1P2-12in

認知バイアス調整機構 LS の Q 学習への実装とその機能

Implementation to Q-Learning of cognitive bias adjustment mechanism LS and the function

清水　隆宏[*1]　横川　純貴[*1]　甲野　佑[*2]　高橋　達二[*1]
Takahiro Shimizu　Junki Yokokawa　Yu Kohno　Tatsuji Takahashi

[*1]東京電機大学　Tokyo Denki University
[*2]東京電機大学大学院　Graduate School of Tokyo Denki University

Loosely symmetric (LS) model is a probabilistic formula defined on a 2×2 contingency table, proposed by Shuji Shinohara in 2007. The model represents nonlogical cognitive biases such as symmetry and mutual exclusivity. It shows the highest correlation with human causal induction data and performs very well in n-armed bandit problems, word learning and game theoretic decision making. In this study, exploiting LS for state-action value calculation (Q values in Q-learning), we make a step toward showing the efficacy in reinforcement learning in general.

1. 概要

本研究は篠原が提唱した緩い対称性モデル (lossely symmetric (LS) model) [篠原 07] を強化学習に用いる事を目的としている．LS は対称性・相互排他性といった人間の非論理的な認知バイアスを持つ価値関数の一種である．そのため LS を強化学習における行動価値の評価に用いることで，方策レベルでの確率的工夫やパラメータチューニングなどの複雑な処理を行わずに，非論理的な認知バイアスを用いた学習を行う事が可能になると考えられる．

本研究では強化学習への応用として強化学習における行動価値関数に LS モデルを適用し，一次元追跡問題のシミュレーションを行う．その結果を通常の Q 学習と比較することで LS モデルを強化学習への適用することの有効性を検証する．

2. 強化学習

強化学習とはエージェントが環境に対して試行錯誤を行い学習を行う枠組みである．教師あり学習とは異なり，ある状態毎に対する正しい行動を明示的に示す教師が存在しない．強化学習ではエージェントの試行によって環境から報酬が与えられる．エージェントは得られた報酬から行動価値を更新し学習を行う．環境から与えられる報酬はノイズが含まれていたり，遅延が存在し，一般的にある時点で得られた報酬だけではその行動が正しいか否かを絶対的に判別できない [Sutton 00].

2.1 Q 学習

Q 学習とは結果との誤差によって学習を行う TD 学習の一種である． Q 学習は各状態において，可能な行動の中から以下に示される行動価値関数において最も値が高い行動を取るように学習を行う．

$$\cdot Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a} Q(s',a) - Q(s,a)] \quad (1)$$

ただし，s は現在の状態，a は行動，s' は遷移後の状態，α，γ は学習率，割引率である．r は報酬の値である. この更新式を用いて，行動を起こした後に更新する．

連絡先: 清水隆宏 e-mail:tkhr.shmz.aaio.iiu[at]gmail.com
東京電機大学理工学理工学科情報システムデザイン学系

3. Loosely symmetric model

緩い対称性モデル (LS) とは，人間の因果帰納等に存在する“ 対称性バイアス ”および“ 相互対称性バイアス ”という 2 つの非論理的な認知バイアスを緩やかに持つ確率的な確信度のモデルである．原因となる事象 p と結果となる事象 q があるとき，対称性バイアスは $p \rightarrow q$ という情報からから $q \rightarrow p$ を導き，相互排他性バイアスは $p \rightarrow q$ から $\bar{p} \rightarrow \bar{q}$ を導く傾向を表す．これらは論理学において逆と裏の関係に当たり論理的には誤りである．しかし人間は因果帰納において度々このような推論を行う事が知られている．しかし人間が常にこのようなバイアスを働かせているとは考え難い．LS は地の不変性などを用いてこれらバイアスを柔軟に変化させる事により人間の因果帰納実験に対して高い相関を持つ事が示されている．[Takahashi 10] 本研究では $p \wedge q$, $p \wedge \bar{q}$, $\bar{p} \wedge q$, $\bar{p} \wedge \bar{q}$ の共起頻度をそれぞれ表 1 上の変数 a, b, c, d で表し，LS は変数を用いて式 (2) のように表される．[篠原 07]

表 1: 共変情報

	q	$\bar{q}$
p	a	b
$\bar{p}$	c	d

a：原因 p，結果 q が共に生起した頻度
b：原因 p のみが生起した頻度
c：結果 q のみが生起した頻度
d：原因 p，結果 q が共に生起しなかった頻度

$$LS(q|p) = \frac{a + \frac{b}{b+d}d}{a + b + \frac{a}{a+c}c + \frac{b}{b+d}d} \quad (2)$$

4. 一次元追跡問題

本研究では追跡問題を一次元に簡素化し，より純粋な問題として扱っている．問題の舞台となる空間は x 軸の長さが 31 マスで右端と左端が壁がった環状の一次元空間として定義する．空間内にはエサとなる逃亡者と，それを追跡する追跡者（ハンター）がそれぞれ 1 体が存在する．本研究ではハンターのみを学習エージェントとして扱っており，ハンターには 1 回行動する毎に -1 を報酬として与える．ただし，ハンターがエサと重なり合った（捕獲した）場合の報酬は 0 とする．ハンターとエサは“ その場に止まる ”と“ 右へ進む ”の 2 つの行動を選択できる．

