ISSN 2186-5647

2-34

Shapingの概念を取り入れた段階的強化学習による自律エージェントの経路探索

日大生産工　○小沼　隆一　日大生産工　山内　ゆかり

1 まえがき

近年、ロボティクスの分野において生物の認知プロセスを模倣し、具体的なメカニズムとして具現することの必要性が唱えられている [1]。その一つとして強化学習(Reinforcement Learning)が挙げられる。その中でも、「Shaping」という動物の調教などで用いられている概念を学習に組み込んだ「Shaping強化学習」は代表的なQ学習に比べ有効なことが分かっている[2]。Shapingとは学習者が容易に実行できる行動からより複雑な行動へと段階的に学習させ、次第に希望の行動系列を形成する概念である。

そこで本研究では、強化学習の代表的なQ学習とShaping強化学習で様々な環境の経路探索問題で比較実験を行い、どのような環境においてShaping要素が強みを発揮するかを検証する。

2 Shaping強化学習による行動獲得

強化学習とは、ある環境内におけるエージェントが、現在の状態を観測し、取るべき行動を決定する問題を扱う機械学習の一種である。エージェントは行動を選択することで環境から報酬を得る。一連の行動を通じて報酬が最も多く得られるような方策を学習する。また、教師なし学習なので未知の学習領域を開拓していく行動と、既知の学習領域を利用していく行動とをバランス良く選択することができるという特徴も持っている。その性質から未知の環境下でのロボットの行動獲得に良く用いられる。代表的な手法のQ学習では、各状態において可能な行動の中で最も行動評価関数の高い行動をとるように学習する。そして、実行するルールに対しそのルールの有効性を示す Q値という値を持たせ、エージェントが行動するたびにその値を更新する。ここでいうルールとはある状態とその状態下においてエージェントが可能な行動とを対にしたものである[3][4]。

Shaping強化学習では段階的に学習するので、複雑な状態でも効率よく学習を行うことができる。しかし、Shaping強化学習では報酬を与えるタイミングや報酬量、与える回数など調教者が行わなければならないことが多い。また、同じ条件下の学習でも調教者により報酬の与えるタイミングの判断など調教者に依存する要素が多いことも課題として挙げられる。

3 提案手法及び実験概要

エージェントは迷路マップを探索しながら進むため、サブゴールを何度も通る可能性がある。そのため何度もMedium報酬を与えてしまうとサブゴール周辺で局所解に陥ってしまう。それを回避するために本手法においてMedium報酬は初めてサブゴールを通過した時にだけ与え、それ以降は通過したとしてもMedium報酬は与えないようにしている。また、適宜与えるのではなく迷路マップに対して予め設定することで調教者の負担を減らしている。Medium報酬はゴール位置に近いものほど多く与えゴール位置より離れるに従って少ない値を設定している。

以下に本研究で扱う迷路のマップを挙げる。

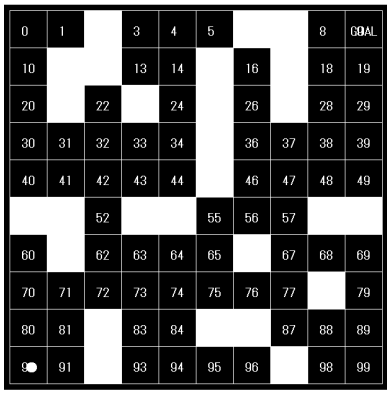

図1　迷路マップの例

図1は迷路マップの一例である。移動不可なマスは白で表記され、エージェントは●で表している。

本手法の条件は以下の通りである。

- 迷路マップの移動環境は10×10マスの100マスとし、障害物にはエージェントは進めないものとする。
- エージェントは上下左右の4つの行動を取ることができる。
- エージェントは迷路マップの外には出ることができない。
- エージェントの目標行動は最短経路でのゴール到達である

Pathfinding of Autonomous Agents by Shaping Reinforcement Learning
Ryuichi ONUMA and Yukari YAMAUCHI

