1E4-4

搾取と探索のトレードオフを解決する適応的強化学習の提案

Adaptive Reinforcement Learning That Solves Trade-off Between Exploitation and Exploration

今井遼太郎 Ryotaro IMAI　吉川毅 Takeshi YOSHIKAWA　野中秀俊 Hidetoshi NONAKA　杉本雅則 Masanori SUGIMOTO

北海道大学大学院情報科学研究科コンピュータサイエンス専攻
Division of Computer Science, Graduate School of Information Science and Technology, Hokkaido University

In this study, we treat the trade-off between exploitation and exploration in reinforcement learning in a dynamic environment. In conventional methods, the trade-offs are often controlled beforehand by parameters or indices, which are highly dependent on tasks. These techniques may spoil advantages of reinforcement learning. We propose a method so that the agent itself controls the parameters for environmental change. As a result, the proposed method gets about 1.2 times more reward per episode compared with previous methods by simulating two-dimensional maze problems, which changes its environment.

1. はじめに

強化学習 [Sutton 88] は，エージェントが試行錯誤を繰り返すことによって環境と相互作用し，環境の状況に基づく行動選択の方策を獲得する機械学習手法である．強化学習では数値化された報酬信号を最大にするために，どのようにして状況に基づく動作選択を行うかを学習する．一般に，強化学習の利点として，エージェントが環境における事前知識や正しい制御規則を与えられることなく学習可能な点，確率的な行動規則の獲得が可能な点や，他の機械学習より設計者の負担が少ない点等が挙げられる．一方欠点として，学習効率が報酬の遅れが大きい状況下で低下する点やパラメータの値に敏感に反応する点が挙げられる．

本研究では，強化学習の重要なテーマの一つである搾取と探索のトレードオフを取り上げる．これは，エージェントが探索によって得られた知識を搾取に用いるタイミングと手段を決定する問題である．搾取と探索のトレードオフを理論的に解決する方法はまだ発見されていない．多くの手法では学習パラメータの設定にエージェントの知識を利用することによりトレードオフを図っている．

強化学習は確率的な行動規則の獲得が可能なため，ノイズが多い環境でも学習がある程度可能となるので，従来より動的環境下での強化学習が扱われてきた．動的環境下では，環境変化前の知識が環境変化後の学習を妨げやすいため，エージェントの知識は必ずしも正しい知識とは限らない．ゆえに，エージェントの知識を利用しても搾取と探索のトレードオフはうまく解決できない場合が多い．この問題点に対し，多くの既存研究は，設計者が事前にパラメータを制御する手法やタスク依存度の高い指標を用いる手法等によって搾取と探索のトレードオフに対処している．しかし，複数のパラメータの制御は煩雑で大きな負担となるだけでなく，強化学習の本来の利点，すなわちタスク非依存性に逆行するものと考えられる．

筆者らは，このような既存研究における問題点を改善するためには，エージェント自身のパラメータ制御と，設計者の負担を少なくして可能な限りタスクに依存しない指標を用いる必要があると考えた．さらに，環境変化に適応するには，環境が変化した際にエージェントが環境変化前の行動に固執せずにその変化を認識して適切な行動の学習を行う必要があると考える．

著者らは，エージェント自身が環境変化を認識し，パラメータや行動価値を修正可能な学習手法を提案した．提案手法は設計者の負担を軽減でき，かつタスクに非依存な手法である．この手法により動的環境下で搾取と探索のトレードオフの解決を図ることができる．提案手法の有効性を環境変化が発生する 2 次元迷路探索シミュレーション実験で確認した．その結果，提案手法が既存手法よりも単位エピソードあたり約 1.2 倍程度の報酬の獲得した．

連絡先: 今井遼太郎，北海道大学，〒 060-0814 北海道札幌市北区北 14 条西 9 丁目，TEL : 011-706-6861, ryotaro@main.ist.hokudai.ac.jp

