3P-03

満足化価値関数を用いて自律的に探索する強化学習手法

牛田　有哉 *1　甲野　佑 *2　高橋　達二 *1

*1東京電機大学理工学部　*2東京電機大学大学院

1. はじめに

我々が対応すべき環境が非常に複雑で把握しきれないこともある．しかし環境と相互作用しながら行動を学習する枠組みである強化学習では，複雑な環境になるほど最適な方策が得られるまでに時間がかかってしまう．このような問題に対して，人間は学習をし続けながらも環境の非定常性や複雑性と上手く折り合いをつけて生活しているといえる．これは人間が満足化と呼ばれる最適化とは異なるルールによって試行錯誤しているからであるとされる．本研究では満足化を実装した価値関数により，自律的に試行錯誤 (探索) の割合を調整する強化学習手法を考案して既存手法と比較する．

2. 満足化方策と強化学習

人間は意思決定において選択肢の評価をある基準に対して “満たす” と “満たさない” に離散化し，基準を満たす選択肢を見つけたらその選択肢に執着する傾向がある．このような選択傾向を満足化方策と呼ぶ．満足化方策は基準値 R というパラメータを持ち，基準値 R を上回る選択肢を見つけたら探索を打ち切ることで探索と利益追求のバランシングを行うことができる．本研究では既存の満足化モデルであり，最も単純な強化学習課題であるバンディット問題において良い成績を示している RS(reference satisficing) 価値関数 [高橋 15] のより広い強化学習課題への適用を行う．

2.1 RS

RS を強化学習に適用させるためには，バンディット問題の試行割合にあたる信頼性を方策の評価にするための新たな信頼性変数の導入と，価値の実数域化が必要となる．信頼性についてはその状態での選択肢の試行回数とその後の行動系列を考慮した信頼性変数として $\tau(s_i, a_j)$ で定義した．

$$\tau(s_i, a_j) = \tau_{current}(s_i, a_j) + \tau_{post}(s_i, a_j) \quad (1)$$

$$\tau_{current}(s_t, a_t) = \tau_{current}(s_t, a_t) + 1 \quad (2)$$

$$\tau_{post}(s_t, a_t) = \tau_{post}(s_t, a_t) + \alpha\Big(\gamma_\tau \tau(s_{t+1}, a_{up}) - \tau_{post}(s_t, a_t)\Big) \quad (3)$$

価値関数は報酬に対する価値と信頼性変数を分離させて重み付き平均で定義することで RS の性質を保ちつつ実数地域への対応を可能にした．

$$RS(s_i, a_j) = w_{ij} Q(s_i, aj) + (1 - w_{ij}) R \quad (4)$$

$$w_{ij} = \frac{\tau(s_i, a_j)}{\sum_k \tau(s_i, a_k)} \quad (5)$$

2.2 R-Timer

満足化方策の問題として基準値 R をどのように獲得するかが存在する．適切な基準値は環境によって異なるため，エージェントが自律的に動的に獲得する必要がある．そこで，報酬の受付期間を L step に引き延ばし累積した値と，報酬の不確実性を考慮した馴化を組み合わせて以下の式で基準値 R を更新するような R-timer を実装した．

$$\text{if } R(s_i) < \sum_{k=0}^{L-1} \gamma^k r_{t+k} \quad (6)$$

$$\text{then } R(s_i) = \sum_{k=0}^{L-1} \gamma^k r_{t+k} \quad (7)$$

$$R(s_i) = R(s_i) + \alpha_{acc}\Big(\max_{a_k} Q(s_i, a_k) - R(s_i)\Big) \quad (8)$$

3. 大車輪シミュレーション

本研究では複雑な環境を持つ強化学習タスクとして大車輪課題を扱った．大車輪課題とは，鉄棒運動をロボットに行わせるもので，自身を回転させるような行動の獲得を目的とする．ロボットは，腰の関節を “曲げる”，“延ばす”，“動かさない” の 3 種類の行動を取り，状態は上半身の角度を 24，上半身と下半身のなす角度を 5，上半身の角速度を 7 に等分割した 840 種である．エージェントは方策に基づいた行動を行い，1,000 step 毎に初期状態に戻る．報酬は初期状態を上半身の角度 $\theta = 0$ として，step 毎に $r = \theta/\pi$ で与えられる．

3.1 設定

RS と既存のモデルの比較をするため，シミュレーションを行った．1 回の行動決定を 1 step とし，学

An Autonomous Search Method Using Satisficing Value Function in Reinforcement Learning.

Yuya Ushida, Tatsuji Takahashi, School of Science and Technology, Tokyo Denki University.

Yu Kohno, Graduate School of Tokyo Denki University.

習時間は 200,000 step とした．比較に用いるアルゴリズムは，既存の研究で良い成績を出していた満足化方策である LS-Q アルゴリズム [Uragami 14]，自律探索アルゴリズムとして知られる VDBE アルゴリズム [Tokic 10]，最も一般的な学習アルゴリズムである Q 学習を用いる．LS-Q と Q 学習では行動選択に ϵ-greedy を用い，ϵ は 1.0 から始まり，徐々に減衰し 100,000 step の時点で 0.0 になるように設定する．RS には基準値の獲得に R-Timer を用いたものと，経験的に成績が良かった基準値 $R = 5.0$ に固定した場合を用いて比較した．

