5U-04

段階的にサブゴールを獲得するStepwise Unified Hierarchical Reinforcement Learningの提案

黒須　宏道†　　　真部　雄介†　　　菅原　研次†
†千葉工業大学　情報科学部

1　はじめに

強化学習とは，エージェントが，観測可能な環境に対する適切な行動系列を確率的に学習する機械学習手法の1つである．エージェントは，設計者が環境に設定した報酬を元に，それを最大にするような行動を選択するように学習する．そのため，試行錯誤的に望ましい行動系列を学習できるが，長期的な戦略を必要とする行動系列の学習は困難であることが知られている．

強化学習に長期的な戦略をさせるアプローチの1つとして，Hierarchical Reinforcement Learning(HRL)[1]がある．HRLとは，事前に用意したサブゴール群から現環境の最適なサブゴールを学習するMeta Controllerと，サブゴールまでの行動系列を学習するControllerの2つから成る階層構造を持った強化学習アルゴリズムである．また，強化学習で解く環境の長さが一定レベルを超えると，HRLが必要となることがAl-Shedivatらの研究[2]で明らかになっている．そのため，近年ではHRLの問題点を解決しようとする研究が増加している．

HRLの問題点として，サブゴールを事前に用意する必要がある点が挙げられる．この問題点に関した研究として，HRLにサブゴール獲得処理を追加したUnified Hierarchical Reinforcement Learning(UHRL)[3]がある．これは，行動により得た経験を記憶し，それらをk-meansを用いてクラスタリングしたクラスタと，行動中に正の報酬を獲得した環境状態をサブゴールとして用いる．経験とは，行動，環境状態，報酬，次のステップ時間の環境状態の総称である．経験を積ませる処理の1つとして，ランダム行動を一定回数させる方法が用いられている．

UHRLの問題点として，1度のサブゴール獲得処理で，環境の全てのサブゴールを獲得できることを前提としている点が挙げられる．そのため，問題解決に強化学習の階層化を必要とするような環境[2]では，上手くサブゴールを獲得できない．この問題は，ランダム行動数を増やすことで対応できるが，その場合，運や環境の影響が大きく，環境ごとに最適なランダム行動数を求める必要がある．

そこで本研究では，この問題を解決可能な新たな強化学習アルゴリズムである，Stepwise Unified Hierarchical Reinforcement Learningを提案する．提案手法は，段階的にサブゴールを獲得する処理を追加することにより，徐々に適切なサブゴールを獲得していく手法である．

A Proposal of Stepwise Unified Hierarchical Reinforcement Learning
†Hiromichi KUROSU　†Yusuke MANABE　†Kenji SUGAWARA
†Faculty of Information and Computer Science, Chiba Institute of Technology

2　提案手法

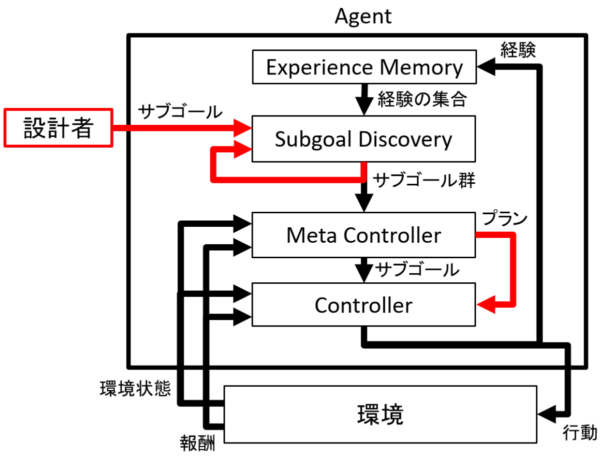

図1　提案アルゴリズムの構成図

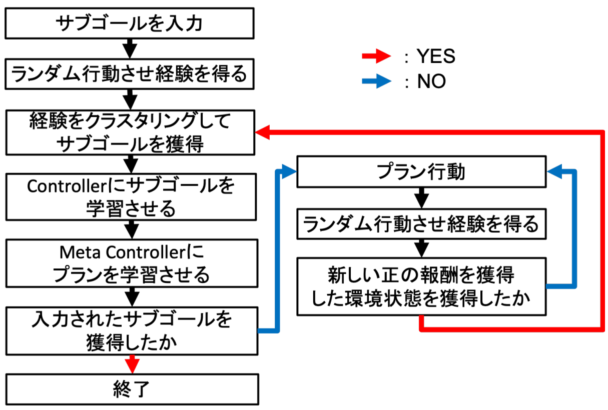

図2　提案アルゴリズムの流れ図

図1に提案アルゴリズムの構成図，図2に処理の流れ図をそれぞれ示す．構成図の赤色の矢印が，本アルゴリズムがUHRLの構成に加えた要素である．

本研究では，前段階で獲得したサブゴールを用いたプランの行動後，新たなサブゴール獲得処理をさせる．サブゴール獲得処理の繰り返しは，設計者が入力したサブゴールを獲得した際に終了する．

ここで，プランはサブゴールの適用順序のことであり，獲得したサブゴール群で構成される．また，プランは，Meta Controllerが学習によって獲得するプランと，サブゴール群から総当りで作られる総当りプランの2つに分けられる．Agentは，Meta Controllerのプランに基づく行動を行った後，総当りプランに基づく行動を実行し，新たなサブゴール獲得のためのランダム行動を行

