ISSN 2186-5647

P-15

強化学習と情動学習を用いた迷路探索問題の最適化

日大生産工（学部）　○下條　涼　日大生産工　山内 ゆかり

1 まえがき

ロボット技術の進展によって、産業用ロボットに加え、人間の生活環境内で作業を行うロボットが精力的に開発されている。人間と同じ生活環境で共生するロボットには、自身の置かれた環境だけではなく、人間やほかのロボットとの関係を考慮した社会的な意思決定が求められる。

意思決定の代表的な学習方法に強化学習がある。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。行動することで報酬や罰が与えられる。エージェントは報酬が最も大きくなるような行動を学習していく。しかし、強化学習はエージェントが複数いる環境において、他のエージェントとの関係を考慮した学習が困難である。そこで本研究では、強化学習と共に情動学習を行うことで他のエージェントとの関係を考慮した学習を行えるようにする。情動とは、心理学の分野で、外部刺激や観念によって生じる快、不快の意識現象である感情のうち特に一過性の強い感情のことである。

情動学習を用いた研究として、三澤らによる強化学習と情動学習に基づく意思決定法(DRE)がある[1]。DREでは、行動決定時にε－グリーディー法を用いる。行動選択プロセスでは、ランダムに行動選択する場合と、強化学習による状態価値に基づいて行動選択する場合がある。行動決定プロセスでは、行動選択プロセスに従って行動する場合と、情動学習による情動価値に基づいて行動決定する場合がある。DREを用いたエージェントが社会的行動である協調行動を創発できることと環境変化に対する高い適応性が確認できた。しかし、この手法では環境が単純すぎることにより、強化学習の本来の目的である最適解の探索を確認することができなかった。

本研究では情動学習に基づいた協調行動と環境変化への対応をし、最適解の探索をすることを目的とする。具体的には環境を迷路化し、最適解という概念を作る。そして一つのエージェントが強化学習を行って最適解を探索したのち、DREで学習する。まず初めに最適解を探索することで、最適解を意識しつつ協調行動を創発できると考えられる。

2 従来研究

2.1 DRE

三澤らは、強化学習と情動学習を組み合わせて、エージェントが社会的行動を取るように情動による割り込み機能(DRE)を実現した。DREを用いたエージェントは強化学習モジュール、情動学習モジュール、行動決定モジュールの3つから構成される。図1にDREを用いたエージェントの概念図を記す。

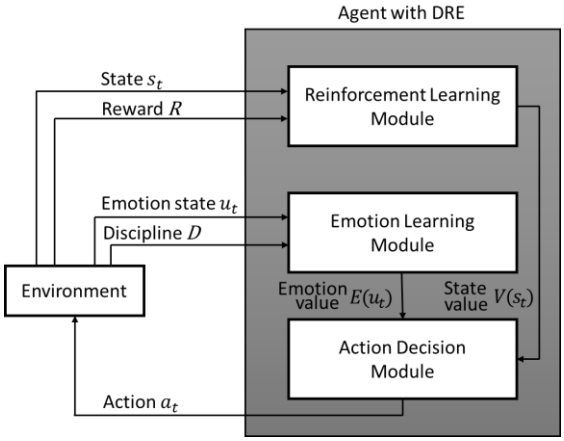

図1.DREを用いたエージェントの概念図

次に各モジュールの詳細について説明する。

2.2 強化学習モジュール

強化学習モジュールでは、タスクを達成するための行動選択に必要となる状態価値を学習する。状態価値の学習にはモンテカルロ法を用いる。

エージェントが行動制限回数以内に目標に到達することができた場合には報酬Rがエージェントに与えられ、状態価値$V(S_t)(t = 1,2,\dots,t')$が次式によって更新される。

$$V(s_t) \leftarrow V(s_t) + \alpha(r_t - V(s_t)) \quad (1)$$

$$r_t = \gamma^{t'-t} * R \quad (2)$$

Optimization in Shortest Path Problem
Based on Reinforcement Learning and Emotion Learning
Ryo SHIMOJO and Yukari YAMAUCHI

エージェントが目標に到達できなかった場合には、次式によって全状態の価値を減衰させる。

$$V(s_t) \leftarrow \eta * V(s_t) \qquad (3)$$

2.3 強化学習モジュール

情動学習モジュールでは、エージェントが経験した恐怖行動を情動価値として学習する。情動価値$E(u_t)$は、恐怖行動を経験した状況を表す情動状態u_tに対して定義される。各エージェントが存在する位置の組み合わせを情動状態とし、情動価値の学習にはモンテカルロ法を用いた。