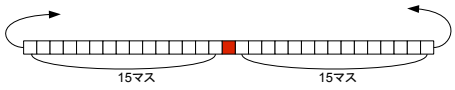

図 1: 一次元空間

4.1 エージェントの設定

ハンターは 2 つの状態をもつ．状態はハンターとエサの位置によって決める．ハンターから見て右側 15 マス内にエサが存在する状態を S_0，左側 15 マス以内にエサが存在する状態 S_1 と表記する．これは一方通行の環状な 31 マスの道を，エサが自分の前後どちらにいるか確認できる状況で追跡を行っている状態に等しい．また，ハンターとエサが重なり合った状態を状態 S_0，状態 S_1 に含まず，その直前の状態を次の状態として扱っている．ハンターの行動は“ その場に止まる ”行動を A_0 とし，“ 右へ進む ”行動を A_1 とする．ハンターが行動し，報酬を得て状態を遷移するところまでを 1 step とする．ハンターは行動価値関数の値を 2×2 分割表として保持し，この表の値を行動から得られた報酬により更新させて学習を行う．行動選択には greedy 法を用いておりエージェントは行動価値関数の値が高い行動を選択する．

4.2 逃亡者の設定

エサはハンターと同一の行動を行う事ができ，学習を行わず一定の確率的で行動を選択する．エサが“ その場に止まる“ を選択する確率 p,“ 右へ進む ”を選択する確率は $1-p$ と定義する．

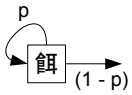

図 2: エサの行動

5. LS の実装

本研究では，この LS モデルを強化学習へ実装するにあたり，行動価値関数の値を 2×2 分割表を表 2 のように分割することで式 (2) に適用させる．

$$\cdot Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{LS} Q(s',a) - Q(s,a)] \quad (3)$$

s は現在の状態，a は行動，s' は遷移後の状態，α，γ は学習率，割引率である．r は報酬の値である．式 (3) では式 (1) での max_a を max_{LS} とし，Q 値の選択に LS 値を用いている．

$max_{LS}Q(s',a)$ を決定するため，2×2 分割表から算出した $LS(A_0|S')$ 値と $LS(A_1|S')$ 値を比較する．LS 値が最も高い行動を状態 s' において最大の Q 値を持つ行動 a として選択し，$max_{LS}Q(s',a)$ としている．

表 2: エージェントが保持する共起情報

	A_0	A_1
S_0	a	b
S_1	c	d

a：状態 S_0 の時に移動しなかった際の報酬
b：状態 S_0 の時に移動した際の報酬
c：状態 S_1 の時に移動しなかった際の報酬
b：状態 S_1 の時に移動した際の報酬

6. シミュレーション設定

LS を実装したエージェントと，従来の Q 学習エージェントを比較するため，前節で説明した一次元追跡問題を用いて終了条件のパターンとエサの行動確率 ($p = 1.0$，$p = 0.3$) の違いから 4 種のシミュレーションを行った．パターン毎に定義した終了条件を満たすまでを 1episode とし，episode 毎にハンターとエサの初期位置はランダムに決めた．学習率を 0.9 ，割引率を 0.9 とし，5000 episode までを 1 試行とした．それぞれのシミュレーション設定で 100 試行し，ハンターの保持する表は施行毎に -1 で初期化した．

6.1 パターン 1

パターン 1 での episode の終了条件はハンターとエサの座標が重なり合った（捕獲した）場合とした．これによりハンターがエサを捕獲する回数は episode 毎に一回に限定され，捕獲するまでの早さ，step 数を観測している．

6.2 パターン 2

パターン 2 での episode の終了条件を step 数が 200 回に達した場合とした．パターン 1 と異なりハンターがエサを捕獲する回数を限定していない．ある一定の時間内，step 数における捕獲回数を観測している．

7. 結果

エサの行動選択確率が $p = 1.0$，$p = 0.3$ の場合に分けて，episode 毎の step 数，または捕獲回数の平均を取ったシミュレーション結果を以下の図に示す．

