The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

Profit Sharing と遺伝的アルゴリズムを用いたハイブリッド学習 -MDPs 環境でのタスク分割性能-

Hybird Learning Using Profit Sharing and Genetic Algorithm -Task Division Performance in MDP Environments-

鈴木 晃平[*1] 加藤 昇平[*1*2]
Kohei Suzuki Shohei Kato

[*1]名古屋工業大学工学研究科情報工学専攻
Department of Computer Science and Engineering Graduate School of Engineering, Nagoya Institute of Technology

[*2]名古屋工業大学 情報科学フロンティア研究院
Frontier Research Institute for Information Science, Nagoya Institute of Technology

Reinforcement learning is generally performed in the Markov decision processes (MDP). However, there is a possibility that the agent cannot correctly observe the environment due to the perception ability of the sensor. This is called partially observable Markov decision processes (POMDP). In a POMDP environment, an agent may observe the same information at more than one state. We proposed a hybrid learning method using Profit Sharing and genetic algorithm (HPG) for this problem. However, Most of real problems can be represented in an MDP environments. In this paper, we improve HPG to adapt to MDPs environments and report the effectiveness of our method by some experiments with mazes.

1. はじめに

強化学習は，学習者であるエージェントが環境との相互作用から目標状態に達する方策の学習を行う手法である．エージェントは目標状態に達したときに報酬が得られ，それを最大化することを目的に学習する．一般に強化学習は，状態を正しく観測できるマルコフ決定過程（MDPs）の環境を想定している．しかし実際には，センサの知覚能力により状態の混同が起こっている可能性があり，エージェントがうまく学習を行えないことがある．状態の混同が発生している環境を部分観測マルコフ決定過程（POMDPs）といい，それが原因で発生する問題を不完全知覚問題という．

不完全知覚問題の解決法は，サブタスクに分割する手法 [Wiering 97, Nomura 15]，Profit Sharing （PS）を用いた手法 [Arai 01, 植村 05]，状態遷移の履歴を用いる手法 [McCallum 95] の３種類に大別できる．これらの手法ではそれぞれ欠点を持っている．サブタスクに分割する手法では，サブゴールの学習とサブタスクの再学習を行う必要があり，学習効率が悪い．PS を用いた手法では，不完全知覚状態が多く存在する環境ではランダム行動と変わらなくなるなど複雑な環境下に弱い．状態遷移の履歴を用いる手法では，環境が大きくなると計算量が発散してしまう．筆者ら [鈴木 17] は，これらの欠点を改善すべく PS と遺伝的アルゴリズム（GA）によりサブゴールを決定し，不完全知覚問題を解決する Hybrid learning using Profit sharing and Genetic algorithm （HPG）を提案した．しかし，実問題では MDPs 環境であることが多いものの，HPG の MDPs 環境下における有効性を検証していなかった．本稿では，MDPs 環境にも対応できるよう HPG を改良し，実験によりタスク分割性能を検証する．

2. 不完全知覚問題

本稿では，図 1 のようなグリッドの環境を想定する．エージェントの観測範囲は近傍の８セルとし，壁の有無のみ知覚でき，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 と状態 B では近傍８セルが同一のものとなっているため，状態の混同が発生し，エージェントは２つを異なる状態だと判別できない．このように異なる状態を同一の状態と観測してしまう環境が POMDPs 環境である．さらにエージェントは，状態 A では右に，状態 B では上に移動しなければならないが，２つを同一状態とみなしているため，ゴールに近い状態 B ですべき上という行動を学習してしまい，状態 A でループに陥ってしまう．このように POMDPs 環境下で正しく学習を行えない問題が不完全知覚問題である．

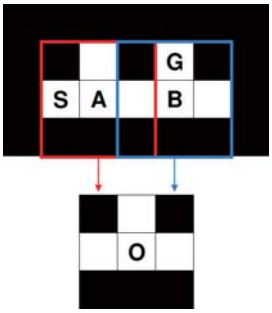

図 1: A POMDP environment

3. Profit Sharing と遺伝的アルゴリズムを用いたハイブリッド学習

提案手法では，GA を用いて不完全知覚問題を解決するエージェントを生成する．各エージェントは配列構造で表現されるサブゴールを設定し POMDPs 環境の分割を行い，サブエージェントが MDPs 環境で強化学習を行うことで不完全知覚問題に対応する．その強化学習の結果に応じ，遺伝的操作を行い環境に適したエージェントが生成される．以下に提案手法の計

連絡先: 加藤昇平，名古屋工業大学，愛知県名古屋市昭和区御器所町，052-735-5625，shohey@katolab.nitech.ac.jp

算手順を示す.

