The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

階層型強化学習における人間のサブゴール知識転移

Human Sub-goal Transfer in Hierarchical Reinforcement Learning

奥戸　嵩登 *1*2　　山田　誠二 *2*1
Takato Okudo　　Seiji Yamada

*1総合研究大学院大学　　*2国立情報学研究所
SOKENDAI　　National Institute of Informatics

Hierarchical reinforcement learning, especially which learn policy with option discovery simultaneously, needs a lot of iterations. This paper investigates how human sub-goal transfer affect to learning speed and performance. we proposes the way to transfer human sub-goals in hierarchical reinforcement learning. To acquire human sub-goal knowledge, we use the problem in interactive machine learning. Supervised learning transforms human sub-goals into initial parameters before learning on hierarchical reinforcement learning. Two experiments, participant experiment and evaluation experiment, are conducted. The participant experiment is to acquire sub-goals of participants. The human sub-goal transfer is evaluated on learning speed and performance after learning in evaluation experiment. The future work is to conduct two experiments and analyze the results.

1. はじめに

近年，深層学習の発達に伴い強化学習の応用範囲も広がり，様々な分野で強化学習が用いられるようになった．特にゲームやシミュレーター環境といったバーチャル環境での利用例が多くなってきている．バーチャル環境で強化学習が用いられる背景として，強化学習は学習初期に危険を顧みない挙動をしたり，膨大な試行錯誤が必要となるというような性質を持つことが挙げられる．また，強化学習は学習したタスクの性質が少しでも変化すると一から学習をし直さなければならないという汎化性能の低さという問題も抱えている．

そこで，状態や方策を階層化する階層型強化学習が提案された．階層型強化学習では，プリミティブな行動をまとめて，マクロ行動の獲得やメタレベルの行動選択により状態空間の探索の効率化が図れる．階層型強化学習ではマクロ行動を自動で得ることが難しく，良いマクロ行動とは何か，自動で獲得するための方法論が研究されていた．Option-Critic アーキテクチャはマクロな行動の数をあらかじめ指定するだけでマクロな行動を自動的に獲得する手法である．しかしながら，マクロ行動の獲得と選択方法，プリミティブな行動選択を同時学習するため学習に要する試行回数が膨大に必要である．そこで，本研究では人のサブゴール知識を転移することで Option-Critic アーキテクチャの試行回数を減らせることができるかを検証する．

また，人のサブゴール知識取得の方法をインタラクティブ機械学習の枠組みの中で提案する．インタラクティブ機械学習は計算機に関する知識がない人でも学習アルゴリズムにドメインの知識を転移することができる枠組みである．学習エージェントの学習プロセス中に人がサブゴールと考える状態でフィードバックを与えてもらうことでサブゴール知識を抽出する．サブゴール知識を Option-Critic アーキテクチャに適した形に変換し，学習前の初期値として用いる方法を提案する．

2. 関連研究

本節では，本研究に関連する研究としてインタラクティブ機械学習とインタラクティブシェイピング問題，強化学習における人の知識転移について説明する．

連絡先: 奥戸 嵩登，総合研究大学院大学，神奈川県三浦郡葉山町湘南国際村，okudo@nii.ac.jp

2.1 インタラクティブ機械学習

インタラクティブ機械学習とは，機械学習の学習プロセスに人が介在することができるような枠組みのことである．インタラクティブ機械学習はドメイン知識を有した人が機械学習の専門家を介さずに直接，機械学習アルゴリズムに知識を転移できるようにすることを目的としている [Amershi 14]．これまで，人がインタラクティブに学習プロセスに介入することを考慮したアルゴリズム [Settles 09] や，人が機械学習アルゴリズムとインタラクションしやすいインターフェース [山田 14] が研究されてきた．次節では，強化学習をインタラクティブ機械学習の枠組みに拡張したインタラクティブシェイピング問題について説明する．

2.2 インタラクティブシェイピング問題

インタラクティブシェイピング問題は，逐次意思決定問題において人から生成された一定の正負の報酬を用いて，そのタスクにおける最適方策を学習する問題である．人は訓練者としてエージェントと環境のインタラクションを観察し，任意のタイミングでエージェントの意思決定に対して評価を与えることができる [Knox 12]．代表的な手法として TAMER や TAMER + RL がある [Knox 12]．インタラクティブシェイピング問題において人のサブゴール知識を転移することを本研究は目指している．

