F-018

POMDPs 環境におけるエントロピと遺伝的アルゴリズムを用いてサブゴール創発を行う強化学習法

Reinforcement Learning Emerging Subgoals by Entropy and Genetic Algorithm for POMDPs Environments

鈴木 晃平 †　**加藤 昇平** †
Kohei Suzuki　Shohei Kato

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を学習する機械学習の一種である．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．強化学習は，設計者がエージェントに目標状態と報酬のみ与えさえすれば，人間が把握できない未知なパラメータがある複雑な環境にも対応でき，また人間より優れた解を見つける可能性もあるため，注目されている [1].

しかし，ロボティクス分野における強化学習の実用化には大きな課題がある．一般に強化学習では，状態が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．しかし実問題では，常に状態を正確に観測できるとは限らない．自律移動ロボットの経路の学習を例に挙げると，センサの故障や障害物の出現などにより状態の混同が発生し，エージェントは正しく学習できないことがある．このように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）[2] とよび，それが原因で起こる問題を不完全知覚問題という．強化学習の実用化のためには，POMDPs 環境下での学習法が必要となる．

筆者ら [3] は，Profit Sharing（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，HPG は PS による不完全知覚状態の判定が不十分であり，GA による学習に時間を要してしまう可能性がある．本稿では，この欠点を改善すべく行動選択のエントロピと GA を用いてサブゴール創発を行う強化学習法を提案し，迷路走行タスクを用いた実験により有効性を検証する．

2　不完全知覚問題

本稿では，Fig. 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．Fig. 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，正しい政策を獲得できない．このように POMDPs 環境下で正しく学習できない問題が不完全知覚問題である．

3　Hybrid learning using Profit sharing and Genetic algorithm （HPG)

HPG 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，サブゴールを創発する．すなわち，サブゴールを持つエージェントを個体とした GA により環境に適したエージェントを生成する．

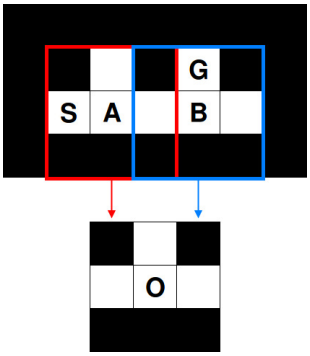


Fig. 1　A POMDP environment

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し．試行回数が増大する．そこで，HPG では強化学習の一種である PS を用いて不完全知覚状態の判定を行い，不完全知覚状態と判定された状態の集合から初期個体のサブゴールを決定する．

3.1　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．

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，不完全知覚状態を判定するため，HPG 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．

3.2　初期集団生成

報酬分配量を考慮した PS で学習を行うと，ある状態において報酬獲得に必要不可欠なルールが複数ある場合，それらの優先度は均一の値に収束していく．ある状態の選択可能な行動の集合を *Action* として全ての行動が報酬獲得に必要だと仮定する

† 名古屋工業大学，Nagoya Institute of Technology

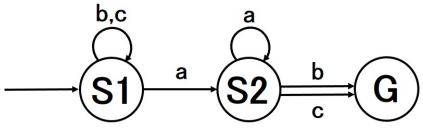


Fig. 2　A example of POMDP environment

と，各行動の優先度は等しくなり，行動選択確率 Pr は次式の値になる．

$$Pr = \frac{1}{|Action|} \tag{3}$$

報酬獲得に必要なルールは 1 エピソードごとに必ず報酬が与えられるため，この式の値を下回ることはない．すなわち，ある状態において，式（3）の値より行動選択確率が大きな行動が複数存在するとき，不完全知覚問題が発生しており，かつ異なった行動を必要としている可能性が高いと考えられる．そこで HPG では，初期集団生成前に PS で学習を行い，行動選択確率が式（3）の値を超えたルールが複数ある状態をサブゴール候補とし，この中から各個体がランダムにサブゴールを選択する．

