2P4-5

POMDPs環境下におけるProfit SharingとGAによるサブゴール創発を行う強化学習法

Reinforcement Learning Emergenting Subgoal by Profit Sharing and Genetic Algorithm for POMDP 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

In a POMDP environment, an agent may observe the same information at more than one state. HQ-learning and Episode-based Profit Sharing (EPS) are well known methods for this problem. HQ-learning divides a POMDP environment into subtasks. EPS distributes same reward to state-action pairs in the episode when an agent achieves a goal. However, these methods have disadvantages in learning efficiency and localized solutions. In this paper, we propose a hybrid learning method combining PS and genetic algorithm. We also report the effectiveness of our method by some experiments with partially observable mazes.

1. はじめに

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 [木村 96]．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，それが原因で起こ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として，POMDPs 環境をサブゴールを用いて MDPs 環境に分割し，各サブエージェントがそれぞれ Q-learning を行うことで解決する HQ-learning[Wiering 97] や，同状態におけるルールの報酬を等しく分配することで確率的に解決する Episode-based Profit Sharing（EPS）[植村 05] が存在する．HQ-learning では，サブゴールの価値を表す HQ 値を用いてサブゴールの学習を行うが，サブゴールが変更するたびにサブゴールに到達する方策を学習し直さなければならず，学習効率が悪い．また，サブゴール数が固定であり，各環境に適したサブゴール数でなければ，不必要な学習が多くなったり，学習ができない可能性がある．EPS では，報酬分配量がゴールまでの距離に依存し，エピソード開始と終盤で報酬の値に大幅に差が生まれる．そのため，行動価値の初期値が大きいと局所解に陥りやすく，小さいと学習が遅くなってしまい，適切な初期値の設定が困難である．また行動価値を累積するため，学習終盤での解の更新は難しく，環境変化に対応できない．本稿では，これらの欠点を改善すべく強化学習の一種である 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新たな手法を提案する．

2. 不完全知覚問題

POMDPs 環境の例を図 1 に示す．図中の円は状態，a と b は行動，矢印は次状態への遷移を表している．この環境では状態 S1 と S2 は異なる状態だが，センサの観測範囲や外部影響などにより 2 つとも状態 S の同一状態と認識し，状態の混同が起こっている．このときゴールである状態 G に到達するには，状態 S1 では行動 a，状態 S2 では行動 b を行わなければならないが，エージェントは状態 S と誤認識しているため，ゴールに近い行動 b を学習してしまいループに陥ってしまう．このような 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．

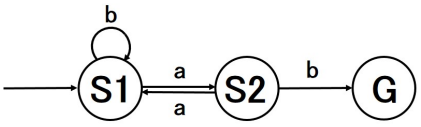

図 1: POMDPs 環境

3. POMDPs 環境下での PS と GA 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．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て POMDPs 環境の分割を行い，サブエージェントが MDPs 環境で学習を行うことで不完全知覚問題に対応する．すなわち提案手法は，サブゴールの設定とサブエージェントの学習を遺伝的アルゴリズムの中で行い，環境に適応したエージェントを生成する．提案手法のおおまかな流れを次に示す．

1. 初期集団を生成する．
   サブゴールと (サブゴール数 + 1) 個のサブエージェントをもつエージェントを生成する．
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 を世代数繰り返す．

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

初期集団生成時において，エージェントはどのような状態が環境に存在するかわからないため，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または存在しないサブゴールが多くなってしまう．それでは学習することなく必要以上にに試行回数を増やし，さらに有効なサブゴールの適応度を下げてしまう可能性があり，学習が非常に遅くなってしまう．そのため，初期集団生成時に有効なサブゴールを設定しなければならない．そこで，提案手法では強化学習の一種である Profit Sharing を用いてサブゴールを生成する．

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．提案手法では，PS のオフライン学習に着目し，報酬分配量を考慮することで不完全知覚が起こってい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，不完全知覚の判定を行うため，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，正の定数とする．これにより，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 A として全ての行動が報酬獲得に必要だと仮定すると，各行動の優先度 p は次式の値になる．

$$p = \frac{1}{|A|} \tag{2}$$

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，式 2 の値より大きい優先度の行動が複数あるとき，不完全知覚問題が起こっていて，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．そこで提案手法では，サブゴール生成に PS を用いて，式 2 の値を超えた優先度のルールが複数ある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ランダムに選択する．

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 [植村 05]．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用なサブゴールかどうか判断でき，学習時間の短縮となる．

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．

3.3 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，ゴールしたかどうかとゴールした際のステップ数をみて更新する．すなわち，学習中の最短ステップ数やゴール達成，未達成は用いない．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．

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

ここで，R はゴール報酬値，Max_step は最大ステップ数，$step$ は報酬獲得までのステップ数，$goal$ は強化学習中のゴール回数，a，b は重みを表す．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る必要がある．そこで式 3 を求めた後，同一のサブゴールをもっているエージェントの適応度を重みで除算する．同様なサブゴールを持つ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．

3.4 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きしまう．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．

