F-037

# 動的環境におけるマクロオペレータと副報酬
# Macrooperator and Sub-rewards in Dynamical Environment

**武藤 真司** † **鈴木 輝彦** ‡ **太原 育夫** ‡
Shinji Muto Teruhiko Suzuki Ikuo Tahara

## 1. はじめに

強化学習は，エージェントが環境から報酬を得ることによって自律的に学習する方法である．強化学習を現実に即した様々な問題に適用する場合，多くの環境は常に変動する可能性があるということを意識する必要がある．

事前に学習した内容のうち環境変化後も利用できる部分を残して環境変化後の学習内容と組み合わせて新たな学習結果を得るという手法にマクロオペレータを用いた強化学習がある [1]．

マクロオペレータ (以下 MO と略記する) を用いることで動的環境における学習の効率化が図れることも示されているが，いくつかの問題点も指摘されている [2]．

本研究では，こうした問題点のうち

1. MO の形成に長期的な事前学習が必要である．
2. 環境変化直後はマクロオペレータに固執してしまうため逆に時間がかかる可能性がある．

といった 2 点に焦点をあて，強化学習において様々な問題解決に役立つとされる副報酬 [3] を導入することによりこれら問題点にどのような影響があるか検討する．

## 2. マクロオペレータの使用

MO は時系列的に連続して適用される複数のルール，または別の MO を構成要素として持つルール群の集まりである．エージェントは，学習する過程で適しているルールをマクロとして保存し，それらマクロを組み合わせて問題解決に利用することができる．

MO は本研究が対象とする動的環境においても，分割再結合を繰り返すことにより柔軟に対応可能である [2]．

エージェントは図 1 に示すような流れで MO を利用する．すなわち，ある状態 $s$ で行動 $a$ を選択し次の状態が $s'$ であるようなルールを “$sas'$” とするとき，このルールが MO の有効な構成要素であればこの MO を利用して行動決定する．そして，図 1 の②において以下の条件を満たしたときその MO に従うことをやめて従来の行動選択法に従って行動決定をする．

- 状態遷移後の環境がルールの予測と異なる．
- 実行すべきルールが unusable な構成要素である．
- MO の持つすべてのルールを実行し終えた．

なお，MO の分割結合は，予測される遷移先への遷移確率や平均報酬に基づいて行われる．この分割結合条件には報酬の概念があるため，副報酬を環境上に与えることで MO の分割結合にある程度の影響を与えることが期待できる．

†東京理科大学大学院理工学研究科情報科学専攻
‡東京理科大学理工学部情報科学科

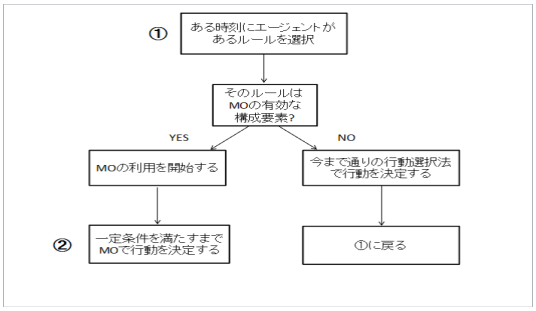


図 1: マクロオペレータの利用

## 3. 副報酬の導入実験

副報酬の MO に与える影響を実験により検討する．通常の MO を用いたエージェント (MA) と MO と環境に副報酬を導入したエージェント (MSA) を用意し，双方に対して同様の実験を行い比較する．実験環境は図 2 に示すような 13 × 13 の迷路環境とする．

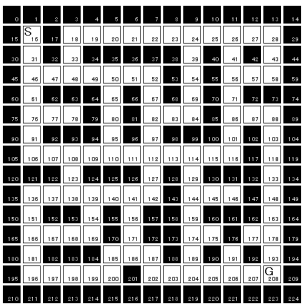

図 2: 実験に用いる基本環境 ($M_1$)

この基本環境に対して以下のように副報酬を与える．そして環境の変化を与える．

1. 基本環境 $M_1$ に対して，副報酬をランダムに 10 箇所与えた環境を $MS_1$ とする．
2. $M_1$，$MS_1$ に対してエージェントは最短経路 (24step)，各迂回経路 (26step，28step，30step) を 100Episode ずつ学習する．
3. $M_1$，$MS_1$ を変化させた $M_2$，$MS_2$ に対してエージェントに 200Episode 学習させる．

ここで，$M_2$ と $MS_2$ は以下のように設定された環境である．

1. 実験 1 - 最短経路 1 つを封鎖

2. 実験 2 - 最短経路 2 つを封鎖
3. 実験 3 - 最短経路 3 つを封鎖
4. 実験 4 - 最短経路 4 つを封鎖
5. 実験 5 - 最短経路 4 つを封鎖し，さらに 3 つの迂回経路を封鎖