3.3　HPG の問題点

HPG における不完全知覚状態の判定は，以下のような環境下で正確ではない．

- 不完全知覚状態において有効ルールが複数存在する環境
- 最適解は不完全知覚問題を解決しなければならないが，その不完全知覚を回避する政策が存在する環境

前者の環境の例を Fig. 2 に示す．図中の円は状態，a,b,c は行動，矢印は次状態への遷移を表し，状態 S1 と S2 は同一状態と観測される．状態 S1 では，行動 a が有効ルールとなるが，状態 S2 においては行動 b，c どちらも有効ルールとなる．この環境で報酬分配量を考慮した PS を用いて学習すると，行動 a は必ず強化されるが，行動 b と行動 c は 1 エピソードでどちらかのみ強化され，均一に優先度を強化できない．そのため，行動 b と行動 c の行動選択確率が式（3）を下回り，不完全知覚状態と判定されない．後者の環境では，不完全知覚状態を回避する政策を学習し，行動選択確率が式（3）の値を超えない可能性があり，不完全知覚状態の判定に失敗する．

4　エントロピと GA を用いた強化学習法

提案手法は，HPG と同様に GA を用いて不完全知覚問題に対応するエージェントを生成する．以下に提案手法の計算手順の概略を示す．

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．サブゴールは観測情報のエントロピから生成したサブゴール候補の中からランダムに決定する．
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．
   交叉によりサブゴールを引き継ぎ，突然変異によりサブゴールを創発する．
4. 以降，2，3 を世代数繰り返し終了．

以下，それぞれの手順について順を追って説明する．

4.1　初期集団生成

初期集団生成では，サブゴールとサブエージェントを所持するエージェントを個体として生成する．サブゴールは HPG と同様に報酬分配量を考慮した PS により決定するが，エントロピを導入することで，HPG で判定できない環境にも対応していく．

4.1.1　エントロピによるサブゴール候補の生成

提案手法は，報酬分配量を考慮した PS の学習後，各状態 s における行動選択のエントロピ En を次式で求める．

$$En(s) = -\sum_{a \in Act(s)} P(a) \log_2 P(a) \tag{4}$$

ここで，Act は選択可能可能な行動の集合，$P(a)$ は行動 a の行動選択確率である．Act において 1 つの行動のみ選択確率が大きくなると式（4）の値は小さくなり，反対に全ての行動の選択確率が等しくなると式（4）の値は大きくなる．すなわち，行動選択のエントロピは政策の不確実性を表現しており，POMDPs 環境では行動選択のエントロピは大きくなる．ある不完全知覚状態 s^* において，2 つの行動 (a_1, a_2) が必要となるとき，これらの行動選択確率は $\frac{1}{2}$ に近づき，その他の行動選択確率は 0 に近づく．このとき，エントロピは次式の値になる．

$$\begin{aligned} En(s^*) &= -\sum_{a \in Act(s^*)} P(a) \log_2 P(a) \\ &= -\sum_{a \in \{a_1, a_2\}} P(a) \log_2 P(a) \\ &= -\frac{1}{2} \log_2 \frac{1}{2} \\ &= 1 \end{aligned} \tag{5}$$

MDPs 環境では，ある状態において高々 1 つの行動の選択確率が大きくなり，エントロピが式 (5) を超えない．一方，POMDPs 環境において少なくとも 2 種類以上の行動が必要となるとき，式 (5) を超える．よって提案手法では，式 (5) を超えるエントロピを持つ状態を不完全知覚状態の判定条件とし，その状態をサブゴール候補とする．これにより，均一に優先度が強化できない不完全知覚状態にも対応することができる．そして，このサブゴール候補の中から各エージェントがランダムにサブゴールを決定する．

4.2　強化学習

サブ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強化学習を行う．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 [4] 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．

報酬分配量を考慮した PS のみでも不完全知覚問題に対応できるが，不完全知覚状態ではランダム行動とほぼ同様の挙動をとるため，良い解を得ることはできない．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．