2.3 強化学習における人の知識転移

強化学習における人の知識転移は学習エージェントに行動選択や価値関数にバイアスを与えることで学習速度を向上させることが示されている [Taylor 18]．人の知識転移の方法としては大きく 3 種類挙げられる．第一にデモンストレーションを与える方法である．代表的な手法としてはエキスパートの軌跡から報酬関数を生成する逆強化学習 [Ng 00] やエキスパートの方策をエージェントの行動決定に確率的に組み込んだ Human-agent Transfer [Taylor 11] がある．第二にオンラインで評価を与える方法である．先述した TAMER や，正負の評価だけでなく，評価しないこともフィードバックに含めた SABL [Loftin 16] が挙げられる．第三に学習エージェントが学習するタスクのカリキュラムを与える方法である．主に，タ

表 1: オプションの構成要素

記号	名称	説明
I	開始集合	オプションを開始できる状態集合
π	方策	行動決定を行う方策
β	終了条件	ある状態でのオプションの終了確率

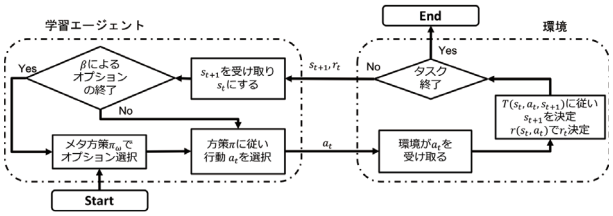

図 1: call-and-return 実行構成

スクのドメイン知識が豊富な設計者がカリキュラムの設計を行う [Taylor 18]．本研究はオンラインで評価を与えることでカリキュラムに関する知識も転移することを目指している．

3. 人のサブゴール知識の転移

本節では人のサブゴール知識を Option-Critic アーキテクチャに転移する手法を提案する．人のサブゴール知識の転移は 2 つのステップを踏む．第一に人のサブゴール知識を取得すること，第二にサブゴール知識の終了条件への変換である．その後，転移された終了条件を初期値として Option-Critic アーキテクチャで学習を行う．はじめに Option-Critic アーキテクチャについて説明を記述し，次にサブゴール知識の取得について記述する．最後にサブゴール知識の終了条件への変換について提案する．

3.1 Option-Critic アーキテクチャ

Option-Critic アーキテクチャはタスクのサブゴールへの分割と方策の学習，メタ方策の学習を一括して学習する手法である．オプション集合とそれらの中から 1 つの方策を選ぶメタな方策という構造を持つオプションフレームワークを踏襲している．1 つのオプションは $< I, \pi, \beta >$ のように定式化される．[Sutton 99] にオプションの構成要素についてまとめた表を表 1 に示す．

Option-Critic アーキテクチャでは実行に call-and-return を採用している．実行の流れを図 1 に示す．

図 1 より，始めに学習エージェントはメタ方策 π_ω でオプションを選択する．選択したオプションが持つ方策 π で行動 a_t を選択する．環境は行動 a_t を受け取り状態 s_t から状態 s_{t+1} へ遷移させ，報酬 r_t を決定する．報酬 r_t と遷移した状態 s_{t+1} を学習エージェントは受け取り，終了条件 β で現在のオプションを終了するかどうかを決定する．終了する場合は，メタ方策 π_ω で次のオプションを選択する．終了しない場合は現在のオプションを継続する．上記の流れを繰り返す．

Option-Critic アーキテクチャの学習則について説明する．Option-Critic アーキテクチャはメタ方策 π_ω，方策 π と終了条件 β を試行錯誤によって学習する．メタ方策は価値反復により方策を更新し，方策と終了条件は方策勾配に基づいて方策を更新する．メタ方策の学習則には intra-option Q ラーニング [Sutton 98] を用いる．Option-Critic アーキテクチャは初期状態 s_0，初期オプション ω_0 で始めた時の全軌跡の元で期待割引収益

$$\rho(\Omega, \theta, \vartheta, s_0, \omega_0) = \mathbb{E}_{\Omega,\theta,\omega}\left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} | s_0, \omega_0\right]$$

の最大化を図るアルゴリズムである [Bacon 17]．ここで，θ, ϑ はそれぞれ方策のパラメータ，終了条件のパラメータである．Ω はオプション集合である．方程式の操作により，下記のように表すことができる．

$$Q_\Omega(s, \omega) = \sum_{a \in \mathcal{A}} \pi_\theta(s, \omega)\left[r(s, a) + \sum_{s'} p^a_{ss'} U(s', \omega)\right] \quad (1)$$

$$U(s, \omega) = (1 - \beta_\vartheta(s))\, Q_\Omega(s, \omega) + \beta_\vartheta(s) \max_{\omega' \in \Omega} Q_\Omega(s, \omega') \quad (2)$$

