5M-05

神経生理に触発された仮想報酬関数を自己改造する強化学習

水戸　亜友美 *1　甲野　佑 *2　太田　宏之 *3　笹川　隆史 *2

*1東京電機大学大学院　*2東京電機大学理工学部　*3防衛医大学校生理学講座

1. はじめに

動物が時間的にスパースな手掛かりに基づいて行動する場合，中長期的なサブゴールの自律的な設定あるいは内発的な動機の維持が必要となり，この学習を目的にエージェント内部の時間ステップ単位を超えた中長期的なサブゴールの設定，及びそのサブゴールに対して仮想報酬を付与する強化学習手法が検討されている [甲野 16]．既に視覚的な手掛かりを元にサブゴールを設定した上で探索する方法が提案されている [Kulkarni 16] が，本研究では線条体ニューロンに関する知見からエージェント内部に異なる時間スケールで動作する異なる二つのレベルの状態を想定し，付随する仮想報酬関数 (遷移確率) を修正することで内発的動機とその遷移関係を形成する手法を考案した．

2. Habit-Former 3.0

内発的な動機 (習慣) の形成を行うアルゴリズムとして，学習した行動系列のある範囲を一つの長期的行動にまとめ，それを開始する状態 (条件状態) に仮想報酬を設ける事で長期的行動の連鎖的誘発を学習する Habit-Former 1.0 (HF1) が存在する [甲野 16]．

本研究では仮想報酬関数の生成と条件状態への付加を応用し，エージェントの主観的な状態認識 (動機，内部状態) の遷移を形成する Habit-Former 3.0 (HF3) を考案した．内部状態 s_I はそれぞれに存在する固有の方策 π_{s_I} が存在し，階層型強化学習でのサブゴール信号に類似しているが，主観的ゆえ学習過渡期では濫造され得るため，その内部状態が必要だという判断＝習慣形成を管理しなければならない，HF3 の根幹は，(1)2 種の状態表現と (2) それに付随する時間感覚と記憶，(3) 遷移確率 (仮想報酬由来) の学習，(4) 習慣化の吟味，にある．本研究では内部状態の遷移関係の獲得を目的とするため方策の学習については扱わない．

2.1 状態モデルと報酬信号の扱い

エージェントはセンサーから得られる外界の認識 s_E と獲得された内部状態 s_I の対によって状態 $s = \{s_I, s_E\}$ を定義する．生成された内部状態 s_{Ik} は遷移の原因となる条件外部状態 s_{Ec} を観測した時に任意の内部状態 s_{Ix} からの遷移確率 $P(s_{Ik}; s_{Ik}, s_{Ec})$ で遷移し，それらの変数と習慣熟練度 h, 仮想報酬 v，報酬源のモデル $R(s_{Ej}, a_i)$ を有する．観測された条件外部状態 s_{Ec} に対する遷移対象の内部状態が複数存在する場合は，それぞれの非遷移確率 $(1-P)$ が等しくなるようにして非遷移の割合も含めたルーレット選択を行うため，遷移確率と呼ぶ変数は厳密な意味での確率ではない．また遷移ルールを持たない中性内部状態 $s_{I\mathrm{null}}$ が存在し，未知の環境は必ず中性内部状態 $s_{I\mathrm{null}}$ から始まる．HF3 では正の報酬しか扱わず，エージェントに対する罰は報酬源での報酬の不在，あるいは最短に対する時間超過分の割引によって表現される．仮想報酬は内部状態遷移時に発生し，仮想報酬と実報酬 r の和である複合報酬 $c = r + v$ は遷移時に $c > 0$ となるため，HF3 における報酬とは “行動の強化” だけでなく “内部状態遷移の証拠” でもある．

