6R-03

# POMDP 環境下でのサブゴール創発による強化学習の動的階層化

野村拓己 † 加藤昇平 †

† 名古屋工業大学

## 1 はじめに

近年，強化学習の研究が盛んに行われている．強化学習は，学習エージェントが試行錯誤を通じて制御則を獲得する機械学習の一種である．学習エージェント自身が制御則を学習・獲得するため，効率的な制御則を発見する可能性も考えられる．そのため，ロボットの自律的な行動獲得などに強化学習が多く使われている．強化学習では観測情報から環境を一意に特定し，1つの行動を結びつけるように学習する．そのため，異なる環境から知覚した観測情報が同一である場合，それらの異なる環境を同一の環境と認識し学習する [1]．同一と認識された個々の環境において最適な行動が異なる場合，学習が混同してしまい正しく学習が行えない．学習器が環境を部分的に観測するとき上記の問題が発生する．このような問題を持つ環境を部分観測マルコフ決定過程（POMDP）と呼ぶ．

POMDP 環境下における問題解決手法として，我々は遺伝的アルゴリズム（GA）を用いたサブゴールの動的生成手法を提案している [2]．しかし前手法ではサブゴールをビットベクトルで表現していたため，観測空間が増大するとサブゴールの探索空間が指数関数的に増加する．加えて，観測が連続値の環境で用いることができない．そこでサブゴールを不等式と論理表現を用いた条件式で表現する手法を提案する．本稿では比較実験により，提案手法の有効性を検証する．

## 2 POMDP

POMDP では不完全知覚により混同が発生する [3]．例を図 1 のネットワークで示す．ノードは状態を示す．ただし状態 $A1$ と $A2$ は観測情報が共に状態 A として観測される．アークは行動の種類を示す．

図 1 において状態 $A2$ では行動 a は非合理的ルールであるが，$A1$ では行動 a は合理的ルールとなる．学習器は $A1$ と $A2$ を同一の状態（状態 $A$）として観測するため，状態 $A$ で行動 a は学習器にとって合理的ルールとされる．その結果 $A2$ と $B$ を往復する非合理的な政策が学習される．

## 3 提案手法

エージェントは $m$ 個 ($m \geq 2$) のサブエージェントを持つ．サブエージェントは図 2 に示すように，サブゴール条件と Q テーブルの対を持つ．サブエージェントの学習には Q-Learning を使用する．行動選択には $\epsilon - greedy$ 法を用い，式 (1) により Q 値の更新を行う．

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max_{a' \in A} Q(s_{t+1}, a') - Q(s_t, a_t)] \quad (1)$$

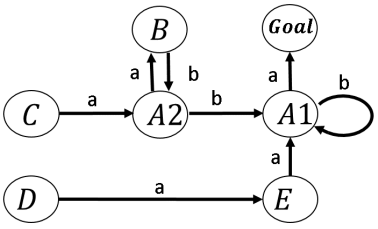


図 1: POMDP における混同

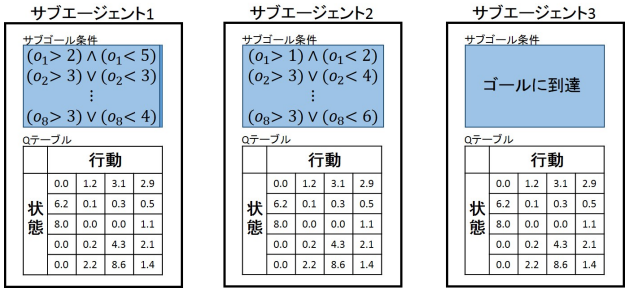


図 2: 遺伝子構造

### 3.1 問題設定

本稿では Grid-world 迷路問題を扱う．環境は壁または道で構成され，周囲 8 方向の壁までのグリッドの数を観測する．観測可能距離内に壁が存在しないとき，グリッド数は観測可能距離+1 とする．学習エージェントは座標を取得できないため場所が異なるが観測情報が同じとなる状況（POMDP）が存在する．選択できる行動は「上」「下」「左」「右」の 4 通りであり，壁に向かって進む場合はその場に留まるが 1 ステップとして数える．