2. 事前知識

2.1 強化学習の目標

強化学習では，状態 s で行動 a を選択する価値 $Q(s,a)$ は以下のように与えられる．

$$Q(s,a) = \mathrm{E}\left[\sum_{i=0}^{\infty} \gamma^i r_{t+i+1} \middle| s_t = s, a_t = a\right] \tag{1}$$

r_{t+i+1} はエージェントがある $t+i$ ステップに受け取った報酬，$\gamma(0 \leq \gamma \leq 1)$ は割引率とよばれる報酬を調整するパラメータである．$Q(s,a)$ は行動価値関数とよばれ，各ステップで通常以下のように更新する．

$$Q(s,a) \leftarrow Q(s,a) + \alpha[Q^*(s,a) - Q(s,a)] \tag{2}$$

$Q^*(s,a)$ は最適行動価値関数とよばれ，強化学習においてはこの $Q^*(s,a)$ の獲得が目標となる．式 (2) 中の $\alpha(0 \leq \alpha \leq 1)$ は学習率とよばれる学習を調整するパラメータである．

2.2 softmax 方策

強化学習では，エージェントが行動価値の推定値を基に確率的に行動を選択する．各ステップにおいて状態 s から取りうる行動 a を選択する確率への写像のことを方策 $\pi(s,a)$ とよぶ．softmax 方策 [Luce 59] は，最善の行動には最も高い選択確率が与えられ，他の行動には行動価値関数の値の高い順に選択確率が与えられる方策である．状態 s における行動 a を選択する確率を以下のように決定する．

$$Pr(s,a) = \frac{\exp(Q(s,a)/\tau)}{\sum_{a' \in \mathcal{A}(s)} \exp(Q(s,a')/\tau)} \tag{3}$$

$\mathcal{A}(s)$ は状態 s で選択可能な行動の集合，$\tau(\tau > 0)$ は温度とよばれる搾取と探索を調整するパラメータである．softmax 方策は，温度が高い場合には探索を，温度が低い場合には搾取を選択しやすくなる．

2.3 Sarsa(λ) 学習

Sarsa 学習 [Sutton 88] は最適行動価値関数を次の状態 s' と次の行動 a' を用いて $r+\gamma Q(s',a')$ と近似する手法である．この Sarsa 学習に状態への訪問，あるいは行動の実行等の事象発生の一時的な記憶として用いられる適格度トレースをトレース減衰パラメータ $\lambda(0 \leq \lambda \leq 1)$ で導入したものが Sarsa(λ) 学習 [Rummery 94] である．

2.4 強化学習パラメータ

ここでのパラメータとは，割引率 γ，学習率 α，温度 τ，トレース減衰パラメータ λ を意味する．表 1 に既存研究におけるエージェントの学習状況に応じた適切なパラメータの制御方法を示す [吉田 01] [尾川 03] [Murakoshi 04].

表 1: パラメータ変数と学習状況の関係

	学習が浅い場合	学習が深い場合
γ	小さくして即時報酬を重視	大きくして将来報酬も考慮
α	大きくして学習速度を向上	小さくして学習精度を向上
τ	大きくして探索を促進	小さくして搾取を促進
λ	小さくして即時報酬を重視	大きくして将来報酬も考慮

2.5 信頼度

エージェントが内部モデルをどの程度信頼できるかを主観的に評価した尺度に信頼度 [Sakaguchi 01] がある．信頼度は内部モデルの予測誤差に基づき更新され，信頼度に比べ誤差が小さければ信頼度は向上し，大きければ低下する．つまり，信頼度が低い場合は学習が浅いことを示し，信頼度が高い場合は学習が深いことを示すものと考える．

信頼度 R を各状態 s と行動 a の組に対して定義するものとして，信頼度の更新規則は以下のように与えられる．

$$R^2(s,a) \leftarrow R^2(s,a) + \alpha_R {}^R\delta, \tag{4}$$

$${}^R\delta = \delta^2 + \gamma_R R^2(s',a) - R^2(s,a) \tag{5}$$

α_R は信頼度の学習率，${}^R\delta$ は信頼度の TD 誤差，γ_R は信頼度の割引率，δ は Sarsa(λ) 学習における TD 誤差である．このとき式 (5) から，信頼度は二乗 TD 誤差の累積値と見なすことができるため，TD 誤差の試行ごとのばらつきが吸収されることによって学習が安定化すると予想される．