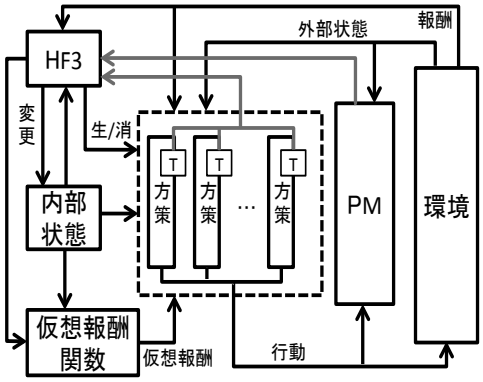

図 1: Habit-Former 3.0 の概略図

外部状態と内部状態の遷移は非同期であるため，外部状態行動対の時系列を利用した内部状態の分離と内部状態遷移に基づく遷移確率の学習には異なる二種の時間感覚と記憶が必要になる．そこで本研究では外部状態行動対の時系列を管理する方策記憶 Policy Memory と 内部状態に対する遷移後の持続的な報酬観測を司る Internal State-Timer を考案する．

2.2 方策記憶と内部状態の新生分離

今までの外部状態行動対の時系列を保存する方策記憶 Policy Memory (PM) は 最長 L_P の状態行動対 $\{s_{Et}, a_t\}$ を記憶 $H_t(k) = \{s_{Et-k}, a_{t-k}\}$ に保存する．

Reinforcement Learning Method to Self-remodel Virtual Reference Function Inspired by Neurophysiology.
Ayumi Mito, Graduate School of Tokyo Denki University.
Yu Kohno, Takafumi Sasakawa, School of Science and Technology, Tokyo Denki University.
Hiroyuki Ohta, National Defense Medical College.

H_t は記憶された外部状態行動対の数が L_P になると最も古い外部状態行動対を削除する．また PM は方策の一貫する範囲を保存するため観測した外部状態行動対 (s_{Et}, a_t) が以前の系列の同様の状態に対して $s_{Et-k} = s_{Et}$ かつ $a_{Et-k} \neq a_t$ だった時，記憶 H からステップ $t-k$ 以前の状態行動対を削除する．

エージェントが現内部状態の既知の報酬源以外で複合報酬 $c > 0$ を観測した時，新たな内部状態 $s_{I\text{new}}$ を生成する．その際，PM が有する最も古い外部状態の記憶が新生内部状態 $s_{I\text{new}}$ への遷移の条件内部状態 s_{Ec} になり，遷移の度に記憶 H_t は全てリセットされる．複合報酬はその際に取った状態行動から $R(s_{Et}, a_t)$ として学習され，仮想報酬関数はその期待値と減衰率 β を用いて $v = \beta E(R(s_{Et}, a_t))$ とする．

2.3 内部状態報酬累積記憶と内部状態遷移の学習

遷移後に出現する報酬を観測累積するため，内部状態毎にタイマー $T(s_I)$ と貯蔵累積報酬 $U(s_I)$ と事前状態記憶 $b(s_I)$ を持たせ，これを内部状態報酬累積記憶 Internal State-Timer (IST) と呼ぶ．エージェントがある内部状態 s_{It} に遷移した際， IST はタイマー $T(s_{It})$ を最大持続数 L_I にセットし，$U(s_I)$ を 0 に初期化，事前状態記憶 $b(s_{It})$ を s_{It-1} とする．その後，内部状態遷移の度に $T(s_{Ik}) \geq 1$ である全ての内部状態に対する $T(s_{Ik})$ を 1 減らし，その際に得られる実報酬も貯蔵する (割引率 $\gamma = 0.99$).

$$U(s_{Ik}) \quad \leftarrow \quad \gamma^{L_P - T(s_{Ik})} U(s_{Ik}) + r_t \qquad (1)$$

タイマーが 0 になった時，あるいは遷移が途切れた時に，$U(s_{Ik}) > 0$ である場合は該当する内部状態 s_{Ik} へ遷移した際の遷移確率 $P(s_{Ik}; s_{Ec}, b(s_{Ik}))$ を強め，逆に $U(s_{Ik}) = 0$ である場合にはそれを弱めることで遷移確率を学習していく．