### 3.2 サブゴール条件

サブゴール条件はサブゴールと判定するために用いられる．本稿では式（2）に示すように，観測情報 1 つにつき 2 つの不等式を論理和または論理積で結合して表現する．

$$(o_i > x1_i) R_i (o_i < x2_i) \quad (2)$$

ここで $o_i$ は $i$ 番目の観測情報を，$R_i$ は論理和または論理積を示す．$x1_i$，$x2_i$ は観測情報と同じ値域の変数である．遺伝的操作により $R_i$ および $x1_i$，$x2_i$ を学習する．$\forall i.[(o_i > x1_i) R_i (o_i < x2_i)]$ 式が真となったときサブゴール到達と判定する．

### 3.3 サブエージェント

サブエージェントはサブゴール条件と Q テーブルの組み合わせを持つ．エージェントは $m$ 個のサブエージェントを配列構造で保持している．学習開始時，1 つ目のサブエージェントが動作する．サブゴール条件をゴールとみなし Q 学習が行われる．このとき 1 回の行動を 1step とする．サブゴール到達と判定されたとき，次のサブエージェントが動作する．最後のサブエージェントがゴールに到達したとき，タスク達成となる．最大ステップ数 (Maxstep) に達したとき，その試行を終了しタスク失敗となる．ここまでの流れを 1 試行と

---

*Dynamic Hierarchy of Reinforcement Learning by Subgoals Emergence under POMDP, Takumi NOMURA†, and Shohei KATO†

†Nagoya Institute of Technology
Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan
{nomura, shohey}@katolab.nitech.ac.jp

する．

### 3.4 適応度

エージェント $i$ の適応度 $F(i)$ を式（3）で定義する．

$$F(i) = \begin{cases} \mathrm{R} + (\mathrm{Maxstep} - step_i)/\mathrm{b} & (\mathrm{complete}) \\ \mathrm{r} + goal_i/\mathrm{a} & (\mathrm{incomplete}) \end{cases} \quad (3)$$

ここで，r は最低報酬値，R はゴール報酬値，$goal_i$ はゴール回数，Maxstep は最大ステップ数，$step_i$ はステップ数，a と b は重みを表す．学習終了後 greedy 法で行動を選択し 1 試行する．このときゴールできたエージェントを学習完了（complete），ゴールできなかったエージェントを学習未完了（incomplete）とする．学習未完了の個体は低い適応度を与えるが，試行中にゴールができた個体は達成できないサブゴールが存在しないため，多少の報酬を与える．また数多くゴールできた場合ゴールしやすいサブゴールが設定されている可能性があるため，ゴール数に応じて報酬を与えた．

### 3.5 交叉

サブゴール条件と，サブエージェントのそれぞれについて交叉を行う．ルーレット選択により親エージェント $i$ および子エージェント $j$ を選択する．サブゴール条件の交叉（交叉 1）は，エージェント $i, j$ のサブエージェントのサブゴール条件 $\mathrm{S_i}$ と $\mathrm{S_j}$ を一様交叉をする．しかしエージェントの持つサブエージェントの数が異なるため，$|\mathrm{S_i}| = |\mathrm{S_j}|$ となるように $\mathrm{S_j}$ を調整する．確率 $\epsilon_{GA}$ でサブエージェントの $R$，$x1$ または $x2$ が突然変異する．この交叉により生成されたエージェントのサブエージェントの Q テーブルは初期化する．

サブエージェントの交叉（交叉 2）は，エージェント $i, j$ のサブエージェントの順序付き集合 $\mathrm{S_i}$ と $\mathrm{S_j}$ を一点交叉する．エージェント $i$ の前部 $\mathrm{S_i[k]}$ $(0 < k \leq p_i)$ とのエージェント $j$ の後部 $\mathrm{S_j[h]}$ $(p_j \leq h < |S_j|)$ を結合する．$\mathrm{p_i}$ は $0 < p_i \leq |\mathrm{S_j}|$ の範囲でランダムに選択する．これによりサブエージェントの数が動的に変化する．確率 $\epsilon_{GA}$ でサブエージェントの追加または削除がされる．この交叉により生成されたエージェントのサブエージェントは親個体の Q テーブルをそれぞれ引き継ぐ．

