1A3-03

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

Adaptive Reinforcement Learning based on Risk Forecast Method under Dynamic Environment

三村 明寛[*1] 加藤 昇平[*1] 伊藤 英則[*1]
Akihiro MIMURA Shohei KATO Hidenori ITOH

[*1]名古屋工業大学大学院工学研究科情報工学専攻
Department of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology

In the existing Reinforcement Learning, it is difficult for the learning agent to adapt itself to dynamic environment. The cause of this problem is that the learning agent keeps exploiting when the barrier appears because of the environmental change. Then we design environmental change perception mechanism that the learning agent can perceive appearance of barrier. We propose a learning method that the learning agent can adapt itself to new environment by promoting the exploration. This method adjusts Q value based on risk in each state when the leaning agent perceives the barrier with the mechanism. We apply the proposed method to two-dimensional maze problems and report on the effect.

1. はじめに

近年，機械学習の一種である強化学習（Reinforcement Learning）[Sutton 98] が注目を集めている．強化学習とは，学習者が試行錯誤を繰り返す中で，適切な制御規則を獲得する学習手法である．強化学習では，学習者自身が制御規則を獲得するので，あらかじめ起こりうる状況を想定してプログラミングする必要はない．また，人間が発見し得なかった制御規則の発見が期待される [木村 99]．設計者は目的となる情報を入力するだけで良いので，未知の環境下での制御規則獲得に有用であると考えられる．

強化学習が他の機械学習と最も異なる点として，正しい制御規則を教示する必要が無いことが挙げられる．強化学習では，報酬関数と価値関数という 2 つの評価関数があり，学習者はこれらの情報を基にして最適な制御規則を学習していく．報酬関数とは，学習者が選択した行動に対して与えられる情報で，即時的な評価を行う．報酬が大きいことは，選択した行動がその時点において良い行動であったことを表す．学習者は，最終的に受け取る報酬の累積値が最大になるように学習を進行させることになる．価値関数は，ある状態や行動に備わった長期的な望ましさを表す．価値が大きいことは，最終的に受け取ることができる予想累積報酬が大きいことを表す．学習者は現在の行動価値の推定値を基にして行動を選択するので，行動価値の推定値を学習により真の価値に近似することを目指す．

しかしながら，強化学習の研究の多くは，学習中における動的な環境の変化を考慮していない．そのため，環境の変化により障害が発生した場合，学習者がそれまでに獲得した制御規則では対応できなくなってしまう可能性がある．本研究では，環境変化による障害発生を学習者自らが感知し，その時点での行動価値の推定値を修正し，新しい環境に適応する学習手法を提案する．ここで，対象とする環境変化とは，障害発生を伴うものとする．

連絡先: 三村明寛，名古屋工業大学，〒 466-8555 愛知県名古屋市昭和区御器所町，TEL：052-735-5625，E-mail：shohey@ics.nitech.ac.jp，mimura@juno.ics.nitech.ac.jp

2. 行動選択規則

2.1 行動選択

強化学習では，学習者自身が行動を選択する．学習者は行動価値の推定値を基にして行動選択を行う．学習者は，搾取（exploitation）か探査（exploration）のいずれかを選択し行動する．搾取とは，それまでの経験で得られた知識を利用することにより最適と思われる行動を選択することである．すなわち，現在最も高い行動価値を持つ行動を選択することで，その行動をグリーディな行動という．一方，探査とはグリーディでない行動を選択することである．

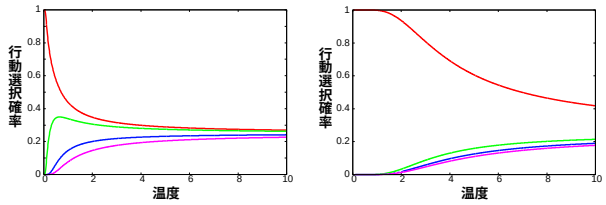
行動価値の推定値が真の価値に近似できている環境，すなわち完全に既知な環境では，搾取をすることにより最適な行動を選択することができる．しかしながら，行動価値の推定値が真の価値に近似できていない環境，すなわち未知な環境では，グリーディな行動が本当に最適な行動であるかどうかはわからない．つまり搾取を行うことにより，結果的に最適でない行動を選択し続けることになる可能性がある．そこで，学習者は，新たな知識を獲得するために探査を行う必要がある．同様に，環境の変化などにより行動の真の価値が時間とともに変化するような場合も探査が必要となる．このように，強化学習では搾取と探査のバランスを取ることが重要であると考えられる．

