ISSN 2186-5647



2-11

# 体験型好奇心探索の提案

日大生産工　○山口　和馬　　日大生産工　山内　ゆかり

## 1. まえがき

強化学習アルゴリズムは、環境から与えられる報酬を最大化することによって、目標タスクを達成するための方針、行動を学習することを目的としている。現在、代表的な強化学習手法として、Q-Learning、深層学習を組み合わせたDeep Q Networkなどがあげられる。そして、Atari2600の一部の環境で人間を上回るスコアを獲得しており、実世界の複雑な問題にも対応可能だと言われている。しかし、実世界においてエージェントが報酬を受け取るということが非常にまばらであるか、まったくないため失敗する傾向にある。

Yuri Burdaらは、ランダムに初期化されたニューラルネットワークを使い、エージェントに好奇心という探索ボーナスを生成するRandom Network Distillation [1]（RND）を提案することで、強化学習にとって難しい問題であるとされるMontezuma's RevengeというAtariのゲームにおいて、人間の平均以上のパフォーマンスを達成したと報告されている。

しかし、従来手法であるRNDでは環境の新規性を好奇心の度合いとして用いている。しかしエージェントがどうやってその場所に来たかという行動には考慮されていない。

そのため本研究では，上記の問題において行動の目新しさを考慮するために、好奇心を出力するためのネットワークに一時刻前の行動も入力としていれることで、様々な状態での好奇心を探索し、局所解に陥りづらくする。これにより、従来手法であるRNDよりも、最適な行動をするようにし、学習時間が短縮できるようにする。

## 2. 従来研究

### 2-1.Q-Network

Q-Networkは強化学習の学習手法の1つであり、環境の状態をネットワークの入力として行動価値を出力する手法である。このアルゴリズムを図１に示す。

学習方法はニューラルネットワークと同様であり、その行動によって得られる報酬を教師とする。学習時に選ばれた行動にはその教師に近づくように学習をし、選ばれなかった行動には学習をさせないよう、出力をそのまま教師として与えるため以下の式のように教師$T$を作る。

・選ばれた場合

$$T_{(s_t,a_t)} = Q_{(s_t,a_t)} + \alpha \times (R + \gamma \times \max Q_{(s_{t+1},a)} - Q_{(s_t,a_t)}) \quad (1)$$

・選ばれなかった場合

$$T_{(s_t,a)} = Q_{(s_t,a)} \qquad a \neq a_t \quad (2)$$

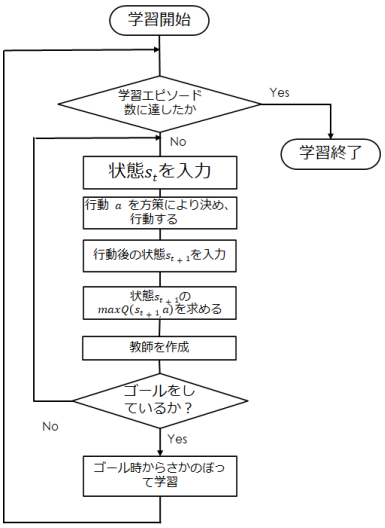


図 1.Q-network のアルゴリズム

### 2-2.Random Network Distillation

好奇心探索[2]は、未知の状態を訪問することで得られる内部報酬を設定し、未知の状態を積極的に探索する手法のことである。Random Network Distillation（RND）は状態を入力とするPredictor NetworkとTarget Networkの2種類のニューラルネットワークを用いて好奇心となる内部報酬を出力する。Target Networkはランダムに初期化後、学習を行わないネットワークであり、Predictor NetworkはTarget Networkの出力を予測し、Target Networkに近づくよう学習するネットワークである。その2種類のニューラルネットワークの予測誤差を内部報酬として生成する。それを図2に示す。

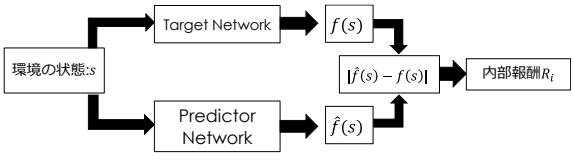


図2.内部報酬の生成

内部報酬と外部報酬の合成値をどのように推定するかは自明ではないとされているが、

Proposal for Experiential Curiosity Exploration
Kazuma YAMAGUCHI and Yukari YAMAUCHI