エージェントが社会的行動を取らずに罰Dを受けた場合、情動価値$E(u_t)(t = 1,2,\dots,t')$が次式によって更新される。

$$E(u_t) \leftarrow E(u_t) + \alpha'(d_t - E(u_t)) \qquad (4)$$

$$d_t = \zeta^{t'-t} * D \qquad (5)$$

罰を受けたエピソード中に経験しなかった情動状態の価値は次式によって更新される。

$$E(u_t) \leftarrow \beta * E(u_t) \quad for \quad u \notin U_d \qquad (6)$$

2.4 行動決定モジュール

行動決定モジュールでは、強化学習と情動学習で学習した価値を基に、行動を決定する。行動決定モジュールは、行動選択プロセスと行動決定プロセスの2つのプロセスから構成される。両方のプロセスにおいてε－グリーディー法を用いる。

行動選択プロセスでは、ε_sの確率でランダムに、1−ε_sの確率で状態価値に基づいて行動を選択する。

行動決定プロセスでは、ε_eの確率で行動選択プロセスにおいて選択された行動、1−ε_eの確率で情動価値に基づいて行動を決定する。

3　提案手法

本研究では、従来研究の最適解を探索していないという問題を解決することを目指し、従来研究の環境を迷路化することと、従来手法にさらに強化学習を追加して学習を行う手法を提案する。

強化学習の学習手法には、迷路探索に優れているQ学習を用いる。

下記に提案手法のアルゴリズムを示す。

1. エージェントを一つ配置し、迷路探索を行い最適解の探索と状態価値の更新を行う。
2. エージェントを複数にし、ランダムに初期位置に配置し、行動を開始する。
3. 与えられた最適解と状態価値を基に強化学習と情動学習を行い、状態価値と情動価値を更新する。
4. ①全てのエージェントが全てのターゲットに到達した場合、②エージェント同士が衝突した場合、③行動制限回数に達した場合を終了条件とする。
5. 2～4を繰り返し実行する。

4　実験環境

本実験では、図2の15×15の2次元平面環境の迷路探索問題を用いて実験を行う。本論文では、2つのエージェントと2つの目標が存在する環境を取り上げる。

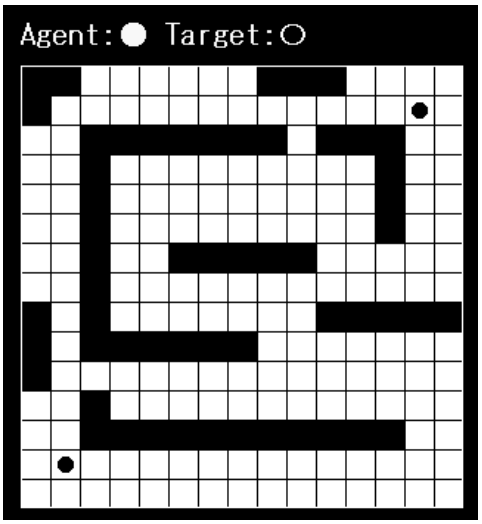

図2.実験環境

初期位置はエピソード毎に①障害物以外、②他のエージェントと異なる場所、③ゴール地点以外の場所からランダムで決定する。エージェントは1回の行動で上下左右いずれかのセルに1マス移動できる。尚、壁や障害物にぶつかる行動を選択した場合は、行動せずその場に留まる。ゴールとはターゲットの上下左右のセルに到達した場合を指す。各エージェントが同じセルに移動した場合を衝突とする。①全てのエージェントが全てのターゲットに到達した場合、②エージェント同士が衝突した場合、③行動制限回数に達した場合を終了条件とする。

各エージェントが目標に到達した場合に協調行動を達成できたとみなし、報酬Rを与える。また、衝突が起きた場合と行動制限回数に到達した場合に協調行動を達成できなかったとみなし、罰Dを与える。

行動制限回数を40とし、これを1エピソードとする。2000エピソードを1試行として、10回の試行を行う。100エピソード毎の衝突回数とゴール回数、ゴールに至るまでの平均ステップ

数の平均値を算出する。ゴール回数を大きくすることと、衝突回数と平均ステップ数を小さくすることを目標とする。

図3に従来手法を実装した協調行動回数の結果を示す。

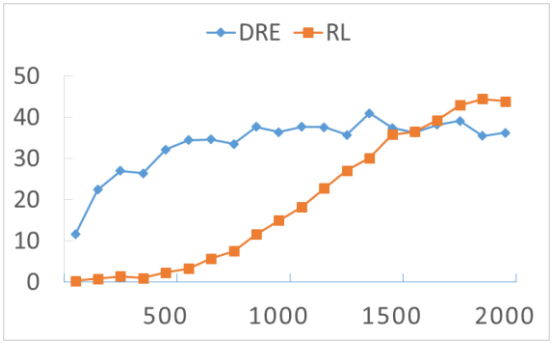

図3.協調行動回数の比較

比較対象として情動学習をしない単純な強化学習(RL)を用いた。RLの序盤は情動学習をしない為、エージェント同士の衝突が頻繁に起こり、協調行動が達成されない。しかし、状態価値の更新をすることで協調行動の達成回数は徐々に増えていく傾向があると考えられる。また、DREは強化学習と情動学習を行う為、RLに比べて学習の序盤に協調行動の達成がされている。