3. 提案手法

3.1 環境変化の認識

エージェントが環境変化を認識する際の指標として，エージェントが受け取る報酬を用いる．k 番目のエピソードにおけるエピソード終了時のステップを T_k ステップとする．k エピソード内で受け取った報酬の平均である EA_k を式 (6) により定義する．

$$EA_k = \sum_{i=1}^{T_k} \frac{r_i}{T_k} \tag{6}$$

ここで，$\{EA_k\}$ を大きさ順に並べ替えたものを $\{EA'_k\}$ とする．つまり，$EA'_1 \geq EA'_2 \geq \cdots \geq EA'_k$ である．

このとき，現在を n エピソード目とした場合，$\{EA'_k\}$ の上位 M 試行の平均値または，下位 M 試行の平均値と比較することで環境変化を認識する．

$$EA_n > \sum_{m=1}^{M} \frac{EA'_m}{M}, \tag{7}$$

$$EA_n < \sum_{m=1}^{M} \frac{EA'_{k-m+1}}{M} \tag{8}$$

式 (7) の場合，最適な行動系列がより短いステップ数である環境に変化した判断し，式 (8) の場合，最適な行動系列がより長いステップ数である環境に変化したと判断する．

3.2 パラメータと行動価値の修正

まず，パラメータの修正について説明する．エージェントの学習状況を表す指標の計算に信頼度を用いる．その際に各行動の行動価値の重みを付ける変数として，行動価値関数 $Q(s,a)$ を基に式 (9) で定義する $\mathrm{Var}[Q(s,a)]$ を用いる．

$$\mathrm{Var}[Q(s,a)] = \mathrm{E}[Q(s,a)^2] - \mathrm{E}[Q(s,a)]^2 \tag{9}$$

$\mathrm{E}[Q(s,a)]$ は行動 a に関する平均を表し，各状態ごとに 1 つの値をとる．このとき，エージェントの学習状況を表す指標 LP を以下の式で定義する．

$$LP = \sum_{a \in \mathcal{A}(s)} \sum_{s \in \mathcal{S}} \frac{R(s,a)\sqrt{\mathrm{Var}[Q(s,a)]}}{|\mathcal{A}(s)||\mathcal{S}|} \tag{10}$$

ただし，$\mathcal{S}$ は全状態を意味する集合である．このように定義した LP を用いて，表 1 に従い以下の式でパラメータの制御を行う．

$$\begin{aligned} &\gamma_t = \gamma_0 \min\left(1, \frac{1}{LP}\right), \lambda_t = \lambda_0 \min\left(1, \frac{1}{LP}\right), \\ &\alpha_t = \alpha_0 \max\left(1, \frac{1}{LP}\right), \tau_t = \tau_0 \max\left(1, \frac{1}{LP}\right) \end{aligned} \tag{11}$$

次に，行動価値の修正について説明する．パラメータの修正のみでは行動価値自体の値は変化しないため環境変化に適応することは難しい [三村 10]．また，環境変化前の最適行動に固執させないためにも適宜行動価値の修正を図る必要がある．エージェントが環境変化を認識した場合，全ての行動価値を式 (12) により修正する．

$$Q(s,a) \leftarrow Q(s,a) + \kappa \frac{R(s,a)[Q(s,a) - \mathrm{E}[Q(s,a)]]}{\max\limits_{\substack{s' \in \mathcal{S} \\ a' \in \mathcal{A}(s')}} R(s',a')} \tag{12}$$

κ は式 (7) の場合に 1 を，式 (8) の場合に-1 の値をとる．

目標状態に到達できると見込める行動系列の行動価値を増加させることにより，それらの行動系列がより多く選択されるように，障害に到達すると見込める行動系列の行動価値を減少させることにより，それらの行動系列がより少なく選択されるように修正を行う．このように行動価値を修正することで環境変化に適応させる．