4.3　遺伝的操作

4.3.1　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．

$$F1 = \begin{cases} R + \dfrac{Max\_step - step}{sub \times a} & \text{(completed)} \\ \dfrac{goal}{b} & \text{(uncompleted)} \end{cases} \quad (6)$$

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（６）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．

4.3.2 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順序でサブゴールに到達する必要がある．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べきである．そこで，交叉は２種類行う．

交叉１

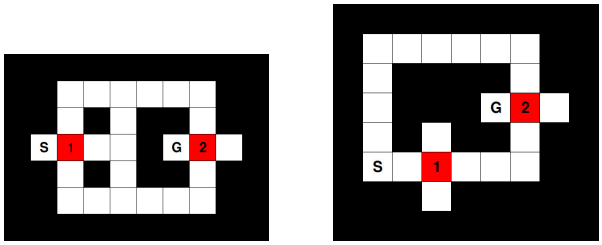
交叉１はサブゴールの一様交叉を行い，学習序盤の意味を成さないサブゴールの順序を入れ替える．また優先度は引き継がずに初期化する．これにより，新たな解が見つかりやすくなり局所解からの脱出もできると考えられる．

交叉２

サブエージェントの一点交叉を行い，サブゴールと優先度をそのまま引き継ぐ．交叉１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，交叉１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効率が低下する．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し，環境に適した数のサブゴール生成を可能とする．

5 関連研究

野村ら [5] は，HQ-learning[6] 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．また arai ら [7] 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 [4] 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．



(a) Maze 1　(b) Maze 2

Fig. 3　Mazes

Table. 1　The number of subgoal candidates on the red cell (the sum of 100 runs)

Method / Maze	Proposed method	HPG
Maze 1	100	46
Maze 2	100	72

6 不完全知覚判定の比較実験

Fig. 3 に示す迷路を用いて，提案手法と HPG による不完全知覚状態の判定の比較実験を行う．エージェントの観測範囲は近傍８セルのみで，行動は「上」,「下」,「左」,「右」の４種類とし，それぞれの迷路は赤いセルにおいて不完全知覚問題が発生する．Fig. 3(a) の迷路では，赤いセルにおいて有効ルールが複数ある環境であり，状態”2”においては「左」が有効ルールとなるが，状態”1”においては「左」以外のどの行動を選択してもゴールに到達することができる．Fig. 3(b) の迷路では，状態”1”において不完全知覚問題が発生するが，それを回避する経路が存在する迷路である．PS の試行数は 1000 回とし，報酬分配量を考慮した PS のパラメータは両手法統一した．

Table. 1 に，100 回の実験で赤いセルが不完全知覚状態と判定された回数を示す．HPG では，3.3 で述べた通り，２つの迷路とも不完全知覚状態の判定が不十分であることがみられる．一方，行動選択のエントロピを用いた手法では，毎回赤いセルを不完全知覚状態と判定できている．これはエントロピを用いることで，行動選択の不確実さ，すなわち有効ルールが複数存在する状態を不完全知覚状態と判定できるからだと考えられる．

7 POMDPs 環境下での性能実験

Fig. 4 に示す Wiering[6] 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．サブゴールは壁を１，道を０として，Fig. 5 のように配列構造で表現する．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，Fig 5 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，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，各強化学習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．提案手法と HPG のパラメータは，Table 2 のように統一した．