- 行動回数には上限を設け、ゴールに至らなかった場合にはその試行は学習しないものとする。
- エージェントの行動選択にはQ値によるルーレット選択を用いる。

比較実験する手法は以下の4つである。
・Q学習(障害物無)
・サブゴール有Q学習(障害物無)
・サブゴール有Q学習(障害物有)
・サブゴール無Q学習(障害物有)

$$\Delta Q(s_t, a_t) = \alpha(r_t + M(t) + \gamma \max Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)) \quad (1)$$

α ：学習率　γ ：割引率　r_t：報酬
M(t)：Medium報酬

式1は実験で使用するQ値の更新式である。これは通常のQ学習の更新式にMedium報酬を加えたものである。

実験1では図1の迷路マップを使用する。この迷路マップでは最短経路が複数ある。実験2では最短経路が一通りの迷路マップを用いる。5回の平均値で比較を行う。

4　実験結果

図2、図3では縦軸は行動回数、横軸は試行回数を表している。図2の実験1では障害物有の迷路マップにおいてサブゴールの有無により最短経路の学習速度に差が出ている。障害物が無い迷路マップではサブゴールの有無による学習速度の差は出ていない。最短経路が一通りの実験2でもサブゴールの有無により学習速度に差が出ている。

5　まとめ

サブゴール地点でMedium報酬を与えた場合、どちらの実験でも同じ程度の試行回数で最短経路を学習している。Medium報酬が無い場合には、迷路マップが複雑になると最短経路を学習するまでに時間がかかっている。このため複雑な迷路マップほど最短経路の学習に強みを発揮すると考えられる。

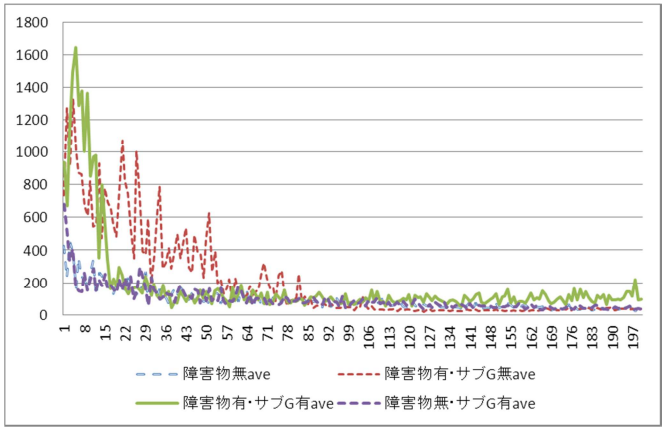

図2　実験1の結果

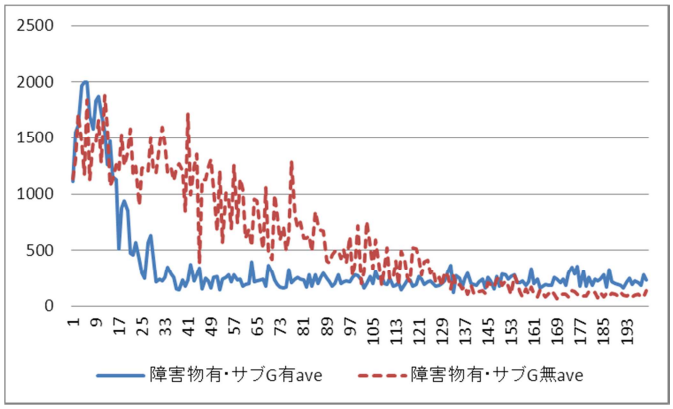

図3　実験2の結果

「参考文献」
[1] 浅田稔，　石黒浩，　國吉康夫，　“認知ロボティクスの目指すもの”，　日本ロボット学会誌，　Vol.17，　No.1，　pp.1-5(1999)
[2] 前田陽一郎，　花香敏，　“Shaping強化学習も用いた自律エージェントの行動獲得支援手法”，　日本知能情報ファジィ学会誌，　Vol.21，　No.05，　pp.722-733(2009)
[3]三上貞芳，　皆川雅章，　強化学習，　森北出版株式会社，　(2000)
[4]濱上知樹，　平田廣則，　学習とそのアルゴリズム，　森北出版株式会社，　第8章，　(2002)