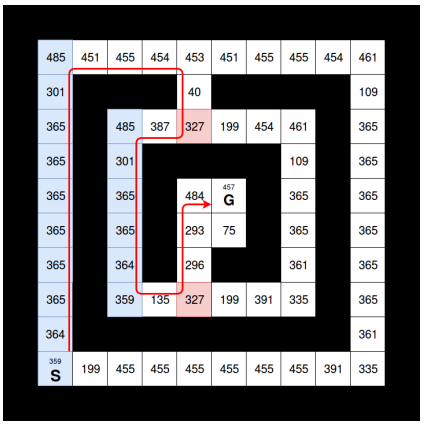

図 2: Wiering の迷路

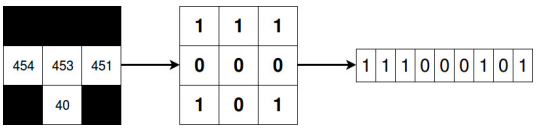

図 3: サブゴール表現

表 1: パラメータ

(a) 提案手法，提案手法（Q 学習）

サブゴール生成時 PS 試行数	15000
世代数	49
エージェント数	50
強化学習試行数	300

(b) サブゴール創発強化学習

世代数	50
エージェント数	50
強化学習試行数	300

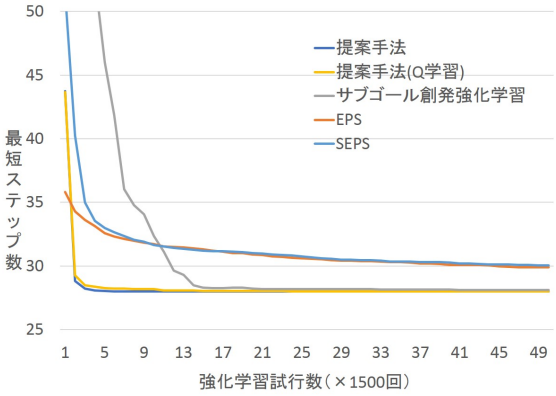

図 4: 実験結果

4. 関連研究

野村らは，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を提案した [野村 16]．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 らは，同状態において等しく報酬分配を行う First Visit Profit Sharing(FVPS)[Arai 01] を提案し，植村らが 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)[植村 05] 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 マスのみで，行動は上下左右の 4 つとする．数字が書いてあるマスが道で，黒いマスが壁であり，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，図 3 のように配列構造で表現する．道のマスに書いてある数字は，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 進数とし，それを 10 進数に変換したものである．この環境で最短経路を得るためには，青いマスと赤いマスで起こっている不完全知覚問題を解かなければならない．他の経路についても必ず不完全知覚問題が複数起こるようになっているため，必ず不完全知覚問題を解決しなければならない．ここでは，提案手法，サブゴール創発強化学習，EPS，SEPS，提案手法の強化学習を Q-learning に変更した 5 種類の手法で比較実験を行う．各手法の試行回数は，表 1 のとおり強化学習を行った回数で統一し 75 万回とする．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 とし，また実験は 100 回行いその平均をとる．

実験結果を図 4 に示す．この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 は局所解に陥いるときもあり，どれも最短ステップで収束していない．EPS では，経験に固執して解の更新ができないため，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．サブゴール創発強化学習では，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，無駄なサブゴールが多々創発され，75 万回では適切なサブゴールが見つからなかったことが考えられる．一方提案手法では最短ステップ数に収束している．また学習の速さに着目すると，EPS の学習の立ち上がりが一番速い．提案手法は EPS には及ばなかったが，同じサブゴールを創発するサブゴール創発強化学習より速い立ち上がりとなっている．学習の収束については提案手法が一番速い．Q-learning を用いた提案手法と比較すると，あまり差はないが PS を用いた提案手法の方が若干速くなっている．PS の学習の速さに加え，POMDPs 環境にも対応できたのが要因だと考えられる．

6. 環境変化に対しての性能実験

環境変化が起こる POMDPs 環境での実験を行い提案手法の有効性を検証する．前実験と同様に Wiering の迷路を用い，強化学習 45 万回した後に図 5 のように最短経路の一部を閉鎖する．その他の設定は前実験と同様である．この実験は，提案手法，サブゴール創発強化学習，EPS，学習を Q-learning に変えた提案手法の 4 種類で行う．環境の変化により，最短ステップを学習していた場合には学習終盤での大幅な経路の変更を要し，また ”455 ”のマスで新たに発生する不完全知覚問題にも対応しなければならない．そのため価値を累積していく通常の強化学習では対応しにくい．

実験結果を図 6 に示す．優先度を累積する EPS は環境変化に対応できず，環境変化後にゴール到達することはなかった．サブゴール創発強化学習と Q-learning を用いた提案手法では，環境変化後はゆっくり学習がすすんでいるように見える．しかし，これは実験 100 回の平均であり，実際は最適解に近いス

