3P-09

習慣学習のための強化学習モデル

水戸　亜友美 *1　甲野　佑 *2　太田　宏之 *3　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　*3防衛医科大学校生理学講座

1. はじめに

我々の行動は目的達成に向かった習慣的行動, すなわち連続した動作系列として自動的に行われている. 例えばドアを開ける程度であれば, 学習済みの “(ドアを開ける一連の) 動作コマンド” を実行し, 次の動作決定まで持続する. しかし, 合目的的行動の学習を担う既存の強化学習モデルでは非常に細かな時間幅での動作決定が一般的である. 強化学習・習慣学習には大脳基底核・線条体が関与している. 最近, 線条体は細かな動作を統合し, 大脳皮質と比べて長い時間幅で機能する事が判ってきた. これによって細かな動作単位を束ねて習慣を学習している可能性が高い. そこで本研究では持続する機構によって細かな動作単位を束ねる強化学習モデルを提案する.

2. 強化学習と脳

強化学習とは報酬を獲得するための試行錯誤的な動物の行動学習に端を発した枠組みである. 強化学習課題には現在の状態における行動の評価 (行動価値関数) に基づき行動し, その行動評価を意味する TD 誤差 (Temporal Difference Error) を用いて学習する TD 学習と呼ばれる手法が一般的に用いられる [Sutton 00]. TD 学習は単なるアルゴリズムとしてだけでなく, 学習中の大脳基底核の挙動 [Schultz 95] が TD 誤差に類似することなどから, 大脳基底核と強化学習のモデルの基礎ともなっている [Houk 95]. また, 線条体の一部が行動価値関数を表現しているとの報告があり, 大脳基底核の挙動を理解する目的で強化学習理論が参照されている. それに対して, 線条体は入力に対する受付時間が長いという特徴が新たに発見された [太田 14]. そのため線条体は皮質から入力される動作単位＋細かい感覚とドーパミン (報酬) を, その長い受付時間によって報酬獲得のスイッチとなる感覚, 運動情報群としてまとめている (習慣形成) のではないかと考えられる. 本研究ではこの新たな知見に基づき, 習慣行動の学習という観点から基礎的な強化学習モデルを考案する.

Reinforcement Learning Model of Habit Learning.

Ayumi Mito, Tatsuji Takahashi, School of Science and Technology, Tokyo Denki University.

Yu Khono, Graduate School of Tokyo Denki University.

Hiroyuki Ohta, National Defense Medical College, Department of Physiology.

3. 習慣の学習と時系列の圧縮抽出

本研究で扱う習慣とは, 例えば部屋から出たい場合, その場からドアまで歩き, ドアを開けて出て行く, つまり部屋を出ようと思った時から, 出るまでの無意識的な一連の動作の事を指す. ラットの実験において, 習慣形成前では課題中ほとんど常に線条体ニューロンが活発に活動するのだが, 習慣形成後では行動の開始時と報酬獲得時のみの活動する. すなわち形成された習慣行動中は状態観測や意思決定に注意が向いていない事が知られている [Smith 13]. つまり習慣の獲得のためには, 一連の動作とその開始条件 (スイッチ) となる状態を報酬によって適切に関連づける必要がある. そのために報酬を獲得した (状態と行動で記述される非同期な) 一連の時系列から関係ない状態を排除して整理できなくてはならない. そこで本研究では神経生理に習った時系列から必要な要素を圧縮抽出という形式で習慣の学習を行う.

4. 時系列圧縮抽出アルゴリズム

本研究では原因状態をスイッチとした行動系列の実行 (習慣学習) を獲得するアルゴリズムとして Habit Former 1.0 を考案した. 習慣行動の獲得には既存の強化学習手法のような細かな時間幅で頻繁な動作決定ではなく, 束ねた大まかな動作コマンドの学習を目指すべきであり, 前述した通りそのためには前述の時系列データの圧縮・抽出が必要となる.

Habit Former 1.0 は大まかな動作コマンドを継続して同じ行動を取り続ける慣性的な行動として扱いう. 習慣は報酬の獲得をきっかけとして形成されていき, 習慣行動テーブルに保存されていく. 非習慣行動中に習慣行動のスイッチ (開始条件) となっている状態を観測した場合, その習慣行動と一致する行動価値関数 Q 値を予測報酬としてエージェントに与える. 最終的に観測された報酬を与えられる状態において, 予測報酬と実際に与えられる報酬の差を各行動価値関数に分配する. このような報酬の出現は過去の生理実験の結果とも符合する [Schultz 95]. ここで行動価値関数 (Q 値) への報酬の分配は Q-Timer 1.0 [太田 14] を用いて行う. Q-Timer 1.0 とは状態行動対の訪問時に起動されるタイマーに基づいた非同期的な行動価

値関数の更新を行う強化学習モデルである．必要のない状態と行動を圧縮するため，予測報酬を獲得した際に，その先の習慣行動と現在の行動が一致する場合，慣性的な行動として習慣行動テーブルにスイッチとなる状態を前倒しした習慣行動を上書きしていく．ここで一致しなかった場合は新たにその中継地点となるスイッチ状態とその際に行った行動を記録する．習慣行動テーブルは推移先と行動，その状態で習慣がすでに形成されているかどうかを保持しており，形成されていれば推移先となる状態に到達するまで慣性的に行動し続ける．ただし，ϵ-greedy 方策を用いて確率 ϵ でランダムな探索行動を併用する等，習慣行動中にそれから外れるような行動を取る事もある．