ここで $r(s,a) = \mathbb{E}\{r_{t+1} | s_t = s, a_t = a\}$，$p^a_{ss'} = Pr\{s_{t+1} = s' | s_t = s, a_t = a\}$ を表している．メタ方策の更新則は

$$Q_\Omega(s_t, \omega) \leftarrow Q_\Omega(s_t, \omega) + \alpha\left[(r_{t+1} + \gamma U(s_{t+1}, \omega)) - Q_\Omega(s_t, \omega)\right]$$

となる．次に，目的関数を $Q_\Omega(s, o)$ とし，方策のパラメータ θ，終了条件のパラメータ ϑ でそれぞれ勾配を算出し，その勾配を用いてパラメータを更新する．方策パラメータの更新則は

$$\theta \leftarrow \theta + \sum_{s,\omega} \mu_\Omega(s, \omega | s_0, \omega_0) \sum_{a \in A} \frac{\partial \pi_\theta(a|s)}{\partial \theta} Q_U(s, \omega, a) \quad (3)$$

$$\mu_\Omega(s, \omega | s_0, \omega_0) = \sum_{t=0}^{\infty} \gamma^t P(s_t = s, \omega_t = \omega | s_0, \omega_0) \quad (4)$$

となる．$\mu_\Omega(s, \omega | s_0, \omega_0)$ は (s_0, ω_0) から始まる軌跡に沿って割引かれる重み付けである．終了条件の更新則は

$$\vartheta \leftarrow \vartheta - \sum_{\omega, s'} \mu_\Omega(s', \omega | s_1, \omega_0) \frac{\partial \beta_{\omega,\vartheta}(s')}{\partial \vartheta} A_\Omega(s', \omega) \quad (5)$$

$$A_\Omega(s', \omega) = Q_\Omega(s', \omega) - V_\Omega(s') \quad (6)$$

となる．Option-Critic アーキテクチャでは開始集合を，終了状態を除いたタスクの状態集合と同じとして扱っている．終了条件 β はメタ方策の切り替えタイミングを提供するので，サブゴールを定義していると解釈できる．

3.2 サブゴール知識の取得

サブゴール知識の取得には，人が訓練者としてエージェントと環境のインタラクションを観察し，任意のタイミングで学習エージェントの意思決定に対してフィードバックを与えることができる枠組みを利用する．人は学習エージェントの学習中の任意のタイミングでサブゴールを指定することができる．人が指定したタイミングで学習エージェントが訪問している状態 s_t をサブゴール状態 s_g として記録する．記録されたサブゴール状態の集合を $\mathcal{S}_g$ で表す．

3.3 サブゴール知識の終了条件への変換

サブゴール状態 s_g における終了確率が 1 に近づくように終了条件の関数のパラメータを更新することでサブゴール知識を転移する．人が設定したサブゴール状態 s_g でオプションの切替えを起こすためにサブゴール状態 s_g における全オプションの終了確率を 1 に漸近させる．確率 1 を教師データとして誤差関数

$$L(s) = \frac{1}{2}(1 - \beta_\vartheta(s))^2$$

を最小化するように勾配法を用いて $\beta_\vartheta(s)$ のパラメータ ϑ を更新する．全てのオプションの終了条件 $\beta_\vartheta(s)$ に対して同様の処理を行う．更新されたパラメータを初期値として Option-Critic アーキテクチャの学習を開始する．サブゴール知識の取得とサブゴール知識の終了条件への変換の擬似コードを Algorithm1 に示す．

Algorithm 1 Human sub-goal transfer

Ensure: $\beta_{\vartheta,*}(s)$
  repeat
    $s \leftarrow s_0$
    Choose ω according to π_ω
    repeat
      Choose a according to π in ω
      Take a in s, observe s', r
      Receive s_g from human trainer
      $S_g \leftarrow S_g \cup s_g$
      Update Q_ω, π_θ, β_ϑ
      if $\beta_{\omega,\vartheta}$ terminates in s' **then**
        Choose ω according to π_ω
      end if
      $s \leftarrow s'$
    until termination
  until predefined number of iterations
  while $\omega \in \Omega$ **do**
    while $s_g \in \mathcal{S}_g$ **do**
      $\vartheta \leftarrow \vartheta - \alpha_g \left(1 - \beta_{\omega,\vartheta}(s_g)\right) \frac{\partial \beta_{\omega,\vartheta}(s_g)}{\partial \vartheta}$
    end while
  end while

4. 実験

本実験の目的は人のサブゴール知識の転移が Option-Critic アーキテクチャの学習スピードと学習後の性能を向上させるかどうかを検証すること．本実験は 2 つのフェーズに別れる．始めに参加者実験を行ないサブゴール知識の抽出を行う．次にサブゴール知識を転移し，学習アルゴリズムの性能の評価実験を行う．

