ISSN 2186-5647

P-11

強化学習による経路学習と情動学習に基づく協調行動の獲得

日大生産工(学部)　○中山 雅隆　　日大生産工　山内 ゆかり

1 まえがき

現代では,ロボット技術の進展により産業用ロボットに加え,少子高齢化社会における労働力として人間の生活環境の作業をサポートするようなロボットの開発が精力的に行われており,様々なサービス分野での活躍が期待されている.人間と同じ生活環境で共生するロボットには,自身の置かれた状況だけでなく,人間や他のロボットとの影響や関係を考慮したうえでの社会的な意思決定が求められている.社会的な意思決定は情動の機能によっても発現されるという考えに基づくことでエージェントの意思決定に情動の概念を利用する.心理学の分野では,外部からの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中で特に一過性に生じる強い感情のことを情動という.従来では,三澤らによる情動学習を用いた研究に,強化学習と情動学習に基づく意思決定法を提案した.1)この方法を用いたエージェントは社会的行動である協調行動を創発できることや環境の変化に対して高い適応性があることが確認できた.しかし,この方法では強化学習の本来の目的である最適経路の探索が充分に出来ていないと考えられる.

本研究では強化学習を優先させ経路学習を行い,その後,情動学習に基づいた協調行動の獲得させる手法を提案し最適経路の探索を目指す方法を検討する.

2 研究背景および従来手法

三澤らは,強化学習と情動学習に基づく意思決定法を提案し,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を行う機能DREを実現した.

図１にDREの手順を示す.

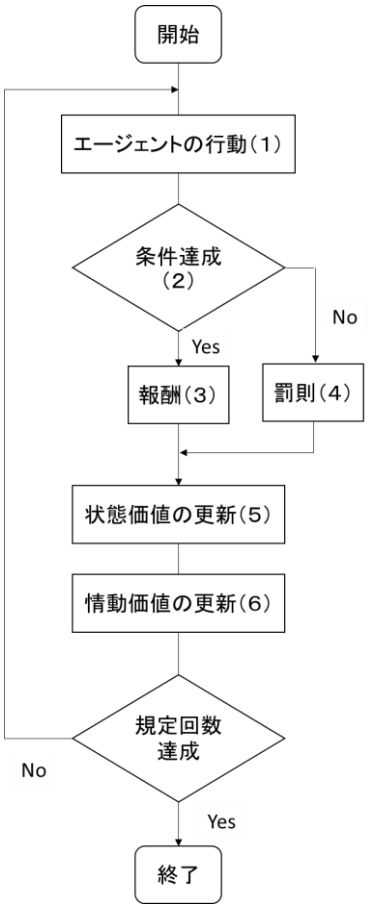

図１　DREのフローチャート

図1の(1)ではエピソードごとに各エージェントを条件の満たすセルにランダム配置を行う.その後,行動決定モジュールの行動選択プロセスによってエージェントはε_sの確率でランダムに行動を選択し,$1-\varepsilon_s$の確率で状態価値に基づいて行動を選択する.行動判断プロセスによってε_eの確率で行動選択プロセスによって選ばれた行動をそのまま行い,$1-\varepsilon_e$の確率で他のエージェントの周囲の情動価値に基づいて行動する.同じ値であった際は,その中からランダム選択される.(2)ではエージェントがゴールに到達もしくは,エージェント同士で衝突,制限回数の上限を超えたかを判定している.(3)では(2)でエージェントがゴールに到達していた場合,報酬Rが状態価値$V(s_t)(t=1,2,\ldots,t')$が下記の式(1),式(2)を用いて更新される.

Acquisition of Cooperative Behavior
Based on Reinforcement and Emotion Learning

Masataka NAKAYAMA　and　Yukari YAMAUCHI

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

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

エージェントが目標に到達できなかった場合には下記の式(3)によって全状態の価値が減衰される.

$$V(s_t) \leftarrow \eta \cdot (s_t) \quad (3)$$

式３のηは減衰率である.

(4)では(2)でエージェントが社会的行動を取らなかった場合,罰Dが与えられ情動価値$E(u_t)(t=1,2,\ldots,t')$は下記の式(4),式(5)を用いて更新される.

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

$$\mathrm{d_t} = \zeta^{\mathrm{t'-t}} \cdot \mathrm{D} \quad (5)$$

罰を受けたエピソード中に経験しなかった情動価値は,最新の情動経験を優先させるために下記の式(6)によって情動価値を変動させる.

$$\mathrm{E(u)} \leftarrow \beta \cdot \mathrm{E(u)} \quad \text{for} \quad \mathrm{u} \notin \mathrm{U_d} \quad (6)$$

罰を受けなかったエピソードでは,すべての情動価値はそのまま保持される.

その後,(5)で状態価値の更新,(6)で情動価値の更新が行われ,次のエピソードでは,それぞれの価値が更新された状態から(1)に戻り,同じことが行われる.

３ 提案手法

従来手法であるDREでは,状態価値と情動価値の更新を同時に行うことで協調行動のエージェント同士が衝突を避ける動きに影響を与えてしまい,その結果,最適経路を求める行動よりも衝突を避ける動きを優先としてしまうことがあるため最短経路の探索が充分に行えていないと考えられる.

そのため本研究では,最短経路を効率的に求めていくためにまず強化学習によって状態価値を求め,その後,強化学習と情動学習による協調行動を行わせながら経路学習させることで最適な経路を求めていく手法を提案する.

４ 実験結果および検討

従来手法であるDREと提案手法として挙げた方法を用いることで従来手法の実験環境である15マス×15マスのセルでのシミュレーションでの比較のほかに,従来手法でのシミュレーション環境に障害物を加えることでより精密な経路学習として状態価値及び情動価値の検証が行えると考え,検討を行っていく予定である.

５ まとめ

従来手法では状態価値と情動価値を同時に行っていくため,提案手法として強化学習を優先的に行うことで効率的に最適経路を求められるようにする.それにより従来手法よりも早い段階で最適経路を求めることが考えられるが,最短経路の探索を優先していく上で協調行動が減り,結果としてエージェント同士の衝突が起こりやすくなると考えられ,ゴールまで到達できずに報酬が与えられないパターンの増加する可能性が高まることが予想される.

【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:利己的な判断による協調行動の創発,知識と情報（日本知能情報ファジィ学会誌）vol.24,No.1,pp.582-591(2012)