2.2 softmax 行動選択規則

行動を選択するときに，強化学習では softmax 行動選択規則（softmax action selection rule）を用いることが多い [石井 02][小堀 05]．softmax 行動選択規則は，各行動の価値の推定値（Q 値）と温度定数 τ により行動選択確率が決められる．各行動には，それぞれの Q 値により重み付けされた選択確率が割り振られる．状態 s における行動 a の選択確率 $P(a|s)$ は，式 (1) により決まる．

$$P(a|s) = \frac{\exp(Q(s,a)/\tau)}{\sum_{a' \in A} \exp(Q(s,a')/\tau)} \quad (1)$$

ここで，A は状態 s で選択可能な行動の集合，$Q(s,a)$ は状態 s における行動 a の Q 値，τ は温度を表す．式 (1) により決定された確率から，各行動は確率的に選択されることになる．



(a) 分散が小さい場合　　(b) 分散が大きい場合

図 1: Q 値の分散の違いによる各行動の選択確率

図 1 に，ある状態における各行動の選択確率の一例を示す．ここで，選択可能な行動数は 4 とする．同図 (a)，(b) はそれぞれ各行動の Q 値の分散が小さい場合，分散が大きい場合の各行動の選択確率を表し，縦軸が行動選択確率で，横軸が温度である．

各行動の Q 値の大きさに応じてそれぞれの行動選択確率が割り振られるので，各行動の Q 値の分散の大きさにより搾取と探査の選択確率が変化する．各行動の Q 値の分散が小さい場合は，各行動に与えられる選択確率の差は小さいので，探査を行い易くなる．一方，各行動の Q 値の分散が大きい場合は，グリーディな行動に大きい選択確率が与えられるので，探査を行いにくくなる．すなわち，グリーディな行動の Q 値が突出して大きい場合に，他の行動がほとんど選択されなくなってしまう．また，行動価値の分散が大きくなる学習後期において，探査がほとんどされなくなる．環境変化によりグリーディな行動の真の価値が下がった場合でも，搾取ばかり行い最適でない行動を選択し続けることになり，新しい環境に適応することが困難となる．

温度は各行動の選択確率を決定するために必要なパラメータである．温度が高い場合には，すべての行動の選択確率がほぼ同じになるように設定され，ランダムに行動を選択するようになる．逆に温度が低い場合には，グリーディな行動を選択しやすくなる．つまり，温度が高い場合は探査，温度が低い場合は搾取をしやすくなる．

3. 環境変化の感知

本研究では，学習者自身が環境変化を感知する機構を設計した．環境変化の感知には，各試行の累積獲得報酬を用いる．通常，定常な環境においては，学習の進行に伴い累積獲得報酬は小さな増減を繰り返しながら徐々に増加していく．しかしながら，環境変化により障害が発生すると，累積獲得報酬は減少すると考えられる．そこで，本研究では，累積獲得報酬の増減の傾向に着目し，累積獲得報酬が減少傾向にある場合，環境が変化したとみなす．