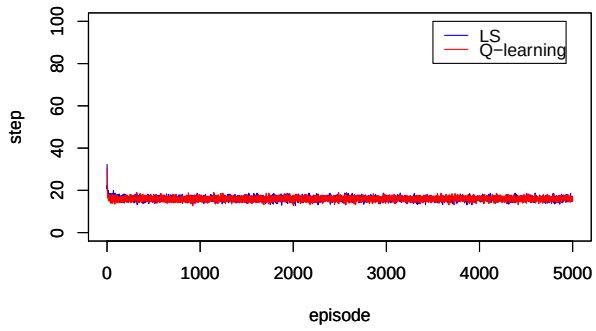

図 3: $p = 1.0$, パターン 1 においての捕獲に掛かる step 数

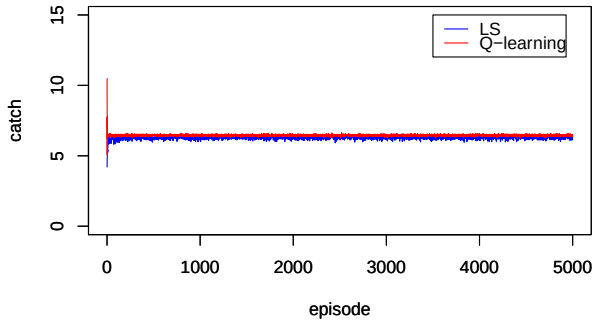

図 4: $p = 1.0$, パターン 2 においての捕獲数

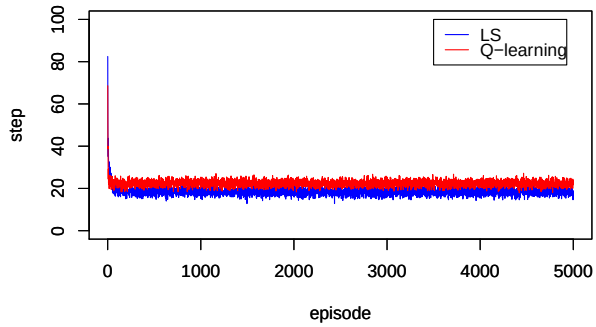

図 5: $p = 0.3$, パターン 1 においての捕獲に掛かる step 数

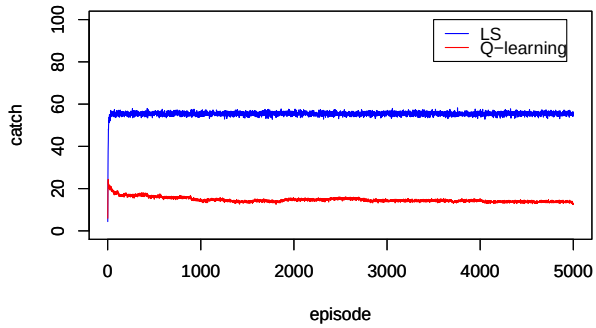

図 6: $p = 0.3$, パターン 2 においての捕獲数

エサの行動選択率 $p = 1.0$ においてパターン 1 のシミュレーションでは LS を適用したモデルは Q 学習モデルとほぼ等しい結果を得た．また，パターン 2 のシミュレーションでの学習初期の成績では LS を適用したモデルの成績は Q 学習モデルの成績より悪くなっている．

しかし，エサの行動選択確率 $p = 0.3$ においてパターン 1, 2 のシミュレーション共に LS を適用したモデルは そうでない Q 学習モデルよりも良い結果を得られた．特にパターン 2 では Q 学習のモデルの成績が変わらないのに対し，LS は捕獲回数を大きく上昇させている．これは LS が非決定的なエサの挙動に上手く適応している事を示している．このような不確実な状況で Q 学習は全ての行動を十分に探索できずに成績を上げられなかったのに対し，LS を適用したモデルでは認知バイアスを持つために通常の Q 学習モデルよりも多く探索を行ったためにより良い成績を得られたのだと考えられる．

8. 考察

本シミュレーションは LS を Q 学習に適応するための初期段階として単純な問題設定を用いたため，Q 学習における LS の有用性を決定的に述べるには今後設定を拡張し，より一般的な枠組みでシミュレーションを行う必要がある．しかし，少なくとも本シミュレーションにおいて LS モデルを非決定的な環境における強化学習に利用することが有用であるとする結果が得られた．LS は強化学習課題の一つである（ただし行動選択肢が 2 つのみで，状態は持たない）バンディット問題において複雑な状況で上手く行動を選択する事が示されている [大用 10]．それらと本研究の結果を交え，行動価値関数としての LS の有用性を検証していく事で，LS によってより柔軟で複雑適応的な強化学習を行う事が可能になると思われる．

参考文献

[Sutton 00] Richard S. Sutton., Andrew G. Barto: 強化学習. 森北出版株式会社. (翻訳) 三上 貞芳, 皆川 雅章

[篠原 07] 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適用，人工知能学会論文誌，Vol.22，No.1，pp.58-68 (2007)

[Takahashi 10] Takahashi, T. Nakano, M. and Shinohara, S.: Cognitive Symmetry:Illogical but Rational Biases, *Symmetry:Culture and Science* 21, 1-3, 275-294. (2010)

[櫻井 08] 櫻井祐輔：マルチエージェント強化学習による協調性獲得の検証－追跡問題を例として－，高知大学大学院理学研究化数理情報科学専攻修士論文 (2008)

[大用 10] 大用庫智, 高橋達二.: 因果推論と意思決定を結ぶ緩い対称モデル, 日本認知科学会第 27 回大会発表論文集, 799-800. (JCSS2010)