3.3 適応的 Sarsa(λ) 学習

環境変化の認識およびパラメータと行動価値の修正を Sarsa(λ) 学習に導入した適応的 Sarsa(λ) 学習アルゴリズムを提案する．

はじめに，全ての状態 s と行動 a に関して行動価値関数 $Q(s,a)$ と適格度トレース $e(s,a)$ を 0 に初期化する．状態 s において式 (3) の softmax 方策から行動 a を選択して，次の状態 s' を観測し状態 s で行動 a を選択した際の報酬 r を受け取る．状態 s' においても同様に式 (3) の softmax 方策から行動 a' を選択して，式 (13) によって定義される Sarsa(λ) 学習の TD 誤差 δ を計算する．

$$\delta = r + \gamma Q(s', a') - Q(s, a) \tag{13}$$

次に，全ての状態行動対に対して行動価値の更新を式 (14) によって行う．

$$Q(s, a) \leftarrow Q(s, a) + \alpha\delta e(s, a) \tag{14}$$

適格度トレース $e(s,a)$ の更新は入れ替え更新トレースの一般化としての式 (15) で行う [Sutton 96].

$$e(s, a) \leftarrow \begin{cases} 1 & (s = s_t \text{ かつ } a = a_t \text{ のとき}) \\ 0 & (s = s_t \text{ かつ } a \neq a_t \text{ のとき}) \\ \gamma\lambda e(s, a) & (s \neq s_t \text{ のとき}) \end{cases} \tag{15}$$

s_t は遷移前の状態，a_t は状態 s_t において選択した行動を表す．式 (7)，(8) により環境変化の有無を調べ，環境変化を認識した場合，式 (12) より全ての状態行動対に対して行動価値を修正する．その後，式 (10) から学習進度 LP を求め，式 (11) においてパラメータの値を更新する．

以上の手続きを試行終了条件を満足するまで繰り返す．

4. 実験

本実験の目的は，提案手法により設計者の負担の少ない手法であってもエージェントが環境変化を認識し，パラメータを自己制御することで搾取と探索のトレードオフを図ることができるかを確認することである．比較手法として，環境変化を認識しパラメータを指数関数により制御する手法，環境変化を認識せず，パラメータも初期値に固定した通常手法を採用した．なお，比較手法ではいずれも方策は softmax 方策，基本的な学習アルゴリズムは通常の Sarsa(λ) 学習を用いる．

4.1 実験環境

Dayan らが提案した迷路探索問題を改良した 2 次元迷路探索問題でシミュレーション実験を行う．エージェントは，図 1-3 における 20 × 20 の格子状の 2 次元迷路内にて下部の赤いマスである初期状態から上部の青いマスである目標状態まで移動する．目標状態に到達した場合にのみ報酬 1 が与えられる．エージェントは上，右，下，左の 4 つの行動を選択できる．1 回の実験を 6000 エピソードとして，各エピソードは固定された初期状態から開始し，エージェントが目標状態に到達するか，最大ステップ数である 500 ステップ進むと終了する．図 1-3 中において白いマスがエージェントが遷移できる状態を表し，黒いマスが壁つまり障害を表す．エージェントは黒いマスに遷移することはできず，白いマスから黒いマスへの行動を選択した場合は-1 の報酬が与えられ，初期状態に遷移する．

本実験では動的環境として，2000 エピソードと 4000 エピソードを境目に迷路の構造を変化させる．なお，強化学習パラメータの値は $\alpha_0 = 0.90$，$\tau_0 = 0.1$，$\gamma_0 = 0.50$，$\lambda_0 = 0.99$，信頼度におけるパラメータの値は $\alpha_R = 0.10$，$\gamma_R = 0.90$ とした．