Q 学習のパラメータは，学習率を 0.5，報酬の割引率を 0.85 に設定し，目的地到着で 100.0 の報酬，壁である状態に遷移しようとした場合は-1.1，副報酬の状態に遷移した場合は 5.0，それ以外の状態では-1.0 の報酬が得られるものとする．

## 4. 実験結果

MA と MSA の双方に対して前述の実験を行ったところ，副報酬の効果が目に見えて分かる部分とそうでない部分があった．表 1 に初期 MO の構成エピソード数を示す．このように各エージェントが全く MO を構成していない状態での学習収束速度，すなわち MO の構成速度は MSA の方が 20Episode 以上高い結果となっている．100Episode 以降の事前学習の速度に関しては MA と MSA にそこまでの違いは見られなかったことから全体的な事前学習時間は MSA の方が僅かに早いものと考えられる．

表 1: 初期 MO の構成エピソード数

| MA | MSA |
|---|---|
| 76.4 | 52.2 |

本実験での迷路環境は，最短経路に関わる状態が目的地到達に一切かかわらない状態に比べて多いので，無作為に副報酬を与えたとしても最短経路，または目的地到達に関わる状態に副報酬が配置される可能性が高い．そのため，副報酬によって多く報酬を得ることにより該当する状態を経由する確率があがり，エージェントはその状態への遷移が有効なものだと捉え MO の一部として組み込むといった流れを経るので，結果的に MO の構成速度が上がったものと考えられる．

つぎに，事前学習後の各実験の結果として最短経路発見までのエピソード数の比較を行った．

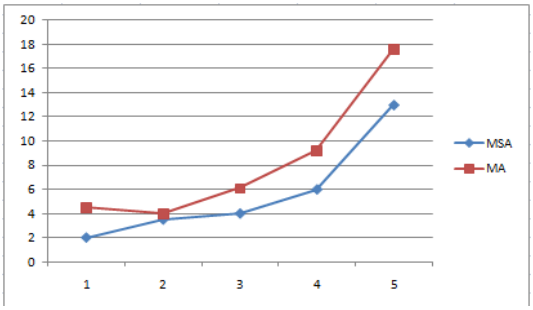


図 3: 各実験における最短経路発見までの Episode 数

図 3 のように全体的に多少早くなっているため，副報酬は環境変化後の再学習の高速化にも影響していると考えられる．

本実験において副報酬を導入した際，以下のような現象が観測された．

1. 通常 MO は目的地付近の状態から形成されていくが，副報酬を無作為に配置することによって思わぬところにマクロが形成される．
2. 環境変化後，道だった部分が壁になりその隣の状態に副報酬が置かれた場合，有効なルールではないにもかかわらず分割が円滑に行われない．

1 の現象は本実験のように環境変化が比較的小規模な場合はあまり重要でないと思われるが，もし規模の大きい環境変化を意識した場合には再学習に役に立つ可能性が大いにある．

2 の現象から袋小路が多いような迷路に対して副報酬を導入すると，MO の分割を妨げ結果的に学習速度が落ちるといったことが考えられるが，副報酬を正の値ではなく負の値にするなど導入する副報酬を変化させることにより対応できると考えられる．しかし，本実験のように行き止まりが比較的少なく最短経路がいくつも存在するような迷路では，過度に分割を行ってしまったり，結合するべきときに結合を行わないといったことが起こるので負の副報酬を導入するのは逆効果である．

## 5. おわりに

動的環境に対する効果的な手法である MO と強化学習における様々な問題に対して有効であるとされる副報酬の関係性について検討した．小規模な環境変化において MO の有効性はすでに示されてきたが，副報酬を導入することでさらに効率化できることが明らかになった．しかし，現実の動的環境問題は決して小規模なものばかりではない．今後は，MO を大規模な環境変化問題に対応させ，また副報酬がそれら大規模な環境変化に対してどこまで影響するのかを調べていく必要があると考えられる．

## 参考文献

[1] 嶋田総太郎，安西祐一郎，“マクロオペレータの部分的再利用による強化学習システムの動的環境への適応能力の改善，” 電子情報通信学会論文誌, D-I, vol.J84-D-I, no.7, pp.1076–1088, 2001.

[2] 高尾大夢，延澤志保，太原育夫，“マクロオペレータを用いた強化学習結果の他環境への適用，” 情報科学技術フォーラム一般講演論文集 3(2), pp.299–300, 2004-08-20.

[3] 菅井克義，延澤志保，太原育夫，“強化学習における副報酬の役割，” 電子情報通信学会総合大会講演論文集 2007 年 情報システム (1), p.105, 2007-03-07.