ここでは，累積獲得報酬の増減傾向を判断するために，パフォーマンス Pf を用いる．Pf は，以下の式で定義される．

$$Pf = R_{ref} - R_{min} \tag{2}$$

ここで，R_{ref} は基準となる報酬，R_{min} は同一環境下における累積獲得報酬の最小値を表す．また，同一環境下での試行とは，学習者が環境変化を感知していない期間における試行を意味する．式 (2) を基にして，学習者の最大のパフォーマンス Pf_{max} と現在のパフォーマンス Pf_{now} をそれぞれ式 (3)，(4) により計算する．

$$Pf_{max} = \bar{R}_{max} - R_{min} \tag{3}$$

$$Pf_{now} = \bar{R}_{now} - R_{min} \tag{4}$$

ここで，$\bar{R}_{max}$ は同一環境下での試行における累積獲得報酬が大きい上位 N 試行の平均値，$\bar{R}_{now}$ は直近 M 試行における累積獲得報酬の平均値とする．Pf_{max} と Pf_{now} の比が閾値 $L(0 < L < 1)$ 未満となる場合，環境が変化したと判断する (式 (5)) .

$$\frac{Pf_{now}}{Pf_{max}} < L \tag{5}$$

学習者は，式 (5) を満たさない限り環境が変化したとはみなさない．すなわち，累積獲得報酬に与える影響が小さい場合は，を学習者自身は環境変化を感知しない．

4. 危険度予測法

従来の強化学習では，各行動の Q 値の分散が大きい場合にほとんど探査が行われず，環境変化に適応することが困難である．[石井 02] では，この問題を解決するために各状態の温度を制御することにより，各行動の Q 値の分散が大きい場合でも探査の促進をする手法を提案している．しかしながら，温度を制御することで行動の選択確率を変化させることはできるが，行動の優先順位自体は変わらない．そのため，本研究で対象としているような障害発生問題，すなわちグリーディな行動の真の価値が大きく変化するような問題には適さない．障害が発生した場合には，障害に至るまでの行動系列の優先順位を下げるべきである．これにより，他の行動すなわちグリーディでない行動の優先順位が上がり，探査を促進することができると考えられる．

4.1 危険度

本研究では，危険度という環境が変化した場合に各状態が受ける影響の大きさを表す指標を用いる．環境変化時の影響の大きさは，各行動の Q 値の分散により判断する．Q 値は，その行動を選択することで将来にわたって得られる累積報酬がどれくらいであるかを表したものである．したがって，Q 値の分散が小さい状態では，どの行動を選択しても将来にわたって得られる累積報酬にあまり差が無いと考えられる．このように Q 値の分散の小さい状態において，環境変化による障害が発生したとしても影響は小さいと考えられるので危険度は小さいと判断する．逆に Q 値の分散が大きい状態では，環境に変化が生じたときの影響が大きいと考えられるので危険度は大きいと判断する．状態 s の危険度 $d(s)$ は，式 (6) により算出される．

$$d(s) = E[Q(s,a)^2] - (E[Q(s,a)])^2 \tag{6}$$

ここで，$Q(s,a)$ は状態 s における行動 a の Q 値，$E[\ ]$ は期待値を表す．

学習が収束した状態では，その状態でどの行動を選択すればより大きい累積報酬を得られるかわかっている．そのため，Q 値の分散もある程度大きいと考えられる．また，周辺の状態によっても Q 値の分散は異なると考えられる．例えば，図 2 のような迷路探索問題において，周りが障壁に囲まれているような状態（赤で囲まれたエリア）では，選択すべき行動は限られる．このような状態でも Q 値の分散は大きくなる．すなわち，選択するべき行動が明確な状態において，環境変化による障害が発生すると多大な影響を受ける可能性がある．したがってこのような状態における危険度は高くなる．逆に周りに障壁が存在しないような状態（青で囲まれたエリア）では，たとえその状態が環境変化により塞がれたとしても他の行動を選択することによって簡単に回避できるので危険度は小さい．

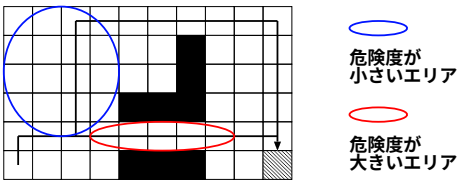


図 2: 状態による危険度の違い

4.2 危険度予測法に基づく Q(λ) 学習

本研究では，基本的な学習アルゴリズムに単純化更新 Q(λ) 学習 [Sutton 98] を採用する．これに前節で述べた環境変化感知機構を組み込み，学習者が環境の変化を感知したときに Q 値を修正する．