1. 初期集団を生成する.
   X 個のサブゴールと $(X+1)$ 個のサブエージェントをもつエージェント Y 個体を生成する.
2. 各エージェントが強化学習を行う.
   各エージェントは，サブゴール到達を切り替え条件に先頭のサブエージェントから順に学習する.
3. 交叉，突然変異を行い，新たなエージェントを生成する.
4. 以降,(2) ,(3) を世代数繰り返し終了.

3.1 初期集団生成

初期集団生成時において，エージェントは環境にどのような状態が存在するか未知である．サブゴールをランダムに生成すると，到達不能なサブゴールや存在し得ないサブゴールが多数出現する．これにより試行回数が増大し，さらに適応度は各サブゴールではなく順序付きサブゴール集合を評価するため，有効なサブゴールをもつエージェントの適応度まで下げてしまう可能性があり，学習が遅くなると考えられる．そのため，初期集団生成時にある程度有効なサブゴールを設定すべきである．そこで，提案手法では強化学習の一種である PS を用いてサブゴールを生成する．サブゴール生成の計算手順を以下に示す.

1. 報酬分配量を考慮した PS で学習を行う.
2. 各状態のルールの優先度からサブゴール候補を決定する.
3. サブゴール候補からランダムに各個体のサブゴールを決定する.

3.1.1 報酬分配量を考慮した Profit Sharing

Profit Sharing （PS）は，各状態ごとの行動の優先度を学習する手法で，報酬獲得した際にエピソード内すべてのルールの優先度を一括に強化するオフライン学習である．状態 s_t における行動 a_t の優先度 $P(s_t, a_t)$ は次式で強化される.

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

ここで $f(x)$ は強化関数といい，x は報酬獲得までの距離を表す．PS は，この優先度に基づくルーレット選択により行動選択を行う．提案手法では，PS のオフライン学習とルーレット選択に着目し，有効ルールである可能性が高い状態を判別し，その状態をサブゴール候補とする.

PS の強化関数は，等比減少関数が使われていることが多い．しかし，有効なサブゴールを生成するため，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強化するのは 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.

$$f(x) = \frac{1}{W} \tag{2}$$

ここで W は，エピソード長である．ゴールに必要不可欠なルールは，不完全知覚状態の有無に関わらず毎回強化されるため，全てのルールの中で最も大きい優先度を持つ．そこで提案手法では，大きい優先度を持つルールの状態をサブゴール候補とし，初期集団生成時にこの中から各個体がサブゴールをランダムに選択する.

3.2 強化学習

初期集団生成時だけでなく各エージェントが行う強化学習にも前述した報酬分配法の PS を用い，学習の高速化を図る．PS は，報酬に至るルールをすべて強化するため報酬獲得に貪欲であり学習の立ち上がりが速い．また植村ら [植村 05] は，複数のルールを必要とする不完全知覚問題において，ランダム選択に劣らない性能をもつためには報酬獲得に必要なルールをすべて同じ確率で選択できれば十分であると示した．提案手法では，同一状態において等しく報酬を与えているため，この十分条件を満たし不完全知覚問題にも対応できる．そのため，サブゴール分割がうまくできず不完全知覚が起こっている場合でも，目標状態に達することができる．これにより早い段階で有効なサブゴールかどうか判断でき，学習時間の短縮となる.

提案手法では，PS でサブゴール候補を生成し，PS によって学習を行うエージェントを GA で選択，淘汰して問題を解決する．つまり PS と GA のハイブリッド学習となる．提案手法は，PS の局所解の陥りやすさと GA の学習の遅さというお互いの欠点を 2 つの手法を組み合せることで解消している.

3.3 交叉

提案手法では，サブゴール候補からサブゴールをランダムに決定するため，サブゴールの組み合わせが無秩序になっている可能性が高い．それにより学習序盤では，遠回りしないように適切なサブゴール到達順序にする必要がある．また学習終盤では，局所解に陥っていた場合に脱出できるようにする必要がある．そこで，交叉は 2 種類行う.