2.4 内部状態の吟味

内部状態が新設される際には，中性内部状態以外の内部状態の中で習慣熟練度 h が低いものが削除される．本研究では習慣熟練度 h は実報酬に由来して生成された場合 $h \leftarrow 10$，仮想報酬に由来では $h \leftarrow 1$, 遷移の度に $h \leftarrow h\gamma_h$, 遷移が成功する (遷移確率が上がる) 度に $h \leftarrow h+1$ とし，習慣熟練度が十分高まることを習慣化と呼んでいる．以上の HF3 の性質はマウスの行動の習慣化 [Smith 07] や大脳基底核における段階的な行動学習の高次化 [Kulkarni 16] に習った．

3. シミュレーションの設定と結果

HF3 が内部状態とその遷移関係を獲得できる事を検証するため，格子空間上で周囲 8 マスの壁 (遷移不可) の有無のみで状態認識を行う部分観測三部屋課題でシミュレーションを行った．課題空間は図 2 の通りであり，ゴールに達した時のみ報酬 $r = 1$ が与えられるエピソードタスクである．中性内部状態時は常に環境が有する最適方策を取り，内部状態に付随する

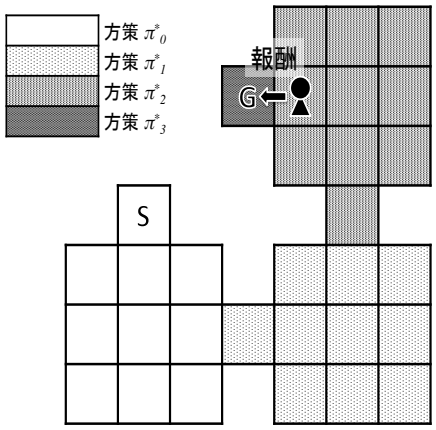

図 2: 部分観測三部屋課題

方策は内部状態を分離した際に，その部屋での最適方策 (図 2 で示される環境が有する内部状態とそれに付随する方策) が与えられる．シミュレーションの結果，PM の 長さ 3，IST タイマーについては長さ 10 遷移確率の初期値は常に 0.5，$U(s_{Ik}) > 0$ である場合に $b(s_{It})$ に対する確率を 0.9 上昇させ，それ以外を 0.1 減少させる (遷移確率はいかなる場合でも最小値 0.05，最大値は 0.95 とした) 場合に，環境が有する方策を $\pi^*_0 \rightarrow \pi^*_1 \rightarrow \pi^*_2$ で使用し，その遷移関係をエピソードを経ても維持できる事を確認した．

4. 結論

本研究では強い条件下で HF3 が実報酬から内部状態を自発的に獲得し，それに付随する仮想報酬関数に誘発されて連鎖的に新たな内部状態の生成とその遷移関係を学習できる事を示した．実用的な行動学習アルゴリズムとするにはより定量的な評価の他に方策の獲得，改善，連続タスクへの対応を行う必要がある．PM の性質上，非効率でも目的達成する方策を得る事は可能だが，その改善は遷移確率を同時に学習する都合上，非常に困難である．その効率化やモデル，各種パラメータの妥当性に関しては理論的解析だけでなく生理学的知見にも求めていくべきだと考えられる．

参考文献

[甲野 16] 甲野佑, 水戸亜友美, 太田宏之, 高橋達二, 笹川隆史: 線条体の動作に触発された習慣形成の強化学習モデル, JSAI 2016(2016 年度人工知能学会全国大会 (第 30 回)) 予稿集, 2N5-OS-03b-4. (2016).

[Smith 07] 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).

[Kulkarni 16] Kulkarni,D.T., Narasimhan,R.K., Saeedi,A., Tenenbaum,B.J.:Hierarchical Deep Reinforcement Learning: Integrating Temporal Abstraction and Intrinsic Motivation,(2016).

[Kulkarni 16] Yin,H.H., Knowlton,B.J. : The role of the basal ganglia in habit formation. Nature Reviews Neuroscience, 7(6), 464-476,(2006).