ISSN 2186-5647



P-17

# 目的地の方向を考慮した強化学習による経路学習

日大生産工　○山崎　涼太　日大生産工　山内 ゆかり

## 1 まえがき

強化学習は未知の環境下で試行錯誤を繰り返すことによって最適な手法を見つけ出すことを目的とする。[1]

状態数が多い環境内では学習するまでに膨大な時間が必要となる。実問題で試行錯誤を繰り返すことは、コストがかかる等の問題があり困難である。よって学習の高速化は極めて重要な課題である。清本らによる部分観測マルコフ決定過程において、エージェントの状態行動をエピソードとして保存する Profit Sharing Plan で位置ベクトルを導入した手法において、迂回行動に対する報酬割り当ての抑制により改善される場合があることが確認されている。[2]

本研究は Profit Sharing Plan において迂回経路の削除をするとともに、ゴールの位置情報を与えることによって学習の高速化を試みる。

## 2 提案手法および実験方法

提案手法の有効性を検証するため、迂回経路の削除をするProfit Sharingと提案手法の2パターンを迷路探索にて比較していく。

### 2.1実験環境

以下に本研究で扱う迷路マップをどのように表銘パターンとして表現するかを説明する。

**表1 迷路の定義**

| -1 | 0 | 2 | 4 |
|---|---|---|---|
| 移動不可能なマス | 移動可能なマス | スタート位置 | ゴール位置 |

上記のようにそれぞれの値に対して対応しているものがある。図１に具体的なマップを示す。

| | | | | | |
|---|---|---|---|---|---|
| -1 | -1 | -1 | -1 | -1 | -1 |
| -1 | 2 | 0 | 0 | 0 | -1 |
| -1 | 0 | 0 | 0 | 0 | -1 |
| -1 | 0 | 0 | 0 | 0 | -1 |
| -1 | 0 | 0 | 0 | 4 | -1 |
| -1 | -1 | -1 | -1 | -1 | -1 |

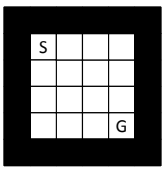


**図 1 表銘の例**

今回の実験に使用したマップは図3に示す。

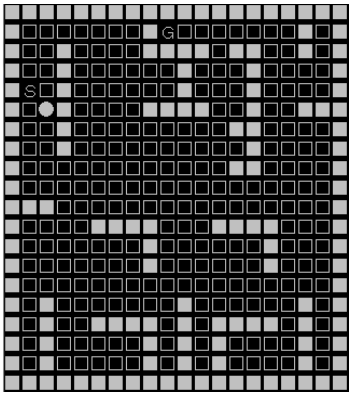


**図 2 マップ環境(サイズ 20×20)**

図 2 はエージェントがマップ内を探索している様子である。移動不可能なマスは白で表記され、エージェントは◯で表している。マップのサイズは縦 20(Y 軸)、横 20(X 軸)の範囲である。スタート地点は Y=4、X=1 の位置で、目的地は Y=1、X=9 の位置である。目的地の情報を与えるためスタート地点からゴールまでに障害物が全く無い場合のマップだと提案手法の方が従来に比べて優位な条件となってしまうため、ゴールの最短経路上に障害物を配置した。

### 2.2 提案手法

本研究では目的地の方向を考慮することで最短経路獲得までの行動回数の削減を目指す。具体的にはエージェントに目的地の方角の情報を与え、行動選択の時利用させる。エージェントの基本性能として行動は上下左右の4方向が可能である。行動選択はルーレット選択とタブー探索を併用する。スタート地点からゴールに着くまでを１試行としゴールに着いたら報酬が与えられ学習し、また試行を繰り返す。この流れを最短経路を獲得するまで繰り返す。

選択行動確率の更新はProfit Sharingに基づき(1),(2)式で行う。

$$\omega_{new} = \omega_{old} + f_i \qquad (1)$$

$$f_i = r\gamma^i \left(0 < \gamma < 1\right) \qquad (2)$$

---

The route learning by reinforcement learning that takes into account the destination