4.1 参加者実験

参加者実験では，タスクにおける参加者のサブゴール知識抽出が目的である．タスクは 2 種類，Fourroom タスクと Pinball Domain タスクを用いる．タスク選定の基準としては，状態空間の表現（離散/連続）の違いである．参加者は始めにタスク理解のために，タスクの説明を読み，制限時間 (5 分) の中でタスクをプレイする．次に，学習エージェントの学習中のプレイを観察し，任意のタイミングでそのプレイに対して正負の評価を与える．実験後，アンケートを実施する．

評価の与え方のインストラクションでは，学習エージェントが目標達成のために良い行動をしたと思う時に正の評価を悪い行動をしたと思う時に負の評価を与えるように伝える．学習エージェントには Option-Critic アーキテクチャを採用する．学習エージェントはプレイを参加者へ表示している間も学習を続ける．参加者に表示されているプレイが終了した時の最新のプレイを次のプレイを参加者へ表示する．

4.2 評価実験

評価実験では，3 つの手法を比較することで学習スピードと学習後の性能を比較することが目的である．3 つの手法の基盤となる手法として Option-Critic アーキテクチャを用いる．それぞれ学習前の終了条件の初期パラメータの与え方が異なる．以下にそれぞれの初期化に用いる情報を示す．

4.2.1 ベースライン

終了条件に人の知識転移を行わないベースラインとなる手法である．[Bacon 17] に従った設定を行う．

4.2.2 ランダムランドマーク

ランダムな状態をサブゴール集合として選択し，終了条件の初期パラメータとする手法である．状態の選択方法は Option-Critic の学習エージェントを用いて学習している間に 30%の確率でサブゴールとして訪問している状態をサブゴール集合に追加する．

4.2.3 人のサブゴール知識転移

本論文で提案する手法である．参加者実験で得たサブゴールを用いて終了条件を初期化する．

5. 結論

本論文では，階層型強化学習においてサブゴール知識の転移の方法と実験方法の提案を行った．今後の課題は提案した実験方法で実験を行い，実験結果の分析を行うことである．さらに，Atari ゲームのような高次元の観測が必要なタスクにおいてサブゴールの知識転移の検証とインタラクティブ機械学習の枠組みでサブゴール知識をオンラインで転移できるアルゴリズムの設計が挙げられる．さらに，人間の相互適応を利用し訓練者も学習エージェントと共に学習する枠組みを設計することを目指す．

参考文献

[Amershi 14] Amershi, S., Cakmak, M., Knox, W. B., and Kulesza, T.: Power to the People: The Role of Humans in Interactive Machine Learning, *AI Magazine*, Vol. 35, No. 4, p. 105 (2014)

[Bacon 17] Bacon, P.-L., Harb, J., and Precup, D.: The Option-Critic Architecture, in *AAAI* (2017)

[Knox 12] Knox, W. B.: *Learning from Human-Generated Reward*, PhD thesis (2012), Dissertation page: http://web.media.mit.edu/ bradknox/Dissertation.html

[Loftin 16] Loftin, R., Peng, B., MacGlashan, J., Littman, M. L., Taylor, M. E., Huang, J., and Roberts, D. L.: Learning behaviors via human-delivered discrete feedback: modeling implicit feedback strategies to speed up learning, *Autonomous Agents and Multi-Agent Systems*, Vol. 30, No. 1, pp. 30–59 (2016)

[Ng 00] Ng, A. Y. and Russell, S. J.: Algorithms for Inverse Reinforcement Learning, in *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pp. 663–670, San Francisco, CA, USA (2000), Morgan Kaufmann Publishers Inc.

[Settles 09] Settles, B.: Active Learning Literature Survey, Computer Sciences Technical Report 1648, University of Wisconsin–Madison (2009)

[Sutton 98] Sutton, R. S., Precup, D., and Singh, S. P.: Intra-Option Learning about Temporally Abstract Actions, in *Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998), Madison, Wisconsin, USA, July 24-27, 1998*, pp. 556–564 (1998)

[Sutton 99] Sutton, R. S., Precup, D., and Singh, S.: Between MDPs and semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning, *Artificial Intelligence*, Vol. 112, No. 1-2, pp. 181–211 (1999)

[Taylor 11] Taylor, M. E., Suay, H. B., and Chernova, S.: Integrating reinforcement learning with human demonstrations of varying ability, in *AAMAS* (2011)

[Taylor 18] Taylor, M. E.: Improving Reinforcement Learning with Human Input, in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, pp. 5724–5728 (2018)

[山田 14] 山田 誠二, 水上 淳貴, 岡部 正幸, インタラクティブ制約付きクラスタリングにおける制約選択を支援するインタラクションデザイン, 人工知能学会論文誌, Vol. 29, No. 2, pp. 259–267 (2014)