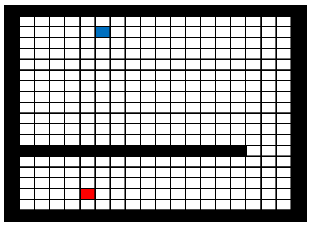

図 1: 1〜2000 エピソードにおける環境

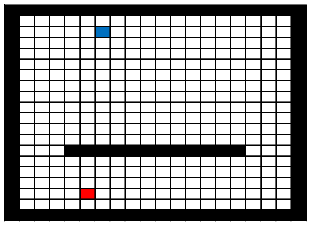

図 2: 2001〜4000 エピソードにおける環境

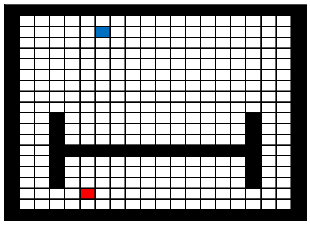

図 3: 4001〜6000 エピソードにおける環境

4.2 実験結果

提案手法，比較手法それぞれ 1000 回の試行を行い表 2 に各手法の実験終了時までに獲得した報酬の結果を，図 4 に累積報酬の推移を示す．(1) は提案手法，(2) は環境変化を認識しパラメータを指数関数により制御する手法，(3) は環境変化を認識する機構もパラメータを制御する機構も持たない通常の手法である．表 2，図 4 はともにある一試行における実験結果であり，累積報酬はエージェントがその時点までに受け取った報酬の和を意味する．数値に多少の誤差はあるものの，他の試行でも同様の結果が得られた．

4.3 考察

提案手法，比較手法ともに図 4 において最初の約 1000 エピソードの間までは探索を選択する回数より搾取を選択する回数が多いために，累積報酬は負の値となっていると考えられる．その後，最初の環境変化前までは学習アルゴリズムとして同等のものを採用しているため提案手法，比較手法 (2) に大きな差異は見られない．比較手法 (3) がこれらよりも，やや高い累積報酬となっているのは，探索よりも搾取をより多く選択しているためと考えられ，学習後半においてもあまり累積報酬が増えないのはそのためであると考えられる．

最初の環境変化が発生する 2001 エピソードから 4000 エピソードの間では，提案手法，比較手法 (2) は今までの環境に比べ目標状態までの最適な行動系列が短いステップ数である環境と判断できたため，1 エピソードから 2000 エピソードの間で

表 2: 提案手法と比較手法の実験結果

	エピソード	(1)	(2)	(3)
目標状態到達回数	[1, 2000]	121	121	210
	[2001, 4000]	585	587	254
	[4001, 6000]	514	311	101
	計	1220	1019	565
衝突回数	[1, 2000]	123	113	110
	[2001, 4000]	8	13	13
	[4001, 6000]	45	47	39
	計	176	173	162
累積報酬		1044	846	403

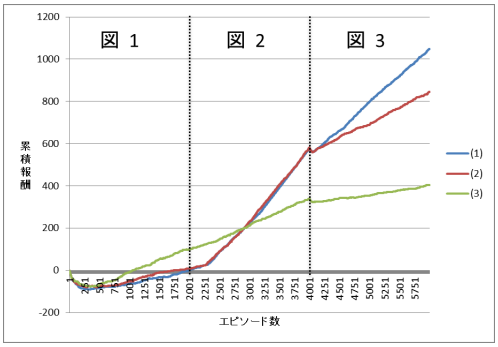

図 4: 提案手法と比較手法での累積報酬の比較

の累積報酬の傾きよりも急になっていると考えられる．

環境変化を認識できない比較手法 (3) は，表 2 から 2001 エピソードから 4000 エピソードの間と 1 エピソードから 2000 エピソードの間での目標状態到達回数がほとんど同じであるため，より短い最適な行動系列が獲得できず，環境変化前と同様に行動していると考えられる．

次の環境変化が発生する 4001 エピソードから 6000 エピソードの間では，エージェントが今まで移動していた状態に遷移できない状態が現れたため，いずれの手法でもにおいて約 200 エピソードの間に負の報酬を受け取り累積報酬が下がったと考えられる．その後，提案手法では 2 度目の環境変化前における累積報酬の傾きとほとんど同じであるが，比較手法では環境変化前よりも傾きが緩やかになった．これは，環境変化の際に提案手法が環境変化前に学習した最適行動に固執することなく環境変化後の最適行動の獲得の達成ができたのに対して，比較手法が環境変化前に学習した最適行動に固執したため環境変化後の最適行動の獲得の達成ができなかったためと考えられる．

