7Q-07

未来考慮型の信頼度に基づく合目的強化学習

Purposive reinforcement learning based on the reliability to estimate the future

有村 柊一 † 南 朱音 † 甲野 佑 † 高橋 達二 †
Shuichi Arimura Akane Minami Yu Kohno Tatsuji Takahashi

東京電機大学 †
Tokyo Denki University

1. 序論

強化学習は, 未知の環境においてエージェントが試行錯誤することで, 収益を最大化する方策の獲得を目標とする機械学習の一分野である. 近年の強化学習は, 深層学習の発展とともにAlpha-Go[1] による囲碁での超人的なパフォーマンスに代表されるように急速に発展してきている. しかし, それに伴い強化学習モデルは複雑化し, 学習のために豊富な計算リソースが必要なことが参入障壁となっている. そのような現状において, 収益の最大化を目標とする既存の強化学習アルゴリズムだけでなく, ある程度の目標を高速に達成できるような強化学習アルゴリズムの必要性が高まっている. そこで, 本研究では最適化を目的とした強化学習ではなく, 目標を満たす解を高速に獲得できる合目的強化学習に注目する.

人間の意思決定傾向を強化学習に適用し, 合目的な意思決定を可能とした例に Risk-sensitive Satisficing (RS) アルゴリズム [2] がある. 同アルゴリズムは Global Reference Conversion (GRC) によって拡張され, 状態遷移を伴う様々な強化学習タスクにおいて高速な合目的解の獲得に成功している. しかし, 現状の RS アルゴリズムにおいて信頼度は単調増加するため, 非定常環境において問題が発生する. また, 強化学習では状態と行動の系列を扱うが, 現在の信頼度は系列からの学習ができていない. そこで本研究では, 深層強化学習で用いられるリプレイメモリを利用した現在状態との照合により, 過去の経験から未来を考慮した信頼度を動的に計算する手法を考案し, 性能の検証を行った.

2. RS アルゴリズム

RS アルゴリズムで用いられる RS 価値関数は, とりうる状態空間とその要素を $s \in S$, 状態 s での行動空間とその要素を $a \in A(s)$, 任意の状態行動価値を $Q(s_i, a_j)$, 状態行動対の信頼度を $\tau(s_i, a_j)$, 各状態に対する希求水準を $\aleph(s_i)$ として定義することで以下の式で算出される.

$$\mathrm{RS}(s_i, a_j) = \tau(s_i, a_j)\Big(Q(s_i, a_j) - \aleph(s_i)\Big) \tag{1}$$

$$\tau(s_i, a_j) = \tau_{\mathrm{curr}}(s_i, a_j) + \tau_{\mathrm{post}}(s_i, a_j) \tag{2}$$

ここで行動価値関数 Q は, 強化学習の代表的アルゴリズムである Q 学習などによって更新される. 信頼度 τ は観測した状態行動をカウントした τ_{curr} と 1 つ先の状態行動の信頼度から算出される τ_{post} の和によって構成され, 時刻 t での価値関数の更新と同時にそれぞれ以下の式で更新される.

$$\tau_{\mathrm{curr}}(s_t, a_t) \leftarrow \tau_{\mathrm{curr}}(s_t, a_t) + 1 \tag{3}$$

$$\tau_{\mathrm{post}}(s_t, a_t) \leftarrow \tau_{\mathrm{post}}(s_t, a_t) + \alpha_\tau\Big(\gamma_\tau \tau(s_{t+1}, a_{t+1}) - \tau_{\mathrm{post}}(s_t, a_t)\Big) \tag{4}$$

連絡先: 高橋達二, 東京電機大学理工学部, 埼玉県比企郡鳩山町石坂, Tel: 049-296-5416, tatsujit@mail.dendai.ac.jp

ここで $\gamma_\tau \in [0, 1]$ は未来の信頼度の割引きであり, $\alpha_\tau \in [0, 1]$ は信頼度学習率である. RS アルゴリズムは式 (1) で計算される RS 価値関数が最大となるような行動を選択するアルゴリズムである.

2.1 GRC

(1) 式で登場する $\aleph(s)$ は, 状態遷移のないバンディットタスクのような環境であればそのタスクにおける希求水準である大局希求水準 $\aleph_G$ と同義であるが, 状態遷移の伴う環境では各状態ごとに基準値が必要となる. そのため, $\aleph_G$ から各状態の希求水準 $\aleph(s)$ への変換を必要とし, その代表的な手法に GRC が存在する. GRC とはタスクにおいて達成したい大局的な希求水準 $\aleph_G$ を, タスク収益の推定値である大局期待収益 E_G, スケーリングパラメータ $\zeta(s_i)$, 状態行動価値 $Q(s_i, a_i)$ を用いて以下のように変形する一連の操作を指す.