サブゴールの交叉

サブゴールの交叉は一様交叉を行う．これにより，学習序盤に意味を成さないサブゴールの順序を入れ替える．また優先度は引き継がずに初期化する．これにより，新たな解が見つかりやすくなり局所解からの脱出もできると考えられる.

サブエージェントの交叉

サブエージェントの交叉は一点交叉を行い，サブゴールと優先度を引き継ぐ．サブゴールの交叉のみでは優先度を引き継がないため，学習が遅くなってしまう．また学習が進むにつれ，サブゴールが適切な組み合わせになっていくが，サブゴールの交叉を行うことで，またサブゴールの順序を乱すことになり学習率が低下する．そこでサブエージェントの交叉を行うことで学習速度を速くする．また親 2 個体の切断箇所は共通ではなく，各個体ランダムに決定する．これによりサブゴール数が動的に変化する.

3.4 適応度計算

適応度は，各エージェントの学習後に greedy 法を用いて行動させ，ゴールしたか否かで評価方法を変更する．ゴール達成した場合は greedy 法によるステップ数，ゴールしなかった場合は学習中のゴール回数をみて更新する．これは不完全知覚問題を，ルーレット選択により偶然解決したエージェントの適応度を上げないようにするためである．適応度を次式で与える.

$$F1 = \begin{cases} R + \dfrac{Max\_step - step}{sub \times a} & \text{(completed)} \\ \dfrac{goal}{b} & \text{(uncompleted)} \end{cases} \tag{3}$$

ここで，R はゴール報酬値，Max_step は最大ステップ数，$step$ はゴール到達までのステップ数，sub はサブゴール数，$goal$ は強化学習中のゴール回数，a，b は重みを表す．重み b につい

ては、$\frac{goal}{b}$ の値がゴール報酬値 R を超えないように設定する．適応度は，最適解に近づけるためステップ数を基準に決めるが，この式で学習を進めると同様なサブゴールのみ残り，多様性がなくなってしまう．局所解に陥らないためにも，遺伝子に多様性を持たせ新たな解を見つける可能性を高めなければならない．そこで式（３）を求めた後，同一の順序付きサブゴール集合をもっているエージェントの適応度を重み c で除算する．同様の順序付きサブゴール集合を所持するエージェントを消すことでも多様性は維持できるが，学習序盤に有効なサブゴールが遺伝しにくくなり学習が遅くなってしまうため，適応度を低くする形をとる．

3.5 突然変異

初期集団を生成する際に，適切なサブゴールがサブゴール候補の中に存在しない可能性もある．その場合に局所解に陥らないためには，突然変異が重要となる．提案手法の突然変異では，配列構造で表現されるサブゴールをランダムに生成する．しかし，ただランダムに生成するだけでは，前述した通り無駄なサブゴールができてしまう．そこで突然変異で生成するサブゴールの一部にドントケアを用いて抽象化する．ドントケアを用いた要素は，観測情報の要素が何であっても真とする．これにより，生成されるサブゴールが有効である可能性を高められる．

4. 関連研究

野村ら [Nomura 15] は，HQ-learning を改良しサブゴールを GA により創発するサブゴール創発強化学習 (SERL) を提案した．しかし，この手法はランダムにサブゴールを生成するため学習が遅い．そこで筆者ら [鈴木 17] は，PS を用いて不完全知覚状態の可能性が高い状態をサブゴール候補とした HPG を提案した．しかし，これは MDPs 環境において有効でないサブゴールを生成してしまい学習効率が悪い．また arai ら [Arai 01] 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，植村ら [植村 05] が FVPS の報酬分配量をエピソードの部分系列を用いて増やした Episode-based Profit Sharing (EPS) を提案した．これらは，確率的に不完全知覚問題を解くため，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，局所解に陥りやすく環境変化にも対応できない．

5. POMDPs 環境下での性能実験

図 2 に示す Wiering[Wiering 97] の迷路を用いて POMDPs 環境下での性能実験を行う．観測範囲は近傍 8 セルのみで，行動は上下左右の 4 種類とする．数字が書いてあるセルが道で，黒いセルが壁であり，壁に移動する行動を選んだ場合は移動を行わず，ステップ数のみ加える．道のセル上の数字は，観測情報を説明上わかりやすく表したもので，図 3 の配列構造を 9 桁の 2 進数とみなし，それを 10 進数に変換した値である．この環境で最短経路を得るためには，青いセルと赤いセルで発生する不完全知覚問題を解決しなければならない．他の経路についても必ず不完全知覚問題が複数発生する環境である．ここでは，提案手法，HPG，SERL，EPS，FVPS の 5 種類の手法で比較実験を行う．各手法の試行回数は，強化学習試行数 75 万回に統一した．この迷路の最短ステップは 28 で，1 エピソードの最大ステップ数は 150，初期サブゴール数 5 とし，また実験は 100 回行いその平均をとる．