図4に従来手法を実装した衝突回数の結果を示す。

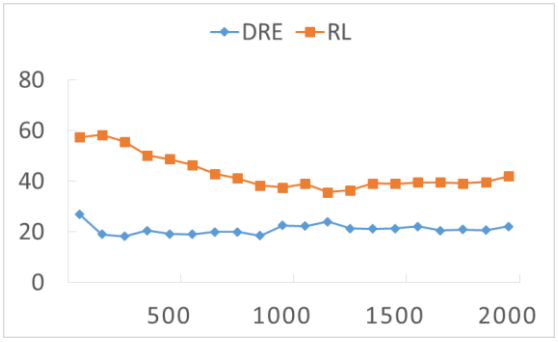

図4.衝突回数の比較

衝突回数については、どちらも減少していると言える。しかし、RL は情動学習を行っていないので、協調行動の達成に伴って減少しただけであると考えられる。

また、図 5 に、ある 1 試行における DRE エージェントの状態価値、図 6 にその時の情動価値を示す。これは 2000 エピソード後の状態価値と情動価値である。

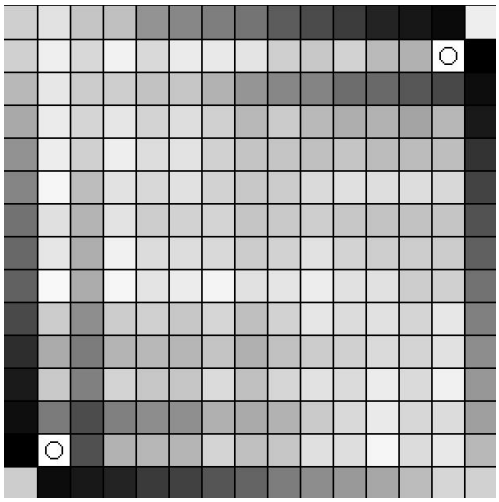

図 5.DRE エージェントの状態価値の一例

状態価値では目標が存在するセルの位置を”〇”で示しており、状態価値が高いものを濃く表示している。(a)から目標付近の状態価値が高いことがわかる。つまり、エージェントは情動価値に基づいて、目標達成のために行動していることがわかった。

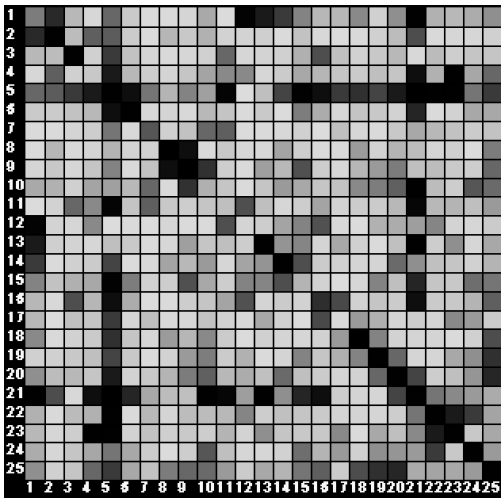

図 6.DRE エージェント情動価値の一例

情動価値が低いものを濃く表示している。5 行目、21 行目、5 列目、21 列目は目標が存在するエリアなのでエージェントが集まり、衝突が多くなっていることがわかる。また、各エージェントが同じエリアにいる場合は衝突が多くなるので、縦軸と横軸のエリア番号が一致している場合、情動価値は低くなる。つまり、エージェントは目標付近と他のエージェントが同じエリア内やその付近にいる場合に情動価値に基づいて恐怖行動を避けて行動していることがわかった。

5 まとめ

本研究では、複数エージェントによる探索問題で迷路環境に対して、従来手法の問題点であった最適解の探索をすることを目的とした学習方法を提案した。

正常に動作した場合、各エージェントは迷路環境において、最適経路の探索を行うと共に、他のエージェントとの関係を考慮した協調行動の創発ができると考えられる。

提案手法の今後の課題は、迷路探索に最適な強化学習の検討と、最適解の探索を優先した場合に起こるエージェント同士の衝突の回避プロセスを実装することである。また、DRE が従来手法の結果に比べて協調行動の達成回数が少ないという問題点があるので、従来手法の結果と同等の結果が得られるように実装をする。

「参考文献」

1) 三澤秀明，松田充史，堀尾恵一，「強化学習と情動学習に基づく意思決定法:利己的な判断による協調行動の創発」，知能と情報，Vol.24, No.1, pp.582-591 (2012)

2) 森山甲一，沼尾正行，「環境状況に応じて自己の報酬を操作する学習エージェントの構築」，人工知能学会論文誌, Vol.17, No.6, pp.676-683 (2002)

3) 強化学習入門：http://blog.brainpad.co.jp/entry/2017/02/24/121500（最終アクセス日時：2017 年 9 月 8 日 5 時 56 分）