Burdaらは式(3)のように環境からの報酬$R_e$と内部報酬$R_i$の合計を全体の報酬$R_t$としている。

$$R_t = R_e + R_i \qquad (3)$$

経験が少なければ誤差が大きく、内部報酬が高くなるため積極的に探索をする。反対に経験を積むごとに誤差が少なくなり内部報酬は減少していくため、環境からの報酬を最大化するように学習をする。

## 3. 提案手法

従来手法 RND では状態の新規性に関して考慮しているが、どこから来たかという行動の新規性を考慮していない。そのため、本研究では従来手法の RND に加え、どこから移動してきたかという情報を入力に合わせ内部報酬を出力する手法を 2 つ提案する。

1 つ目の提案は、現在の状態空間の入力に一時刻前の行動を加えたものを入力とする。そうすることで、どこから移動してきたかを入力として表現する手法である。

2 つ目は、現在の状態空間の入力に加え、一時刻前の状態を加えたものを入力とする。そうして疑似的にどこから移動してきたかという情報を入力として与える手法である。

## 4. 実験および検討

本研究の実験手法として、9×9 のトーラス上のグリッドワールドで行う。この実験環境を図 3 に示す。

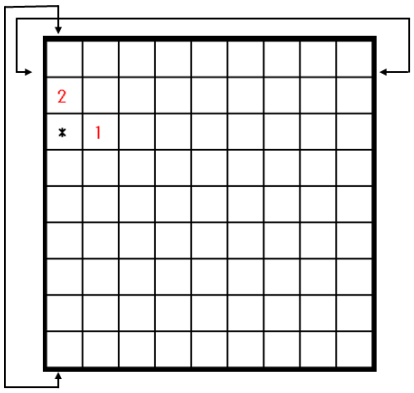


**図3.実験環境**

図3のようにハンター2体、エネミー1体をランダムに配置し、ハンターがエネミーの周囲を囲むことでゴールとし報酬を得る。エージェントの最大行動回数を1000回とし、5000エピソード学習を行う。その学習を好奇心無し、従来手法、提案手法1、提案手法2の 4 種類で10回ずつ繰り返し、学習時間と総行動回数の平均を求めそれぞれ比較する。実験結果を以下の表1に示す。

**表 1. 学習結果の 10 回平均**

| | 学習時間(秒) | 総行動回数 |
|---|---|---|
| 好奇心無し | 533.1 | 826465.4 |
| 従来手法 | 1059.5 | 781532.0 |
| 提案手法1 | 1066.1 | 780660.3 |
| 提案手法2 | 1090.3 | 804003.6 |

今回の実験結果より、提案手法1,2の両方とも好奇心なしと比べると良い結果が出ているため、好奇心としての役割は果たしている。しかし、従来手法と比べると提案手法1では総行動回数はわずかに減少している事がわかるが、学習時間は入力を増やしたため、その分伸びていると考えられる。また、提案手法2では従来手法よりも総行動回数が多く、学習時間が伸びている。そのため提案手法2のやり方はあまり効果がないと考える。

## 5. まとめ

本研究では、従来手法であるRNDに、どうやって来たかという一時刻前の行動の情報を加えることで、様々なパターンの組み合わせを試し、局所解を回避することで、効率的に学習させることを目的として行った。結果として、提案手法1ではわずかに減少していることが確認できた。しかし、提案手法2では総行動回数も、時間も従来手法より増えていた。そのため、本研究の手法である一時刻前の行動の情報を加え、入力情報を増やすという手法はそれほど効果がないと考えられる。

**【参考文献】**

[1] Yuri Burda, Harrison Edwards, Amos Storkey, Oleg Klimov., “Exploration by Random Network Distillation”, arXiv:1810.12894 [cs.LG], 30 Oct 2018, (2018)

[2]岩科亨, 森山甲一, 松井藤五郎, 武藤敦子, 大塚信博, マルチエージェント強化問題への好奇心探索の適用, 人工知能学会第二種研究会資料（2021）