Fig. 6 に実験結果を示す．グラフの縦軸は最短ステップ数，横

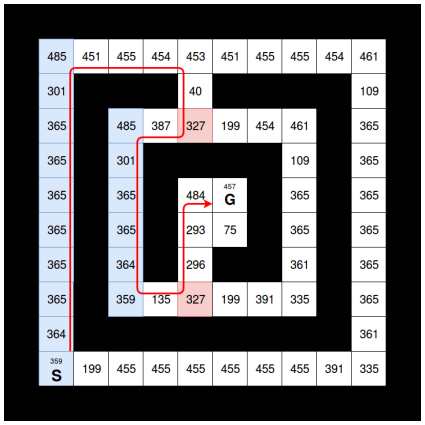


Fig. 4　The maze of Wiering

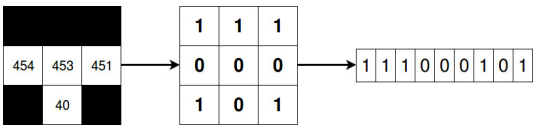


Fig. 5　Subgoals expression

Table. 2　Parameters

The number of PS trials at initial population	15000
The number of generations	49
The number of agents	50
The number of reinforcement learning trials	300
Elite preservation（%）	30
Crossover 1（%）	20
Crossover 2（%）	50
Mutation rate（%）	5
R (fitness calculation)	100
a (fitness calculation)	0.2
b (fitness calculation)	3

軸は強化学習回数である．FVPS と EPS は局所解に陥いることがあり，最短ステップで収束していない．これらの手法は，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 は最短ステップ数に収束しており．PS による不完全知覚状態判定の有効性がみられた．その中，提案手法は 6 世代目に，HPG は 14 世代目に最適解に収束しており，半数以下の試行数で学習を完了している．

Table. 3 に，提案手法と HPG の各手法が 100 回の実験中，最短経路上の不完全知覚状態をサブゴール候補とした回数を示す．両手法とも，過半数の実験で不完全知覚状態を正確に判定している．しかし，HPG では，「301」や「485」などの不完全知覚問題を回避できる状態においては，サブゴール候補となる割合が低くなっている．一方，提案手法では 9 割程度サブゴール候補となっている．これにより，提案手法は HPG より早く最適解が獲得できると考えられ，行動選択のエントロピを用いた有効性が確認された．

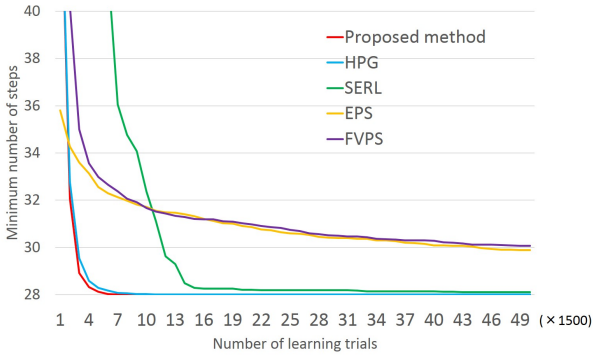


Fig. 6　The result in a POMDP environment

Table. 3　The number of appropriate subgoal candidates (the sum of 100 runs)

State ID \ Method	Proposed method	HPG
301	86	59
327	100	100
359	81	77
364	81	74
365	100	100
485	90	58

8　おわりに

本稿では，行動選択のエントロピと GA によってサブゴール創発を行い，不完全知覚問題に対応する手法を提案した．実験により，関連研究よりも早く最適解を獲得でき，また HPG より不完全知覚状態をより正確に判定できることが確認できた．今後の課題として，不完全知覚状態の判定の精度を性能を向上させること，さらに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

[1] 木村　元，宮崎和光，小林重信：創発システム研究の新たなる展開強化学習システムの設計指針，計測と制御， Vol. 38, No. 10, pp. 618–623 (1999).
[2] Whitehead, S. D. and Ballard, D. H.: Active perception and reinforcement learning, *Neural Computation*, Vol. 2, No. 4, pp. 409–419 (1990).
[3] 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).
[4] 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).
[5] 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).
[6] Wiering, M. and Schmidhuber, J.: HQ-learning, *Adaptive Behavior*, Vol. 6, No. 2, pp. 219–246 (1997).
[7] Arai, S. and Sycara, K.: Credit assignment method for learning effective stochastic policies in uncertain domains, *Proceedings of the 3rd Annual Conference on Genetic and Evolutionary Computation*, Morgan Kaufmann Publishers Inc., pp. 815–822 (2001).