$$\delta_G = \min(0, E_G - \aleph_G) \tag{5}$$

$$\zeta(s_i)\delta_G = \max Q(s_i) - \aleph(s_i) \tag{6}$$

$$\aleph(s_i) = \max Q(s_i) - \zeta(s_i)\delta_G \tag{7}$$

ここで大局満足度 δ_G は, エージェントが希求水準を満たせていると考えていれば最大値 0 を示し, 水準未満であると考えていれば負の値を取る. また, 大局期待収益 E_G は以下の式で更新される.

$$E_G \leftarrow \frac{E_{\mathrm{tmp}} + \gamma_G E_G N_G}{1 + \gamma_G N_G} \tag{8}$$

$$N_G \leftarrow 1 + \gamma_G N_G \tag{9}$$

ここで, E_{tmp} は直前 1 エピソードの収益を表している.

3. 提案手法

RS は前述の通り希求水準, 行動価値, 信頼度によって定義されるアルゴリズムであるが, 現状の信頼度の定義はいくつかの問題を抱えている. 本節ではその問題点を示し, 解決策となる定義を提案する.

3.1 信頼度の問題点

信頼度は選択した状態行動対の回数を上限なくカウントしていくため, これまで正しかった行動が誤った行動になってしまう非定常な環境において, 間違った選択の信頼度が高いという問題が発生する. また, 現状の信頼度は 1 ステップ先の状態行動しか考慮できていないことから, 深層強化学習で広く用いられる軌跡からの学習ができておらず, サンプル効率が悪いという問題がある.

3.2 軌跡を利用した信頼度の定義

前述の問題を解決するため, 経験照合による未来考慮型の信頼度を提案する. これは深層強化学習などで用いられる手法である Experience Replay[3] に使用されるリプレイメモリと現状

の信頼度の更新であるカウントを組み合わせたものである．経験再生では各時間ステップにおけるエージェントの経験を $e_t = (s_t, a_t, r_t, s_{t+1})$ とし，リプレイメモリ $D_t = \{e_1, e_2, \ldots, e_t\}$ に格納していく．このメモリは first-in first-out のリングバッファで構成される．

経験照合型の信頼度を定義する準備としてインデックス集合 I を以下のように定める．ここで e_i^a や e_i^s のような記述はリプレイメモリ上の i 番地の経験 e_i の行動 a_i，状態 s_i をそれぞれ意味する．

$$\mathrm{I} = \{i \in \mathbb{N} | e_i^s = s_t, e_i^a = a_t, i \leq t\}$$

この集合 I に入るインデックス i は，タイムステップ t における状態行動と，リプレイメモリ上の i 番地の状態行動が一致することを意味する．未来考慮型の信頼度 $\tau(s_t, a_t)$ は経験 e_i，インデックスの集合 I を用いて以下のように算出される．

$$\tau(s_t, a_t) = \tau_{\mathrm{curr}}(s_t, a_t) + \tau_{\mathrm{post}}(s_t, a_t) \tag{10}$$

$$\tau_{\mathrm{curr}}(s_t, a_t) = |\mathrm{I}| \tag{11}$$

$$\tau_{\mathrm{post}}(s_t, a_t) = \sum_{i \in \mathrm{I}} \sum_{k=1}^{\mathrm{L}} \gamma_\tau^k \tau_{\mathrm{curr}}(e_{i+k}^s, e_{i+k}^a) \tag{12}$$

式 (11) に登場する |I| は集合 I の要素数であり，式 (12) の L は経験の先読みの数を意味し，γ_τ は経験の割引を意味する．この定義では，固定長のリプレイメモリの中に登場する回数と，現在の状態行動とリプレイメモリ上の状態行動が一致する点から L ステップ先までの信頼度の割引和から計算される．価値が低い状態行動は学習が進むにつれ選ばれにくくなりリプレイメモリから消えていくため，信頼度が無限に増加してしまうという問題を解決することができる．また，現在から L ステップ先までの信頼度を考慮できるため，軌跡からの学習ができずサンプル効率が悪いという問題も同時に解決される．

4. 実験

未来考慮型の信頼度と既存の定義による信頼度を非定常なグリッドワールドタスクで比較した．また，ベンチマークとして softmax 方策を用いた Q 学習を使用した．この Q 学習で用いた方策の温度パラメータは $\tau = 1$ とし，その他のパラメータについては後述する．グリッドワールドタスクは格子状のマスに区切られた環境であり，1 ステップで隣接する上下左右 4 マスのいずれかに確率 1 で遷移する．実験に使用したグリッドワールドの形状を以下の図 1 に示す．

