1R-02

人の認知バイアスを応用した強化学習手法による適応的な学習について

樋口　将大　　浦上　大輔

日本大学生産工学部

1. はじめに

認知バイアスとは，人間の判断や意思決定に影響を与える認知の偏りのことで，ヒューリスティクスに関係があるとされている．本研究で取り扱う認知バイアスは称性バイアスと呼ばれ，「p ならば q」から「q ならば p」を推論する傾向性のことである．篠原らは，対称バイアスの程度を数値的に表現した LS モデルを考案し，n 本腕バンディット問題に応用して優れた成績を示すことを報告している[1]．我々の研究グループでは，LS モデルをより一般的な学習課題に応用する手法として「LS-Q」を提案し，鉄棒ロボットの運動獲得をテスト課題としてその有効性を検証してきた[2]．これまでの研究成果として，LS-Q は状態分割が粗い学習環境においても，適応して学習することが明らかになっている．一方，鉄棒ロボットの状態遷移は複雑で，学習過程を詳細に観察することには不向きであるため，LS-Q の特性には不明な点がある[3]．そこで本研究では，鉄棒ロボットより単純で解析が容易なツリーバンディット課題[4]に LS-Q を適用し，その特性を調査する．

2. LS-Q

LS-Q の学習アルゴリズムは「Q 学習」，「C-Table」，「LS モデル」の 3 つの要素から構成される[2]．それぞれについて説明する．

Q 学習　ある状態 s の下で行動 a を選択することの価値 Q(s, a) を，次の式で更新することで学習する．

$$\mathrm{Q}(s_t, a_t) \leftarrow (1-\alpha)\bullet \mathrm{Q}(s_t, a_t) + \alpha \bullet \left[r_{t+1} + \gamma \bullet max_a \mathrm{Q}\left(s_{t+1}, a\right)\right] \quad (1)$$

ここで，rは報酬で，α は学習率，γ は割引率と呼ばれる学習パラメータである．ある状態において Q(s, a)の値（Q 値）が最大になる行動を greedy な行動と呼ぶ．学習の途中では，greedy な行動とそうでない行動の両方を選択する．

C-Table　表 1 のように，Q 値の意味で greedy な行動とそうでない行動を選択した回数を状態ごとに記録したもの．行動 1 を選択してその行動が greedy な行動であった回数が a で，そうでなかった回数が b である．行動 2 を選択した場合についても同様に c, d として記録する．

表 1.　C-Table

行動	1	2
greedy	a	c
not greedy	b	d

LS モデル　表 1 において，行動 1 を選択したときにその行動が greedy である確率は，条件付き確率として次式ように計算される．

$$\mathrm{P}(\mathrm{greedy}|1) = \frac{a}{a+b} \quad (1)$$

一方，LS モデルでは，同じ事象に対する主観的な信頼度のようなものを，対称バイアスを加味して次式で計算する[1]．

$$\mathrm{LS}(\mathrm{greedy}|1) = \frac{a + b\cdot d\ /(b+d)}{a + b\cdot d/(b+d) + b + a\cdot c/(a+c)} \quad (2)$$

行動 2 を選択した場合についても同様に，次式のようになる．

$$\mathrm{LS}(\mathrm{greedy}|2) = \frac{c + b\cdot d\ /(b+d)}{c + b\cdot d/(b+d) + d + a\cdot c/(a+c)} \quad (3)$$

LS-Q では，LS(greedy|1)と LS(greedy|2) を比較して，値が大きい方の行動がより価値が高い行動と判断する．

3. ツリーバンディッド課題

学習の対象となる「ツリーバンディット課題」とは，n 本腕バンディット問題をツリー構造上に拡張し，状態遷移を追加することで遅れ報酬を発生させたものである[4]．今回扱うツリーバンディット課題は図 1 のように状態が 7 つ，行動が 1 つの状態につき 2 つ存在する．状態 1 からスタートして行動選択を 2 回行い，状態 4, 5, 6, 7 のいずれかに到達して報酬を受け取る．それを学習の 1 エピソードとして 150 エピソード繰り返し学習を行う．報酬は確率的に与え，当たり時の獲得報酬は 1 とした．当たり確率の設定は，単高設定（表 2）と拮抗設定（表 3）の 2 パターン用意した．単高設定では 40 エピソードを境に報酬を確率 1 から確率 2 に変更した．

On Adaptive Learning by Reinforcement Learning Method with Human Cognitive Bias
Higuchi Nobuhiro and Daisuke Uragami
College of Industrial Technology, Nihon University

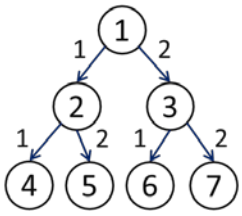

図 1. ツリーバンディッド

表 2. 報酬確率(単高設定)

状態	4	5	6	7
確率 1	0	0.1	0.5	0
確率 2	0	0.1	0	0.5

表 3. 報酬確率(拮抗設定)