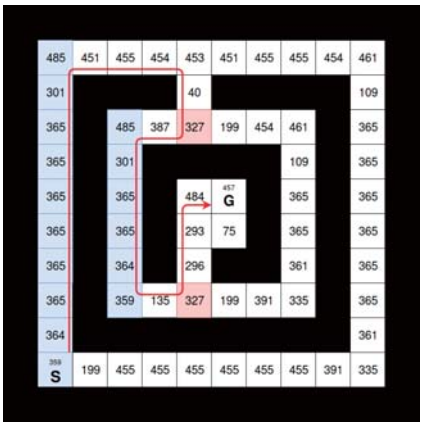

図 2: The maze of Wiering[Wiering 97]

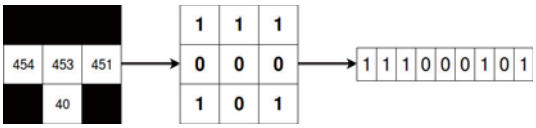

図 3: A example of subgoal

図 4 に実験結果を示す．グラフの縦軸は最短ステップ数，横軸は強化学習回数である．EPS と FVPS は局所解に陥いることがあり，最短ステップに収束していない．これらは，経験に固執して解の更新がされにくいため，学習序盤に最適解が見つからない限り局所解に陥ってしまう．SERL では，ほぼ最適解に収束している．最適解にはならなかった原因として，初期に無駄なサブゴールが多く創発され，75 万回では適切なサブゴールが発見できなかったことが考えられる．一方提案手法と HPG では，最短ステップ数に収束している．提案手法は，HPG より学習の立ち上がりが遅いが，先に最短ステップに収束した．これはサブゴール候補が多くなった分，適切なサブゴールを見つけることに時間がかかったが，解の多様性が高くなり，最適解を獲得しやすくなったと考えられる．

6. MDPs 環境下への適応性実験

図 5 に示す Sutton[Sutton 99] の迷路を用いて MDPs 環境下への適応性実験を行う．この迷路は，ゴールまでの経路が複数存在するが，最短経路を獲得するためには”25”と”62”のセルを通過しなければならない．ここでは，提案手法，HPG，SERL，FVPS，Q-learning の 5 種類の手法で比較実験を行う．各手法の試行回数は，強化学習試行数 1 万回に統一した．この迷路の最短ステップは 17 で，1 エピソードの最大ステップ数は 150，初期サブゴール数 3 とし，また実験は 100 回行いその平均をとる．

図 6 に実験結果を示す．Q-learning は学習速度は速いが，局所解に陥っており，FVPS では，1 エピソードあたりの報酬が少ないため学習が遅い．提案手法，HPG，SERL を比較すると，提案手法が最も学習速度が速い．これにより初期集団生成時に有効ルールをサブゴール候補にした有効性がみられた．また Q-learning と比較すると，学習速度は遅いが，最適解を獲得している．このことから，簡単なタスクにおいては提案手法はオーバーヘッドとなってしまうが，局所解から逃れることが

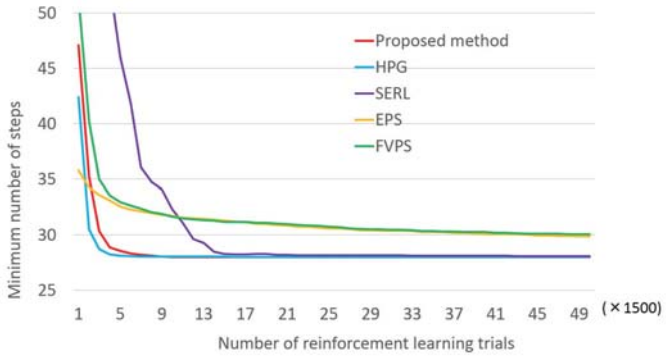

図 4: The result in the POMDP environment

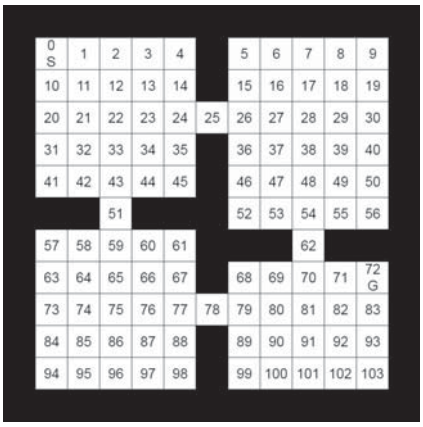

図 5: The maze of Sutton[Sutton 99]