う．上記のAgentによるサブゴール獲得処理は，ランダム行動により新しい正の報酬を持つ環境状態を獲得するまで実行され，総当たりプランが逐一変更される．

例えば，獲得されたサブゴール群が2つの地点（位置Aと位置B）だった時，総当りプランは(A), (B), (A→B), (B→A)の4通り作成される．(A→A)や(B→B)のような2つ以上で全てが同じサブゴールで構成されているプランは用いない．

新たなサブゴールを探す際に，Meta Controllerが学習したプラン，すなわち「現在のサブゴール群での最高報酬を得る訪れるサブゴールの順序」を用いて新たなサブゴールを探すため，運などの不安定な要素が絡まず，環境の問題解決に必要なサブゴールを獲得可能だと考える．

3 実験・評価方法

実験環境として，UHRLで用いられていたグリッドワールド環境を用いて，HRLが必要な環境[2]と問題解決に必要なサブゴール数が同じ環境f-rooms(図3)を作成した．行動主体が可能な行動は上下左右であり，行動主体がゴールに辿り着くことで問題解決となる．

サブゴールとして用いることができる正の報酬を持つ環境状態として，鍵，ドア，ゴールがあり，ドアは，同じ色の鍵を取得していない場合通ることができない．そのため，偶然ランダム行動中に鍵を獲得してからドアを訪れなければ，ドアをサブゴールとして獲得できず，1度のランダム行動では全てのサブゴールを上手く獲得できない．

本実験では，k-meansのクラスタ数に4，サブゴール獲得処理を終了するサブゴールにゴールを用いる．

この環境で，最終的に獲得したサブゴールを用いて問題解決が可能か確認することで，本アルゴリズムの評価を行う．

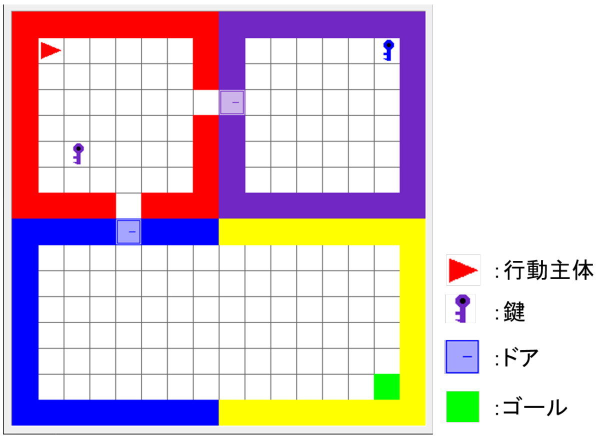

図3 f-rooms

4 実験結果

1段階目のサブゴール獲得処理で得たサブゴール群を図4に示す．それらを用いてMeta Controllerが学んだプラン(4→6)をサブゴール群に緑色で示す．これに総当たりプランを足し合わせたプランの行動をさせた後，新たなサブゴール獲得処理をさせる．

1段階目でMeta Controllerが学んだプラン(4→6)に総当たりプラン(0)を足し合わせたプランの行動後，サブゴール獲得処理をして得たサブゴール群を図5に示す．それらを用いてMeta Controllerが学習したプラン(4→6→2→8)をサブゴール群に緑色で示す．

実験結果により，提案アルゴリズムでHRLが必要な環境で問題解決に必要なサブゴール群を獲得できることがわかった．

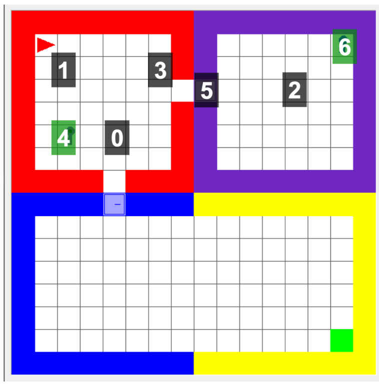

図4 1段階目のサブゴール群とMeta Controllerが学んだプラン(4→6)

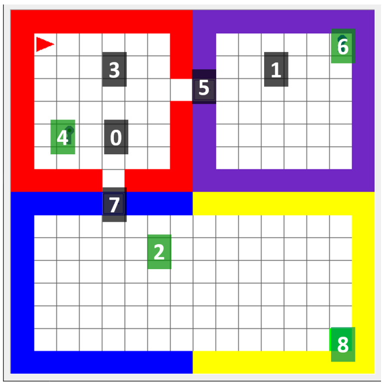

図5 最終的に獲得したサブゴール群とMeta Controllerが学んだプラン(4→6→2→8)

5 おわりに

本研究では，UHRLのサブゴール獲得処理を段階的にする手法を提案した．実験の結果，HRLが必要な環境で問題解決に必要なサブゴールを獲得できることがわかった．

今後は提案手法を他の環境を用いて実験を行う．

参考文献

[1] T.D. Kulkarni, et al. “Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation.” *Advances in neural information processing systems 29*. 2016. pp. 3675-3683.

[2] M. Al-Shedivat, et al. ”On the Complexity of Exploration in Goal-Driven Navigation.” arXiv preprint arXiv:1811.06889. 2018.

[3] J. Rafati, et al. “Unsupervised Methods For Subgoal Discovery During Intrinsic Motivation in Model-Free Hierarchical Reinforcement Learning.” *33rd AAAI Conference on Ar- tificial Intelligence (AAAI-19). Workshop on Knowledge Extraction From Games. Honolulu, HI, USA*. 2019.