4S-03

POMDPs環境下での群知能を導入した階層型強化学習

鈴木晃平† 加藤昇平†‡

†名古屋工業大学 大学院工学研究科情報工学専攻
‡名古屋工業大学 情報科学フロンティア研究院

1 はじめに

強化学習を実用化するためには課題がいくつか残っている．その一つに，POMDPs環境下で異なる状態を同一の状態と認識してしまう perceptual aliasing（知覚の見せかけ問題）が存在する．強化学習は現在の状態を正しく観測できなければ，良い方策を獲得できない．そこで本稿では，知覚の見せかけ問題を解決するため，群知能を導入した階層型強化学習を提案する．

本稿では，図1のような未知環境のグリッド迷路を扱う．エージェントの観測範囲は近傍の8セルとし，エージェントは，その近傍8セルを現在の状態として観測する．行動は「上」「下」「左」「右」の4種類とする．図1の環境では，スタートの状態「S」からゴールである状態「G」に到達するために，状態「A」「B」を通過しなければならない．しかし，状態「A」と「B」では近傍8セルが同一のものとなっているため，エージェントはそれらを異なる状態だと判別できない．このような問題が知覚の見せかけ問題である．

2 群知能を導入した階層型強化学習

提案手法は，強化学習を階層型に拡張し，サブゴールを用いて POMDPs 環境を MDPs 環境に分割する．そして，それぞれのサブエージェントがサブゴールまでの経路を学習することで，知覚の見せかけ問題を解決する．そのサブゴールは，サブエージェントが強化学習を用いて学習する．提案手法では,「サブゴール候補生成フェーズ」と「学習フェーズ」の2つのフェーズがあり，まずサブゴール候補生成フェーズで有効なサブゴールを探査し，その後，学習フェーズでサブゴールと経路を学習する．それぞれのフェーズに群知能を導入し，良い解の獲得と学習の効率化を図る．

それぞれのフェーズで強化学習の一種である Profit Sharing (PS) を用いる．PS は，報酬獲得後に優先度を一括に更新するオフライン学習であり，行動はルーレット選択により決定する．状態 s_t における行動 a_t の優先度 $P(s_t, a_t)$ の更新式を以下に示す．

$$P(s_t, a_t) \leftarrow P(s_t, a_t) + f(x), \tag{1}$$

ここで $f(x)$ は強化関数であり，x は報酬獲得までの距離を表す．

提案手法では，有効なサブゴールを生成するため，報酬の分配方法を変更する．提案手法では同一状態で行われたルールの報酬分配量を等しくする．そのため，1エピソードにおいて各ルールを更新するのは1回のみとし，各ルールの強化関数は報酬獲得までの距離 x に依存せず，次式とする．

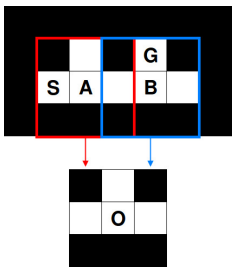

図 1: A POMDP environment

$$f(R, W) = \frac{R}{W}, \tag{2}$$

ここで，R は報酬，W はエピソード長である．

2.1 サブゴール候補生成フェーズ

サブゴール候補生成フェーズでは，PS を用いて有効なサブゴールになり得る状態を判別し，その状態をサブゴール候補とする．複数のエージェントが PS を行い，優良なエージェントの優先度をその他のエージェントに引き継ぐ．優良な優先度を持った複数のエージェントが各々ランダム性のある行動選択をすることで，よい解が早く獲得でき，有効なサブゴールを生成できると考えられる．サブゴール候補生成フェーズの計算手順を以下に示す．

1. 複数のエージェントを生成する．
2. 各エージェントが定められたエピソード数だけ PS を実行する．
3. エージェント間で優先度の情報を交換する
4. 一定回数，2 と 3 を繰り返す．
5. PS の結果によってサブゴール候補を決定する．

2.1.1 優先度の情報交換

Iima ら [1] は，評価値が最大であるエージェントの行動価値を他のエージェントに引き継ぐことが有効であると示した．提案手法では，エージェントをゴールまでの最短ステップで評価し，多様性を維持するため，その評価値が高い上位 N 個のエージェントの優先度をその他のエージェントに引き継ぐ．

2.1.2 サブゴール候補の生成

サブゴール候補は PS の結果によって生成する．PS は1エピソード内のルールに等しい報酬を与え，さらにゴールにたどりつくために必要なルールは毎回のエピソードで必ず行われるため，PS を行うとそれらのルールの優先度は等しく，全ルールの中で最大値をとる．そこで提案手法では優先度が最大となるルールをもつ状態をサブゴール候補とする．これにより，ゴールまでの経路上に知覚の見せかけ問題が発生している状態があるとき，その状態はサブゴール候補になり，MDPs 環境に分割することができる．

2.2 学習フェーズ

学習フェーズは，サブゴール候補生成フェーズと同様に，複数のエージェントが PS を行い，ゴールまで

Hierarchical Reinforcement Learning Introducing Swarm Intelligence under the POMDPs

Kohei SUZUKI† and Shohei KATO†‡
†Dept. of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology
‡Frontier Research Institute for Information Science, Nagoya Institute of Technology
†‡Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan
{suzuki, shohey}@katolab.nitech.ac.jp

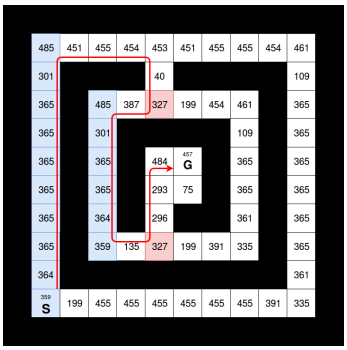

図 2: The maze of Wiering [2]

