ステップごとに確率的に選択する．これら３つのバランスを適切に取ることにより，類似したタスクの政策を利用した場合には短時間で終端状態にたどり着き，類似しないタスクの政策を利用した場合には過去の政策に固執せずに探査を行うことが重要である．

3.2. 行動価値観数の利用

探査戦略のバランスを取るため，過去のタスクで学習した行動価値関数を利用する．行動価値観数はエピソード的タスクにおいて，現在状態が終端状態にどれほど近いかを表す指標になりうる．このため，過去の政策に加え，政策に従うような行動に対応する行動価値関数

$$V_\pi(x) = Q(s, \pi(s)) \qquad (1)$$

をエージェントに与えることにより，過去タスクの終端状態への近さを評価できる．ただし，$V_\pi(s)$ の値は 0 から 1 の値を取るよう正規化される．

提案手法では各ステップ毎に，そのステップtでの状態 s_t に対する $V_\pi(s_t)$ を，過去の行動価値を蓄積する変数 μ に加算してゆく．このμの値は，過去タスクの終端状態から遠い状態で行動した場合は緩やかに，近い状態で行動した場合は急激に増加してゆく．

探査戦略は，μを使って以下のようにバランスを取る：

- 確率 $\big(1-\sigma(x)\big)$ で，過去の政策に従う
- 確率 $\sigma(x)$で，$\varepsilon-\text{greedy}$ 戦略に従う

ただし $\sigma(x)$ はシグモイド関数

$$\sigma(x) = \frac{1}{1 + exp\,(-a(x-b))} \qquad (2)$$

を表す．また，$\varepsilon-\text{greedy}$の$\varepsilon$には，$\big(1-\sigma(x)\big)$を代入する．この探査戦略は，エピソード序盤においてはμはほとんど増加しないため，過去の政策に従って行動する．また，エピソード中盤においても，過去タスクの終端状態に近い状態

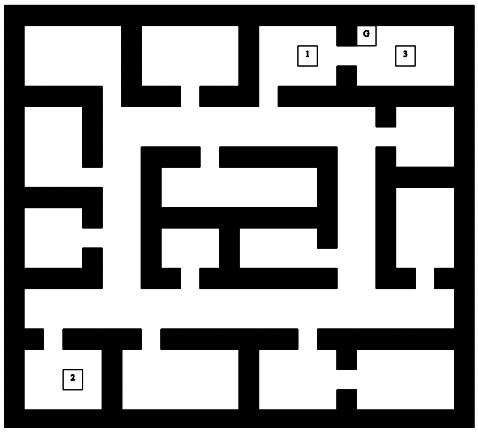

図 1:地図とゴール位置

を探査していない状況では μ の値は小さいため，過去のタスクに従って行動する．一方で過去タスクの終端状態に近い状態を探査した場合には η の値が大きいため，以降は $\varepsilon-\text{greedy}$ に従って行動する．

4. 実験

4.1. 環境

実験環境として，grid-based robot navigational domain[3]を用いる．エージェントは図 1 に表される地図上でナビゲーションタスクを行う．地図は24×21の広さの中に，それぞれ大きさ1×1の壁，通路，ゴールが配置される．エージェントは「東」「西」「南」「北」へ移動することができ，その移動距離は 0.8 から 1.2 の一様分布に従う．エージェントの座標は実数値を離散化し，24×21の領域として表現する．エージェントの移動が壁に阻まれる場合には，移動は失敗し移動前の位置にとどまる．

エージェントの目標は，エピソード開始から 100 ステップ以内に地図上のゴール位置 G へ到達することである．エピソード開始時に，エージェントは迷路上のランダムな位置に配置される．エージェントがゴールに辿り着いた場合に報酬 1 が与えられ，そうでない場合には報酬 0 が与えられる．100 ステップ経過するか，ゴールへ到達することで，エピソードが終了する．1 つのタスクは 2000 のエピソードによって学習される．

4.2. 比較方法

従来手法のπ-reuse[3]と比較実験を行なう．π-reuse は与えられた過去の政策を再利用し，新たなタスクを学習するアルゴリズムである．実験では，ゴール位置 G へ到達するタスクを学習するために，図 1 に表される 1,2,3 の位置へ到達するタスクから学習した政策π_1,π_2,π_3を与える．提案手法ではπ_1,π_2,π_3に加え，行動価値観数も与える．そして，従来手法と提案手法において，各政策を再利用したときの平均割引報酬を比較する．

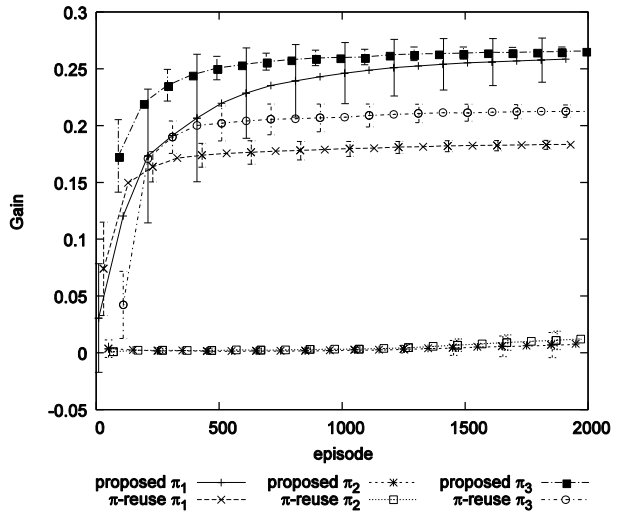

図 2:実験結果

4.3. 実験結果

共通パラメータとして $\alpha=0.05,\gamma=0.95$，π-reuse では $\psi=0,\upsilon=0.95$，提案手法では $a=0.4,b=5$ を用いて実験を行った．図 2 は各政策を再利用して学習したときの平均割引収益を表している．横軸はエピソード数であり，縦軸の Gain は平均割引収益である．実験は 10 回行い，標準偏差を計算した．実験結果から，ゴール位置同士が近いタスクの政策(π_1,π_3)を再利用した場合には，従来手法よりも多く報酬を得ていることがわかる．一方でゴール位置が遠いタスクの政策(π_2)を再利用した場合には，ほぼ同等の結果である．

5. 今後の課題

提案手法は与えられた政策を再利用し，効率的に学習できることを示した．今後は提案手法を発展させ，適切に選んだ一つの政策ではなく，複数の政策を再利用する手法を提案することを目指す．

参考文献

[1] J. Carroll and T. Peterson. Fixed vs. dynamic sub-transfer in reinforcement learning. In Proceedings of the International Conference on Machine Learning and Applications, 2002.

[2] M. G. Madden and T. Howley. Transfer of experience between reinforcement learning environments with progressive difficulty. Artificial Intelligence Review, 21:375–398, 2004.

[3] Fernández, F., and Veloso, M. 2006. Probabilistic Policy Reuse in a Reinforcement Learning Agent. In *Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multi-Agent Systems AAMAS '06*