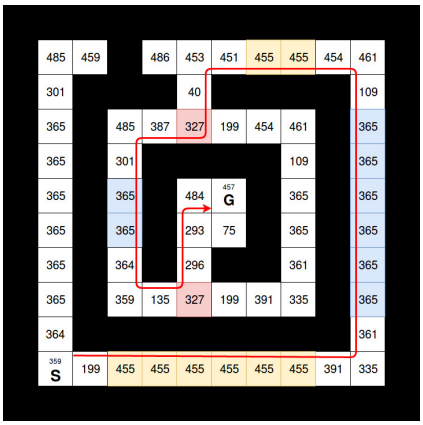

図 5: 環境変化後の迷路

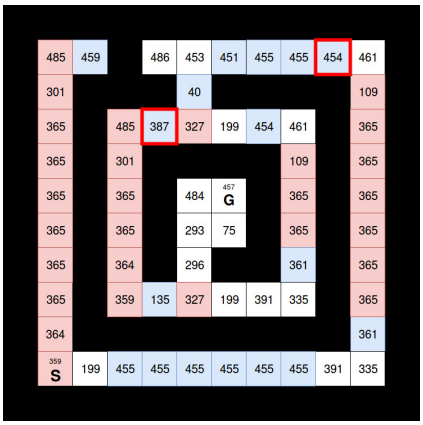

図 7: 生成されたサブゴールの分布

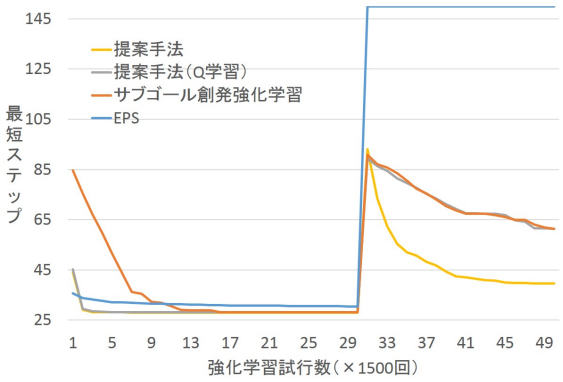

図 6: 実験 2 の結果

テップ数かゴールしていないかのどちらかになっている．環境変化後は，不完全知覚が起こっている”455”のマスの間にサブゴールを作らなければならないが，双方とも環境変化前の最短経路を学習しているため，そのサブゴールをもつエージェントは存在しない．そのため突然変異で新たに創発する必要がある．突然変異でうまく生成できたときに最適解に近くなり，反対にできなかったとき最大ステップ数となる．これは提案手法でも同様のことである．しかし提案手法では，ほぼ最短ステップ数を獲得している．これは報酬分配量を等しくした PS によって不完全知覚に対応できたことが大きな要因と考えられる．Q-learning では，不完全知覚があると学習中にゴールできないため，サブゴールが有効かどうかわからない．しかし PS では，学習中にゴールできるため適応度が上がり，有効なサブゴールかどうかは判断できる．そして交叉を行うたびに有用なサブゴールが組み合わさり，MDPs 環境に分割できた結果，最適解に近い解となった．

図 7 に提案手法で生成されたサブゴールの分布を示す．赤いマスが初期集団生成時に用いるサブゴール候補であり，青いマスが最終世代のエージェントが所有していたサブゴール候補に存在しないサブゴールである．また赤く囲ってあるマスは，最終世代の優良個体が所有していたサブゴールを示す．サブゴール候補においては不完全知覚が起こっている状態をほぼ網羅している．突然変異では，環境変化後にサブゴール候補に存在しないサブゴールを生成する必要があったが，必要なサブゴールが生成され，優良個体のサブゴールも無駄なく MDPs 環境に分割できている．また図 7 に示したサブゴールは最終世代のみのサブゴールを示しているが，環境変化後の 20 世代全体では，突然変異により迷路に存在する状態がすべてサブゴール生成されたことも確認された．

7. おわりに

本稿では，不完全知覚問題において新たな解決法を提案した．実験により，従来手法より速く最適解に収束し，また環境変化にも対応でき，提案手法の有効性を確認した．今後の課題として，次の 2 つが挙げられる．1 つ目は，初期集団生成時に適切なサブゴールが存在しない状況の対処法である．提案手法では突然変異のみで対処していたが，ランダムなため確実に適切なサブゴールを生成する保証がない点について検討する必要がある．2 つ目は環境変化実験においての学習法である．提案手法は環境変化後に学習し直していることが多く学習効率が悪い．GA の個体に多様性が存在しゴールまでの経路が複数あれば，学習する必要はなくなる．提案手法では同様のサブゴールをもつエージェントの適応度を低くしているが，他の方法を検討しなければならない．

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)

[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)

[木村 96] 木村元, 山村雅幸, 小林重信：部分観測マルコフ決定過程下での強化学習: 確率的傾斜法による接近, 人工知能学会誌, Vol. 11, No. 5, pp. 761–768 (1996)

[野村 16] 野村 拓己, 加藤 昇平：問題分割と遷移規則の創発による POMDP 環境下での強化学習, 研究報告知能システム (ICS), Vol. 2016, No. 19, pp. 1–6 (2016)