の経路とサブゴールを学習する．エージェント間でサブゴールの価値を情報交換することで，探索空間の大きいサブゴールの学習時間を短縮する．学習フェーズの手順を以下に示す．

1. サブエージェントを所有するエージェントを複数生成する．サブエージェントはサブゴールの価値テーブルと優先度テーブルをもつ．
2. サブエージェントを初期化する．優先度を初期化し，ルーレット選択を用いてサブゴールを決定する．
3. サブエージェントが PS を定められたエピソード数行う．先頭のサブエージェントから順に強化学習を行う．サブゴールに到達したとき，次のサブエージェントに切り替わる．
4. 各エージェントが greedy 選択を用いて 1 エピソード試行する．
5. エージェント間でサブゴールの価値を情報交換する．
6. 一定回数，2 5 を繰り返す．

サブゴールの価値は，行動選択のランダム性を取り除くため，greedy 選択を用いた 1 試行に基づいて更新する． s_t におけるサブゴールの価値 $SP(s_t)$ の更新式を以下に示す．

$$SP(s_t) \leftarrow SP(s_t) + \frac{R}{W}, \tag{3}$$

ここで，R は報酬，W はエピソード長である．エージェント間での情報交換では，サブゴール候補生成フェーズと同様に，評価値の高い上位 N 個のエージェントのサブゴール価値を引き継ぐ．

3 関連研究

Wiering[2] らは，Q-learning を階層型に拡張した HQ-learning を提案した．HQ-learning は強化学習を用いてサブゴールを学習するが，大規模環境では状態数が多く，学習効率が悪い．著者ら [3] は，HQ-learning を改良しサブゴールを遺伝的アルゴリズム（GA）により創発する Hybrid learning using profit sharing and genetic algorithm（HPG）を提案した．しかし，この手法は GA を用いるため学習が遅い．

Arai ら [4] は，同状態において等しく報酬分配を行う First Visit Profit Sharing (FVPS) を提案した．この手法は，知覚の見せかけ問題を確率的に解決するため，状態の混同が多く発生している環境ではランダム行動に近くなってしまう．また価値を累積しているため，局所解に陥りやすい．

4 迷路走行実験

図 2 に示す Wiering[2] の迷路を用いて POMDPs 環境下での性能実験を行う．セル上の数字は，観測情報

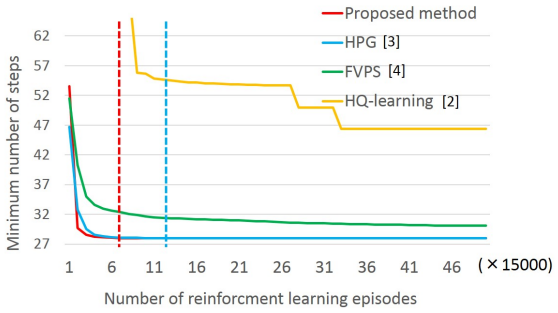

図 3: The result in maze of Wiering [2]

を表現している．この環境でゴールにたどり着くためには，知覚の見せかけ問題を必ず解決しなければならない．赤の矢印で示された最短経路上では赤いセルと青いセルで知覚の見せかけ問題が発生している．提案手法，HPG，HQ-learning，FVPS の 4 手法で比較実験を行う．各手法の試行回数は，強化学習のエピソード数 75 万回に統一されている．この迷路の最短ステップは 28 であり，強化学習の最大ステップ数は 150，提案手法のエージェント数 20，サブエージェントル数 3 とした．実験は各手法ごとに 100 回行い，結果は 100 回の平均をとる．

図 3 に実験結果を示す．グラフの縦軸は獲得した解の中での最短ステップ数，横軸は強化学習のエピソード数であり，点線は最短ステップを獲得したエピソード数を示している．提案手法は，全手法の中で，最も早く最適解を獲得した．一方，FVPS はゴールまでの経路が複数あるため局所解に陥り，HQ-learning は学習すべき状態が多く，サブゴールの学習に時間がかかっている．HPG は最適解を獲得しているものの，GA を用いているため，提案手法より 2 倍近くのエピソード数を要している．以上から，サブゴール候補の生成と群知能導入の有効性を確認できた．

5 おわりに

本稿では，群知能を導入した階層型強化学習法を提案した．Wiering の迷路を用いた実験により，POMDPs 環境において関連研究 [2, 3, 4] より早く最適解を獲得した．今後の課題として，エージェント間の情報交換方法を改善し，さらに車輪型ロボットを用いた実環境への適応が挙げられる．

参考文献

[1] Hitoshi Iima and Yasuaki Kuroe. Swarm reinforcement learning algorithms based on sarsa method. In *SICE Annual Conference, 2008*, pp. 2045–2049. IEEE, 2008.

[2] Marco Wiering and Jürgen Schmidhuber. Hq-learning. *Adaptive Behavior*, Vol. 6, No. 2, pp. 219–246, 1997.

[3] Kohei Suzuki and Shohei Kato. Hybrid learning using profit sharing and genetic algorithm for partially observable markov decision processes. In *International Conference on Network-Based Information Systems*, pp. 463–475. Springer, 2017.

[4] Sachiyo Arai and Katia Sycara. Credit assignment method for learning effective stochastic policies in uncertain domains. In *GECCO*, pp. 815–822. Morgan Kaufmann Publishers Inc., 2001.