### 3.6 世代交代

次世代に残す個体はエリート個体，交叉 1 および交叉 2 によって生成された個体である．これらの個体の構成比は，エリート個体の割合を増やすと多様性を保持しやすい反面学習効率が低下する．交叉 1 により生成された個体の割合を増やすと新しいサブゴールの発見率が増加する．交叉 2 により生成された個体の割合を増やすと前世代に近しいルートの探索率が増加する．

## 4 実験

Wiering ら [4] が作成した 12 × 12 迷路（図 3）を利用して比較実験を行った．パラメータは学習率 $\alpha$=0.9, 割引率 $\gamma$=0.7．次世代に残す個体率はエリート保存が 0.3，交叉 1 が 0.2，交叉 2 が 0.5．世代数 50，エージェント数 50，エージェントの Q-Learning の試行数 300，最大ステップ数 100，突然変異確率 $\epsilon_{GA}$=0.5．初期サブエージェント数 4 とした．比較対象として前手法 (previous) と前手法と同等となるよう条件文を等式としたもの (proposal2) を用いる．図 4 に実験結果を示す．proposal1 は収束速度が劣っているが最終的にはともに最適解である 28 に収束した．これは観測範囲が狭いため不等式による表現で容易に全範囲がカバーされるため

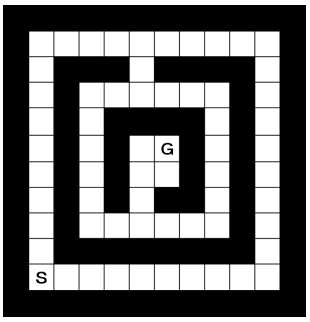


図 3: 12 × 12 迷路

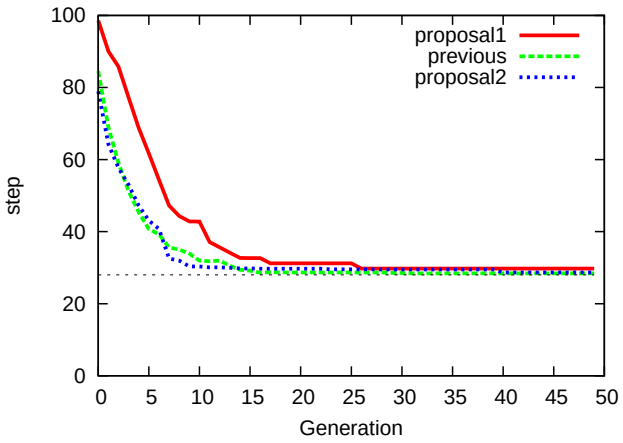


図 4: 実験結果 50 回平均

と考えられる．previous と proposal2 はほぼ同じ曲線が得られた．これは proposal2 が previous と同じ学習がされたことを示している．これにより本手法は前手法を包含できていることが分かる．

## 5 おわりに

本稿では POMDP に対して，サブゴール条件及びその組合せを GA で自律獲得する手法を提案した．本稿では観測範囲が狭い実験のみを行ったが，今後は範囲を広げた実験を行い不等式によるサブゴール条件の有効性を調べていく．またサブゴール条件が範囲で指定できるため連続値を扱うことができる．今後は観測空間が連続値となる環境でも実験を行う．

## 謝辞

本研究は，一部，文部科学省科学研究費補助金（課題番号 25280100，および，25540146）の助成により行われた

## 参考文献

[1] WHITEHEAD, S. D.: Learning to Perceive and Act by Trial and Error, *Machine Learning*, Vol. 7, pp. 45–83 (1991).
[2] 野村拓己, 加藤昇平：POMDP 環境下での強化学習における GA による問題分割, 第 13 回情報科学技術フォーラム, Vol. 13, No. 2, pp. 61–66 (2014).
[3] 宮崎和光, 荒井幸代, 小林重信：POMDPs 環境下での決定的政策の学習, 人工知能学会誌, Vol. 14, No. 1, pp. 148–156 (1999).
[4] WIERING, M.: HQ-learning, *Adaptive Behavior*, Vol. 6, No. 2, pp. 219–246 (1998).