Q 値の更新は，1 ステップごとにすべての状態行動対に対して行われる．まず，状態 s で softmax 行動選択規則（式 (1)）に従い，行動 a を選択する．行動 a を選択し，報酬 r と次状態 s' を観測し，式 (7) により TD 誤差 δ を計算する．

$$\delta = r + \gamma \max_{a' \in A} Q(s', a') - Q(s, a) \tag{7}$$

ここで，γ は割引率，A は状態 s' で選択可能な行動の集合を表す．次にすべての状態行動対に対して，Q 値の更新を行う．更新式を式 (8) に示す．

$$Q(s, a) \leftarrow Q(s, a) + \alpha \delta e(s, a) \tag{8}$$

ここで，α は学習率，$e(s, a)$ は状態 s の行動 a に対する適格度トレースを表す．適格度トレースは，すべての状態行動対が持っているパラメータである．1 ステップごとに訪問された状態行動対の適格度トレースはインクリメントされ，訪問されなかった状態行動対は減衰する．適格度トレースの更新式を式 (9) に示す．

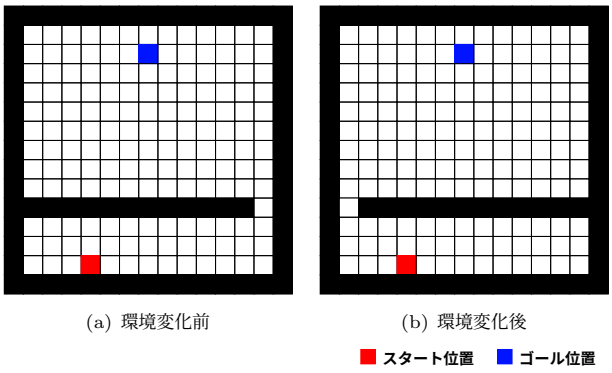
$$e(s, a) \leftarrow \begin{cases} \gamma\lambda e(s, a) + 1 & (s = s_t \text{かつ } a = a_t \text{のとき}) \\ \gamma\lambda e(s, a) & (\text{それ以外のとき}) \end{cases} \tag{9}$$

ここで，λ はトレース減衰率，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．以上の手続きを試行の終了条件を満たすまで繰り返す．試行の終了条件とは，学習者が目標状態に遷移する，あるいは，ステップ数が上限に達することである．

試行が終了したときに，学習者はその試行で獲得した累積報酬から学習者の現在のパフォーマンス Pf_{now} を式 (4) により更新する．また，累積獲得報酬が上位 N 試行に入る場合は最大のパフォーマンス Pf_{max} を式 (3) により更新する．算出された Pf_{now}，Pf_{max} が条件式 (5) を満足する場合，環境が変化したとみなし，Q 値を修正する．

環境変化を感知したということは，累積獲得報酬が減少傾向にあることを表し，環境変化が学習者に与える影響が大きい．すわなち，危険度の高い状態で環境が変化したと考えられるので，危険度の高い状態の Q 値を大きく修正するべきであると考えられる．従って，Q 値の修正は，各状態の危険度を用いてすべての状態行動対に対して行う．まず，式 (6) により，すべての状態の危険度 $d(s)$ を計算する．次に，式 (10) により，すべての状態行動対の Q 値を修正する．

$$Q(s, a) \leftarrow Q(s, a) - \frac{d(s)}{\max_{s' \in S} d(s')} \left[Q(s, a) - \frac{\sum_{a' \in A} Q(s, a')}{|A|} \right] \omega \tag{10}$$



(a) 環境変化前　(b) 環境変化後

図 3: 環境変化のある 2 次元迷路

ここで，S は現在の環境下におけるすべての状態集合，A は状態 s で選択可能な行動の集合，ω は危険度の影響力を調節する重みを表す．式 (10) により，すべての Q 値はそれぞれの状態の Q 値の平均値に近づくように修正される．危険度が大きい状態では，それぞれの Q 値が平均値に近い値に平滑化される．これにより，探査を促進し，新しい環境に適応できると考えられる．