Ryota YAMAZAKI and Yukari YAMAUCHI

この時の、ωは重み、rは報酬、γは学習率である。

エージェントの今いる位置をx座標とy座標で表し、ゴールのx座標y座標に向かって行動させるようにとする。そのため目的地の方向への行動が選択されやすいようにゴールの方向に行動する選択確率を2倍へと増加させた。スタート地点から目的地への方向をエージェントに教えてしまうと初めから目的地への方向へと進みやすくなるが、途中で障害物にぶつかるマップでは最適解がでずに局所解へと陥りやすくなるという問題がある。そのため、ある程度行動をしてから目的地の方向を知らせることとした。目的の位置をエージェントが知覚できるのは目的地からマップサイズの半分あたりに差し掛かった時に知らせるものとした。式を以下の(3)に示す。

$$\frac{\text{マップの縦}\times\text{横}}{(\text{マップの縦}+\text{横})/2} \qquad (3)$$

以下に実験に用いた各係数を以下に示す。

報酬 1
学習率 0.8
最大行動回数 10000
最大試行回数 500

## 3 実験結果および検討

Profit Sharingと提案手法での比較実験を行った。比較対象となるデータは各試行回数での行動回数、最短行動回数を出すまでの総行動回数、最短行動回数を出すまでの平均試行回数の3つである。スタートから最短行動回数を出すまでを1シミュレーションとし、それぞれの各データを10シミュレーションの平均を取り、図3、表2、3に示す。

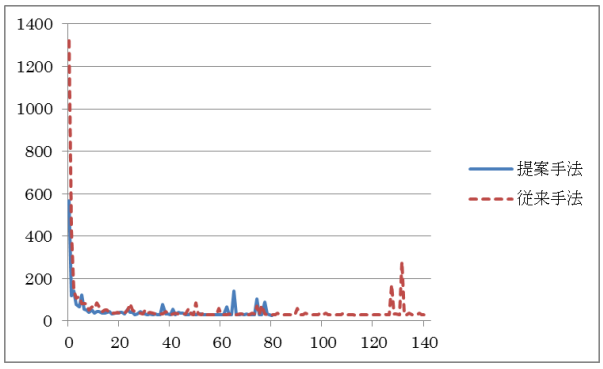


**図 3 各試行回数における行動回数**

図3はY軸が行動回数、X軸が試行回数を示している。図3を見ると提案手法、従来手法ともに早い試行回数で学習していることが読み取れるが、最短行動回数を出すまでに提案手法の試行回数は最多で約80回、従来手法は最多で約140回となっている。

**表2 最短行動回数を出すまでの総行動回数**

| | 従来 | 提案 |
|---|---|---|
| 総行動回数 | 6940,5 | 3635 |

表2は従来手法と提案手法の最短行動回数が出るまでの総行動回数の比較データである。総行動回数も従来に比べると約3300ステップも減少できていることが確認できる。この事から早い段階で学習していることが推測できる。

**表3 最短行動回数を出すまでの試行回数**

| | 従来 | 提案 |
|---|---|---|
| 平均試行回数 | 90.53 | 74.42 |

表3は従来手法と提案手法の最短行動回数が出るまでの試行回数の比較データである。平均試行回数も従来と比べると約15減少していることがわかる。

上記の3つ全てにおいて従来手法より提案手法の方が良い結果を得ることができた。

## 4 まとめ

本研究では従来のProfit Sharingに目的地の情報を与え学習の高速化を試みる提案をした。試行回数、総行動回数ともに減少した結果が得られた。今後の課題としてマップサイズを大きくした場合にも試行回数が減少するかを検証していきたい。

「参考文献」

[1]Richard S.Sutton、Andrew G.Barto(著)、三上 貞芳、 皆川 雅章(訳)、 強化学習、 森北出版(株) (2000/12)

[2]清本盛明、亀井且有,部分観測マルコフ決定過程における位置ベクトルを用いた強化学習手法の提案,システム制御情報学会論文誌、Vol.14,N0.2,pp.86-91, 2001;

[3]強化学習の基礎
東京工業大学 精密工業研究所 科学技術振興士業弾 ERATO 川人学習動態脳プロジェクト;
http://www.jnns.org/niss/2000/text/koike2.pdf
（最終アクセス日時 2013/10/27）