5. シミュレーション

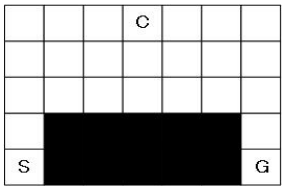

図 1: 条件付き崖歩き

崖歩き課題と条件付き崖歩き課題でシミュレーションを行った．本シミュレーションにおいてエージェントが取りうる行動は，a_N:上へ行く，a_W:左へ行く，a_S:下へ行く，a_E:右へ行くの 4 種類である．初期状態 (S) から始まり，崖歩き課題ではゴール (G) に着いた場合，条件付き崖歩き課題では，報酬条件状態 (C) を通ってからゴールに着いた場合にのみ報酬 (100/ ゴールまでに掛かった step 数) を与える．条件付き崖歩き課題はどの状態が報酬の条件なのか認識できないため，通常の TD 学習では困難な非マルコフ課題に分類される．どちらも (黒く塗りつぶした) 崖領域に落ちた場合，ペナルティとしてゴールまでかかった時間を 100 増加させる．シミュレーションは全 1,000 エピソードを 1,000 回行い，平均を取った．また習慣の形成の開始は 500 エピソード以降とした．比較のために Habit Former 1.0 に加えて，Sarsa，Sarsa($\lambda = 0.9$)，Q-Timer でシミュレーションを行った．各エージェントは学習率 $\alpha = 0.05$，割引率 $\gamma = 0.9$ のパラメータで学習を行う．行動の決定には ϵ-greedy 方策，ランダム行動率 $\epsilon = 0.1$ を用いた．

5.1 結果および考察

評価には，獲得報酬と，どれだけ習慣に基づいた行動を取ったかという習慣行動使用率を用いる．習慣行動使用率の図 3，5 は黒に近い程，その後の一貫した動作 (上下左右) を引き起こすスイッチとなっている事を表しており，形成された習慣がどの状態と結びついているかを意味する．図 3 から，曲がる部分が濃くなっており，間の行動はほとんど無視されているので

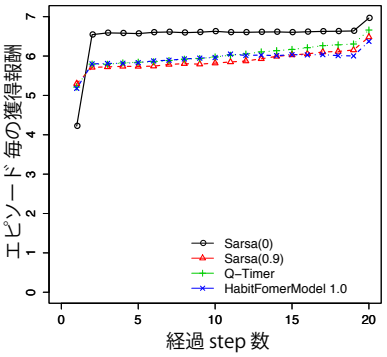

図 2: 崖歩きでの獲得報酬

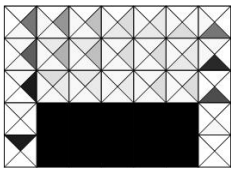

図 3: 崖歩きでの習慣行動使用率

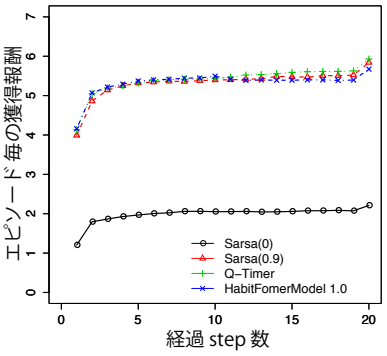

図 4: 条件付け崖歩きでの獲得報酬

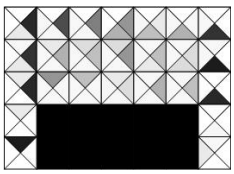

図 5: 条件付き崖歩きでの習慣行動使用率

習慣の形成が行われていることがわかる．また図 4 から，条件付き崖歩き課題のような非マルコフ環境でも学習できることが示された．

6. 結論

本研究では習慣学習のために時系列を圧縮抽出して，記憶・実行する Habit Former 1.0 を考案した．これにより従来細かな意思決定で記述されてきた行動の獲得を，習慣という比較的長い間隔で捉える事が可能になった．それにより習慣行動の組み合わせ等を行うシステムと併用して高度な学習を創発するような手法の考案に寄与すると思われる．

参考文献

[Houk 95] Houk, J.C., Adams, J.L., Barto, A.G.: A model of how the basal ganglia generate and use neural signals that predict reinforcement, In Models of Information Processing in the Basal Ganglia, Houk, J.C., Davis, J.L., Beiser, D.G., Eds. MIT Press, 215-232. (1995).

[太田 14] 太田 宏之, 甲野 佑, 高橋 達二: 線条体ニューロンの持続的発火と強化学習, JSAI 2014(2014 年度人工知能学会全国大会 (第 28 回)) 予稿集, 2N5-OS-03b-4. (2014).

[Schultz 95] Schultz, W., Romo, R., Ljungberg, T., Mirenowiczm, J., Hollerman, J.R., Dickinson, A.: Reward-related signals carried by dopamine neurons, In Models of Information Processing in the Basal Ganglia, Houk, J.C., Davis, J.L., Beiser, D.G., Eds. MIT Press, 233-248. (1995).

[Smith 13] Smith, K.S., Graybeil, A.M.: A Dual Operator View of Habitual Behavior Reflecting Cortical and Striatal Dynamics, Neuron, Vol.79, No.2, 361-374. (2013).

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