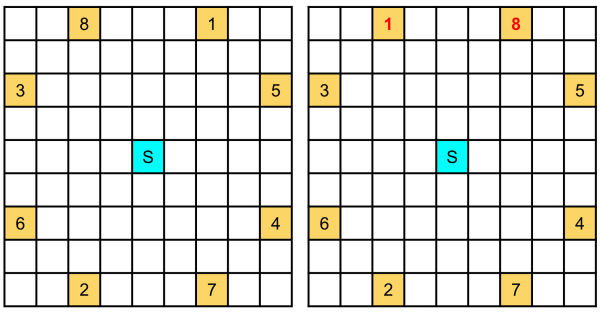

図 1: 左：初めの環境，右：1 度目の環境変化後

実験は左図の中央 S からスタートし，いずれかの数値の書かれたゴールマスに到達するとその量の報酬を受け取り，エピソードが終了する．本実験における非定常とは，1000 エピソードごとに最大の報酬と最小の報酬の座標が入れ替わることである．2000 エピソードより長いエピソード数実験を行う場合，一度図 1 右の環境になった後，再び左の環境に変化するため，エージェントには 1 度悪いと判断したゴールに再び到達する探索力が求められる．

特に断りのない場合，全手法で共通のハイパーパラメータを使用している．行動価値関数 Q の更新は Q 学習を用いて行い，そのパラメータは $\gamma = 0.9$, $\alpha = 0.1$ を用いた．また，大局希求水準 $\aleph_G = 8$，スケーリングパラメータ $\zeta(s) = 1\ \forall s \in S$，信頼度学習率 $\alpha_\tau = 0.1$，信頼度割引率 $\gamma_\tau = 0.99$，考慮する未来のステップ数 $L = 5$，信頼度のカウント部分 $\tau_{curr}(s, a) = 1\ \forall s, a$，大局期待収益の割引率 $\gamma_G = 0$，リプレイメモリのサイズ $M = 2000$ にそれぞれ設定した．

5. 結果・考察

3000 エピソードを 100 シミュレーション行った報酬変化のシミュレーション平均を図 2 左に示す．また，この結果の最初 1000 エピソード分を抜粋した結果を図 2 右に示す．

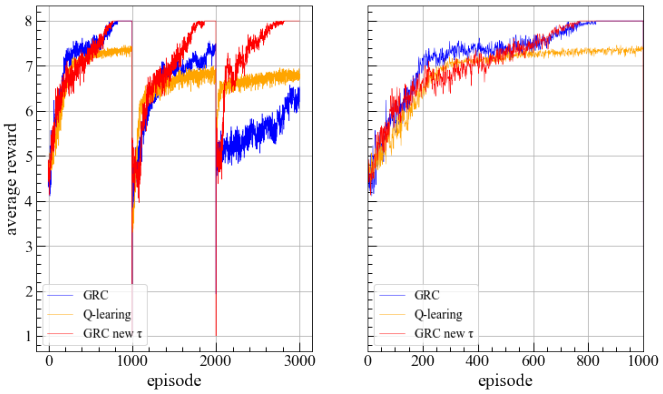

図 2: 報酬平均の時間変化

図 2 左より，環境の変化が起こる 1000 エピソード以降の性能は提案手法が既存手法を上回っている．また既存手法の 2000 エピソード以降の性能を見ると，Q 学習よりも低い性能となっており，3.1 節にて述べた既存手法での問題点を確認することができる．提案手法は 1000 エピソード以降も 2000 エピソード以降も性能が高いことから，非定常環境での追従性を獲得している．これは固定長のメモリを使用することで記憶の忘却が発生し，環境変化前の記憶を忘れられるようになったためであると考えられる．

図 2 右より，環境の変化が起こらない定常な環境での性能は既存手法が提案手法をわずかに上回っている．しかし，最高報酬である 8 への到達速度はほぼ同等である．初期の性能が既存手法よりも下回っている要因をとして，軌跡から学習する提案手法は 1 ステップ先のみから学習する既存手法と比べて，探索段階の最善でない系列の影響を受けやすいためと考えられる．

6. 結論

本稿では合目的強化学習の手法である RS アルゴリズムの課題を指摘し，改善策の提案，検証を行った．結果から，提案手法は非定常環境への追従性を獲得したうえ，定常環境でもほとんど同程度の性能を示した．今後はメモリサイズの変更による性能差の検証や，様々なタスクでの検証を行っていく．

参考文献

[1] Silver, D., Huang, A., Maddison, C. et al. Mastering the game of Go with deep neural networks and tree search. Nature 529, 484–489 (2016).

[2] 高橋 達二, 甲野 佑, 浦上 大輔：認知的満足化一限定合理性の強化学習における効用, 人工知能学会論文誌, Vol. 31, No. 6, pp. AI30-M-1-11 (2016).

[3] Long-Ji Lin : Self-improving reactive agents based on reinforcement learning, planning and teaching, Machine Learning 8, 293–321 (1992)