5. 適応学習実験

動的環境下における危険度予測法の効果を確認するために比較実験を行った．比較対象として，環境変化時に行動価値を修正しない従来の手法を用いた．双方とも学習アルゴリズムは Q(λ) 学習，行動選択は softmax 行動選択規則を採用した．また，学習率，割引率，トレース減衰率，温度などの定数パラメータは同一の値を用いた．

5.1 実験条件

実験はシミュレーション上で行い，問題環境は文献 [Sutton 98] で用いられた障害が発生する二次元迷路探索問題の状態空間を拡張したものを用いた．図 3 に実験で用いた迷路を示す．同図において，各マスが状態で，白いマスは学習者が遷移することができる状態，黒いマスは遷移することができない状態（障壁）を表す．同図 (a)，(b) はそれぞれ環境変化前，環境変化後の状態空間である．最初はゴールに到達するために右側の通路を通過しなければならないが，環境変化により右側の通路は塞がれ，左側に新しい通路が出現する．

1 回の実験を 100 試行とし，50 試行終了した時点で環境が変化する．1 試行は，ゴールに到達する，あるいは 500 ステップ行動したときに終了する．1 ステップで，上下左右のいずれかに 1 マス移動することができ，1 ステップごとに-1 の報酬が与えられる．学習者が壁に衝突するような行動を選択した場合は，状態の遷移は行われない．学習者は，ゴールに到達するまでマイナスの報酬が与えられ続けることになる．学習者は累積獲得報酬が最大になるように学習を進めるので，より少ないステップ数でゴールに到達できる制御規則の獲得を目指すことになる．

5.2 実験結果

図 4 に，提案手法と従来手法の平均累積獲得報酬の推移の比較図を示す．同図において，提案手法は環境変化を感知するまでは通常の Q(λ) 学習と同じ振る舞いをするので，50 試行までは従来手法との差は見られない．また，約 30 試行経過したところで累積獲得報酬に大きい変化は見られなくなり，価値関

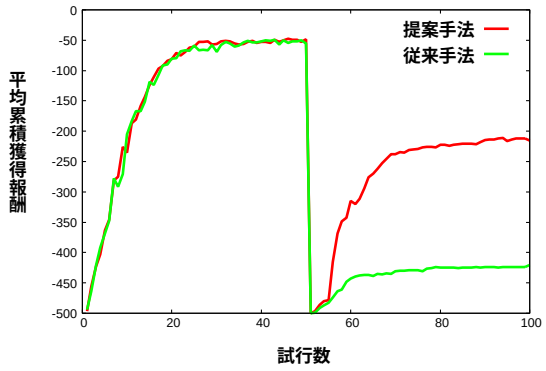


図 4: 平均累積獲得報酬の推移の比較

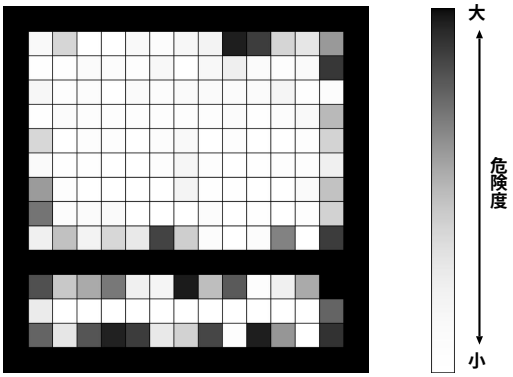


図 5: 環境変化発生直前の各状態の危険度

数の推定値がある程度収束していると考えられる．環境変化が発生する 51 試行以降は大きい差異が確認できる．提案手法が従来手法と比較して累積獲得報酬が大きいことが確認された．

図 5 に，環境変化発生直前 50 試行終了時点での各状態の危険度を表したものの一例を示す．同図において，暗い色になっている状態ほど危険度が大きいことを表す．中央の障壁よりも下側の状態や障壁に隣接する状態において危険度が大きいことがわかる．また，右側の通路付近の状態では特に大きい危険度を示していることがわかる．