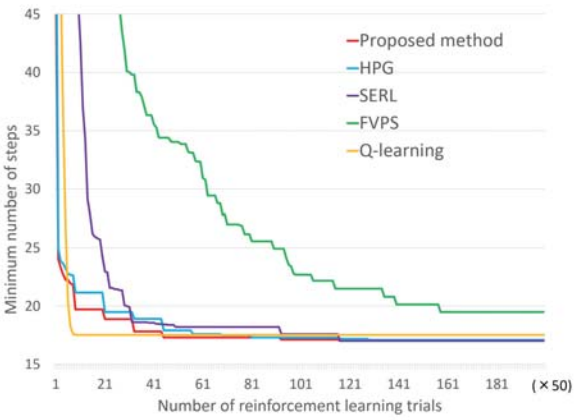

図 6: The result in the MDP environment

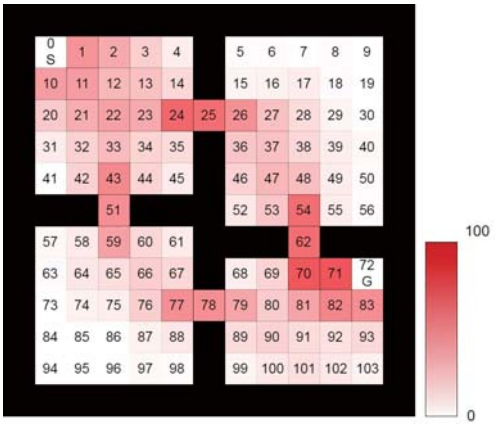

図 7: The distribution of subgoal candidates

できると考えられる．

図 7 に初期集団生成時のサブゴール候補の分布を示す．100 回の実験でサブゴール候補となった回数に赤色の濃さが対応している．最短経路上の状態がほぼ網羅されているが，”25”と”62”のセルはそれぞれ 100 回中 58 回と 62 回サブゴール候補になっており，毎回サブゴール候補になっているわけではない．しかしながら図 6 では最短経路を獲得しており，GA によるタスク分割の有効性がみられる．

7. おわりに

本稿では，不完全知覚状態の有無に関わらずタスクを解決する手法を提案した．実験により，提案手法は POMDPs 環境におけるタスクでは，他の手法より早く最適解を獲得し，また MDPs 環境下におけるタスクでは，サブタスクに分割する手法として最適解を獲得し有効性を示した．今後の課題として，複雑な MDPs 環境下での性能実験に加え，連続状態空間への拡張と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

[Arai 01] Arai, S. and Sycara, K.: Credit assignment method for learning effective stochastic policies in uncertain domains, in *Proceedings of the 3rd Annual Conference on Genetic and Evolutionary Computation*, pp. 815–822Morgan Kaufmann Publishers Inc. (2001)

[McCallum 95] McCallum, R. A.: Instance-based utile distinctions for reinforcement learning with hidden state, in *ICML*, pp. 387–395 (1995)

[Nomura 15] Nomura, T. and Kato, S.: Dynamic subgoal generation using evolutionary computation for reinforcement learning under POMDP, *International Symposium on Artificial Life and Robotics*, Vol. 20, pp. 322–327 (2015)

[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)

[Wiering 97] Wiering, M. and Schmidhuber, J.: HQ-learning, *Adaptive Behavior*, Vol. 6, No. 2, pp. 219–246 (1997)

[植村 05] 植村 渉, 上野 敦志, 辰巳 昭治：POMDPs 環境のためのエピソード強化型強化学習法, 電子情報通信学会論文誌 A, Vol. 88, No. 6, pp. 761–774 (2005)

[鈴木 17] 鈴木晃平, 加藤昇平：不完全知覚問題に対する Profit Sharing と遺伝的アルゴリズムを用いたハイブリッド学習, 電気学会論文誌 C, Vol. 137, No. 12, pp. 1591–1599 (2017)