状態	4	5	6	7
確率	0.3	0.4	0.5	0.2

4. シミュレーション結果

前述のツリーバンディッド課題に，LS-Q と Q 学習および CQ 学習を適用してシミュレーションを行った．ここで，CQ 学習とは，C-Table から通常の条件付き確率を計算して行動価値を決定する手法である．Q 学習と CQ 学習を比較すると C-Table の効用がわかり，LS-Q と CQ 学習を比較すると，LS モデルの効用がわかる．

図 2 は，単高設定(表 2)でのシミュレーション結果で，各手法のエピソード毎の獲得報酬を示したグラフである(1000 試行の平均値)．10 エピソードまではランダムに行動選択し，11 エピソード以降から 150 エピソードまでは greedy 法で行動選択をするようにしている．40 エピソード以前では，LS-Q と CQ 学習は，Q 学習より良い結果を示している．これは，C-Table によって確率的な報酬設定に適応しているためと思われる．報酬設定が変更される 41 エピソード以降をみると，LS-Q は CQ 学習よりも早く獲得報酬が回復していることがわかる．これは，LS モデルによって素早く評価値を切り替えることの効用であると思われる．ただし，Q 学習に比べると獲得報酬の回復が遅い．これは，C-Table の履歴が残っていることが要因であると考えられ，忘却率のようなものを導入すれば改善できると見込まれる．また，このシミュレーションでは学習率 α を 0.9 と高く設定しているため，Q 学習では以前の学習結果が改められやすくなっている．しかし，Q 学習の獲得報酬を期待値に十分に近付けるためには，学習率を小さくする必要がある．その傾向は，当たり確率の差が小さい場合，つまり問題が難しい場合により顕著になると思われる．

そこで，表 3 のように報酬確率が拮抗している場合でのシミュレーションをおこなった．今回は報酬確率の変更は行わず，150 エピソードでの獲得報酬の合計(収益)を比較した．図 3 は，学習率と収益の関係である．Q 学習の学習率 α を 0.1～0.9 の間で 0.1 刻みに変更した結果と，LS-Q と CQ 学習で学習率 α を 0.9 とした結果を記載している．Q 学習は学習率 α=0.5 のときの収益が最も大きいが，LS-Q はそれと同程度の収益であることがわかる．

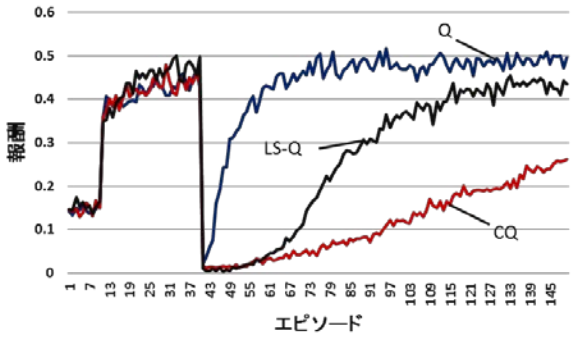

図 2. 学習曲線(単高設定)

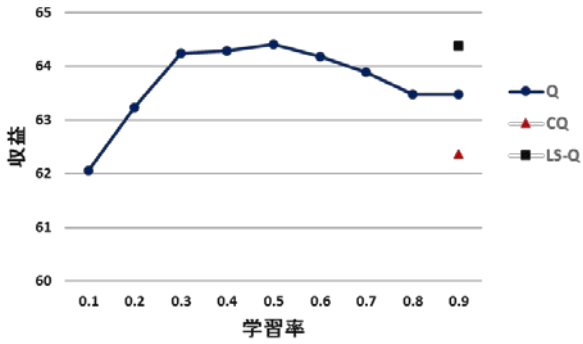

図 3. 収益(拮抗設定)

5. まとめ

本研究では，LS-Q を解析が容易なツリーバンディット課題に適用することにより、その学習能力を検証した．その結果，LS-Q は，学習の途中で報酬の設定値を変更した場合や，状態ごとの報酬の設定値の差が小さい場合においても適応して学習できることが確認された．

参考文献

[1] 篠原修二，田口亮，桂田浩一，新田恒雄: 因果性に基づく信念形成モデルと N 本腕バンディット問題への応用，人工知能学会論文誌 22 巻 1 号 G, pp.58-68, 2007.

[2] Uragami, D., Takahashi, T., Matsuo, Y., “Cognitively inspired reinforcement learning architecture and its application to giant-swing motion control,” BioSystems 116, pp. 1-9, 2014.

[3] Uragami, D., Kohno, Y., Takahashi, T., Matsuo, Y., ″Robotic Action Acquisition with Cognitive Biases in Coarse-grained State Space,” BioSystems 145, pp.41-52, 2016.

[4] 牛田有哉, 甲野佑, 浦上大輔, 高橋達二, "探索割合を自律調節する強化学習手法," 2016 年度人工知能学会全国大会, 1M2-3, 2016.