5.3 考察

環境変化が発生した 51 試行目以降，提案手法と従来手法とで平均累積獲得報酬に大きい差異が見られる．これは環境変化後に行動の真の価値が大きく変化したためと考えられる．環境変化が発生する前の 50 試行までに学習者は右側の通路からゴールに到達する経路を学習する．環境変化により右側の通路は塞がれ，左側に新しい通路が出現する．これにより，学習者が 50 試行までに学習した経路上に障害が発生することになる．障害発生により，行動の真の価値が変化することになり，それまでの学習により獲得した Q 値を基に行動選択をしてもゴールには到達できなくなる．すなわち，障害発生後に Q 値を早急に修正する必要があると考えられる．提案手法では，学習者が環境変化を感知し，危険度予測法に基づいて各状態の Q 値を変化させているため，比較的スムーズに新しい経路を発見することができている．一方，従来手法では，環境変化が発生しても通常の Q 値の更新しか行わないため，新しい解を発見することが困難になっていると考えられる．

また，この実験で用いた問題環境は，環境変化により最初に通れていた経路とは逆側の経路が通れるようになるものであった．すなわち，環境変化後にそれまで Q 値が低かった行動を選択しなければ新しい解を発見することはできないということである．学習がある程度収束してくるとスタートからゴールまでの経路は絞られてくるため，Q 値の分散は大きくなってくる．そのため Q 値が最大でない行動が選択される可能性は著しく低下してしまう．提案手法では，このような Q 値の分散が大きい状態において，環境変化を感知した際に行動価値の修正を行うので，Q 値が最大でない行動も選ばれやすくなっている．その結果，提案手法が従来手法と比較してより良い経路を発見することができ，環境変化に適応することができたと考えられる．

6. まとめ

本研究では，各試行の累積獲得報酬の増減の傾向から環境変化を感知する機構を設計した．環境変化感知機構により学習者が障害発生を感知したときに，Q 値を各状態の持つ危険度に応じて修正することにより新しい環境に適応できる学習手法を提案した．また，提案手法を二次元迷路探索問題に適用し，提案手法を用いることで障害発生を伴う環境変化に適応できることを確認した．

本研究では環境変化として，それまでの最適経路上に障害が発生する場合を考慮した．しかしながら，環境変化としては，それまでの最適経路よりも更に良い経路が出現する場合も考えられる．すなわち，障害により目標状態への遷移が難しくなるのではなく，新しい解の出現により目標状態への遷移が容易になるような環境のことである．このような環境変化に適応するためには，exploration ボーナス [Sutton 90] などにより探査を促す必要があると考えられる．今後は，障害発生を伴わない環境の変化にも適応できるような学習手法へ拡張する予定である．また，障害を感知したときにすべての状態行動対に対して Q 値の修正を行ったが，必要とされる状態行動対の Q 値だけを修正することで更に学習の効率が向上すると考えられる．そのために，障害発生の感知だけではなく，障害が発生した状態の特定ができるように環境変化感知機構を改良したい．

参考文献

[Sutton 98] Richard S. Sutton, Andrew G. Barto：Reinforcement Learning, The MIT press (1998).

[木村 99] 木村 元，宮崎 和光，小林 重信：強化学習システムの設計指針，計測と制御，Vol. 38，No. 10，pp. 618–623 (1999).

[石井 02] 石井 信：強化学習におけるランダムさの自己調節，日本神経回路学会論文誌，Vol. 9，No. 4，pp. 254–262 (2002).

[小堀 05] 小堀 訓成，鈴木 健嗣，ハルトノ ピトヨ，橋本 周司：尤度情報に基づく温度分布を用いた強化学習法，人工知能学会論文誌，Vol. 20，No. 4，pp. 297–305 (2005).

[Sutton 90] Richard S. Sutton：Integrated Architectures for Learning Planning and Reacting Based on Approximating Dynamic Programming, The seventh International Conference on Machine Learning, pp. 216–224 (1990).