提案手法は，4001 エピソード以降で環境変化によって学習経路が変化した際に，その学習経路の変化を認識して適切な行動の学習を行うことで，今までの最適行動系列の行動価値を下方修正し累積報酬の減少を抑制したと考えられる．しかし比較手法 (2)(3) では環境変化前に学習した最適行動の知識の影響により，環境変化後の最適行動を学習できていない．一時的に提案手法や比較手法 (2) よりも多くの報酬を得ていた比較手法 (3) は，環境変化に適応できた提案手法に比べて，環境変化に適応できなかったため最終的な累積報酬では提案手法や比較手法 (2) を下回る結果となったと考えられる．

以上の結果から，提案手法は環境変化を認識することができ，エージェント自身が適宜パラメータや行動価値を修正することで，環境変化前の学習に固執することなく環境変化後の最適な行動を学習することができ環境変化に適応できたと考えられる．

5. おわりに

本研究では，設計者の負担を少なくするとともにタスク非依存性である指標を用いて動的環境下において環境変化の認識とエージェント自身によるパラメータ制御及び行動価値の修正により，動的環境での搾取と探索のトレードオフ問題の解決を図った．また，実際に環境変化を認識し，環境に適応できるか，そしてパラメータを自己制御できるかどうかを確認するシミュレーション実験を行った．その結果，エージェントは環境変化後の最適行動を獲得でき，単位エピソード当たりの報酬も既存手法に比べて増加したことを確認した．以上の結果より，設計者の負担を少なくする提案手法を用いることで動的環境下において搾取と探索のトレードオフを図ることができたと考える．

本報告では，環境変化を認識する際に環境が変化したということだけを認識する比較的単純な機構を導入したが，さらに，環境変化の前後の状態を特定できるような機構に改良していきたい．また，信頼度を用いる際のパラメータの制御についても定数値ではなく動的に調整できるようにしていきたい．

参考文献

[Doya 02] Doya K.: Metalearning and neuromodulation, Neural Networks, 15, pp.495-506(2002).

[三村 10] 三村 明寛, 加藤 昇平, 伊藤 英則: 動的環境下における危険度予測法に基づく適応的強化学習, 第 24 回人工知能学会全国大会, pp.1A3-3, June 09-11(2010).

[Luce 59] Luce, R. D.: Individual Choice Behavior, New York, Wiley(1959).

[Murakoshi 04] Murakoshi K, Mizuno J: A parameter control method in reinforcement learning to rapidly follow unexpected environmental changes, Biosystems, Nov.77(1-3), pp.109-117(2004).

[尾川 03] 尾川順子, 並木明夫, 石川正俊: 学習進度を反映した割引率の調整, 電子情報通信学会技術研究報告, NC, ニューロコンピューティング 102(628), pp.73-78(2003).

[Rummery 94] Rummery G. A., Niranjan M.: Online Q-Learning using Connectionist Systems, technical report no.166, University of Cambridge, Engineering Department(1994).

[Sakaguchi 01] Sakaguchi Y., Takano M.: Learning to switch behaviors for different environments: A computational model for incremental modular learning, Proc. 2001 Int. Symp. Nonlinear Theory and its Applications (NOLTA2001), pp.383-386, Oct(2001).

[Sutton 88] Sutton R. S., Barto A. G.: Reinforcement Learning: An Introduction, MIT Press, Cambridge, MA(1988).

[Sutton 96] Sutton R. S.: Generalization in Reinforcement Learning: Successful Examples Using Sparse Coarse Coding, Advances in Neural Information Processing Systems 8, pp.1038-1044, MIT Press(1996).

[吉田 01] 吉田 和子, 石井 信: 強化学習における exploration と exploitation の制御, 電子情報通信学会技術研究報告. NC, ニューロコンピューティング 101(154), pp.41-48(2001).