3.2 結果および考察

シミュレーションの結果として，1,000 step 毎の報酬の総和の時間発展を図 1 と図 2 に，1,000 step ごとの greedy な選択をした割合の時間発展を図 3 と図 4 に示す．図 1 と図 2 の結果から Q 学習や LS-Q は学習率 0.9 のときのみ高い成績となっており，さらに ϵ を用いて探索を促さなければ学習が行えないことがわかる．それに対して RS は学習率の値が 0.1 では高い成績を，0.9 でもある程度の成績が出せていることから共に学習ができていることがわかる．これは図 3 と図 4 の結果から，RS がパラメータの違いに合わせて探索と利益追求のバランスを上手く調節しているためであると考えられる．また，RS の中では $R = 5.0$ の場合に最もよい成績となっている．これは，仮に動的に適切な R を設定することができれば，より良い成績を出すことができることを示していると解釈できる．

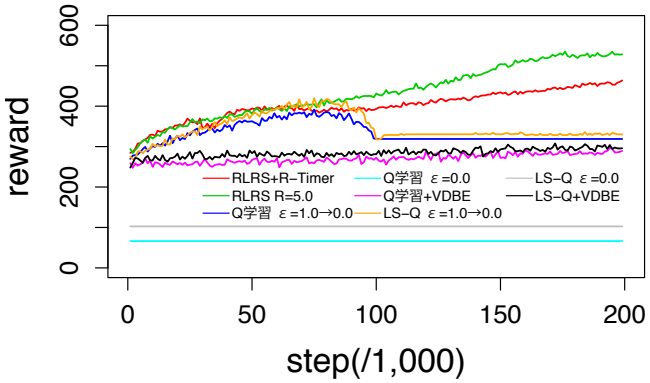

図 1: 獲得報酬の時間発展，学習率 $\alpha = 0.1$

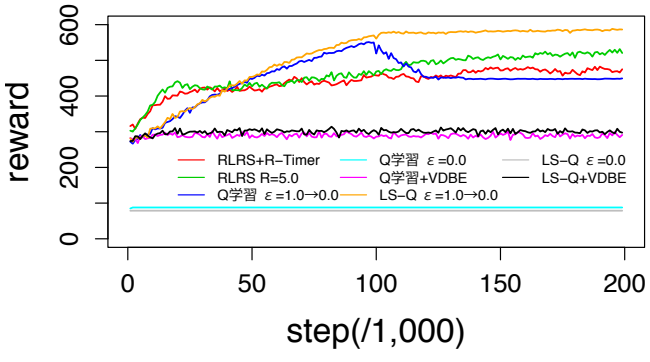

図 2: 獲得報酬の時間発展，学習率 $\alpha = 0.9$

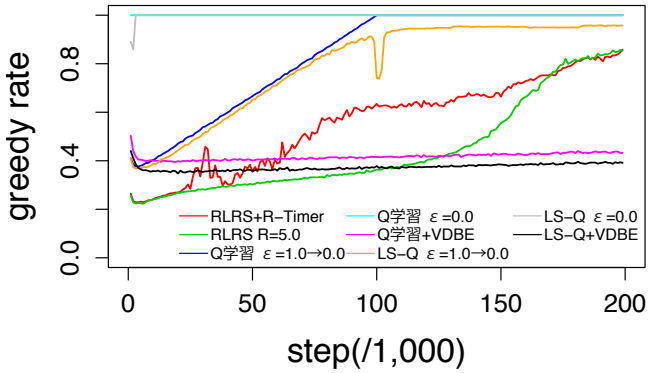

図 3: greedy 選択率の時間発展，学習率 $\alpha = 0.1$

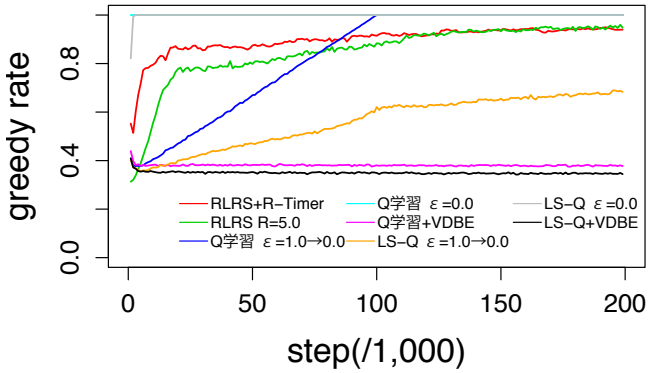

図 4: greedy 選択率の時間発展，学習率 $\alpha = 0.9$

4. 結論

本研究では，意思決定における人間の特性である満足化を強化学習へ応用した．結果から，RS は自律的に探索と利益追求のバランスを調節し，乱数を用いずに効率よく学習が行えることを示した．また，パラメータへの依存性が低いことから，より柔軟なアルゴリズムであると言える．さらに決め打ちではあるが，適切な基準値を設定することによってより良い成績をだすことができることを示した．今後は適切な基準値を動的にどのように獲得するかが課題となる．

参考文献

[Uragami 14] Uragami, D., Takahashi, T., Matsuo, Y.: Cognitively inspired reinforcement learning architecture and its application to giant-swing motion control,BioSystem, 116, 1-9(2014).

[高橋 15] 高橋達二, 甲野佑, 大用庫智, 横須賀聡：不確実性の下での満足化を通じた最適化, JSAI2015(2015 年度人工知能学会全国大会 (第 29 回))

[甲野 15] 甲野佑，高橋達二:満足化とその基準の動的な更新による強化学習の促進,JSAI2015(2015 年度人工知能学会全国大会 (第 29 回))

[Sutton 00] Sutton, R.S. and Barto, A.G.: 強化学習, 森北出版,(三上, 皆川訳) (2000).

[Tokic 10] Tokic, M : Adaptive ε-greedy Exploration in Reinforcement Learning Based on Value Differences, KI'10 Proceedings of the 33rd annual German conference on Advances in artificial intelligence,203-210.