多目的強化学習のマルチエージェントシステムへの適用

○西田公己　山田和明（東洋大学）

概要　本研究は，マルチエージェントシステムに発生する競合状態を回避するために，各エージェントの意思決定機構に多目的強化学習を導入するアプローチを提案する．従来の強化学習は，報酬をスカラーで与える必要があったが，多目的強化学習は，報酬をベクトルで与えられるため報酬系を設計し易いという利点がある．本稿では，狭い通路を２台のエージェントが通過する狭路問題に提案手法を適用することで，提案手法の有効性を検証する．

キーワード: マルチエージェントシステム，多目的強化学習，競合回避，狭路すれ違い問題

1　はじめに

マルチエージェントシステム（Multi-Agent System: MAS）[1]は，多数の自律エージェントから構成されている．MASは中央集権的な管理機構を持たず，各エージェントが環境や近傍のエージェントとの相互作用を通してシステム全体の秩序を形成するという特徴を持つ．そのため，MASはシステム内外の環境変化に対して頑健であるとされている．しかし，多数のエージェントが相互作用するため，エージェント間に複雑なダイナミクスが発生する．そのため，設計者が予めエージェントが遭遇するすべての状況を想定して，適切な行動をエージェントに組込むことは極めて困難である．この課題に対し，各エージェントの意思決定機構として強化学習[2]を用い，協調行動や競合回避行動を学習させるマルチエージェント強化学習（Multi-Agent Reinforcement Learning: MARL）[3),4),5)]が注目されている．

強化学習は機械学習の一手法であり，学習エージェントは観測した状態に対して行動を実行し，その評価として環境から与えられる報酬に基いて，目的を達成するために必要な状態-行動間の関係を試行錯誤的に学習する．強化学習をシングルエージェント問題に適用する場合，エージェントは与えられた目的を達成したときにスカラ型の報酬を与えることで，目的達成に必要な行動規則を学習することができる．一方，マルチエージェント問題に適用する場合，各エージェントは自身の目的とシステム全体の目的の両方を達成することが求められる．しかし，多くの場合，各エージェントの目的とシステム全体の目的は矛盾する．例えば，Fig. 1のような狭路すれ違い問題において，指示された目的地に到達するという個々のエージェントの目的と，両方のエージェントが目的地に到達するというシステム全体の目的は矛盾する．

仮に，従来の強化学習をそのままマルチエージェント問題に適用したとすると，設計者は，両方のエージェントが目的を達成したときのみスカラ型の報酬を与え，各エージェントが上記の矛盾を解消する行動規則を学習するよう問題を設定することができる．しかし，MARLでは，学習初期，2台のエージェントはランダムに環境中を探索しながら行動規則を学習する．そのため，両方のエージェントが，偶然，目的地に到達して報酬を得る確率は極めて低く，学習の収束には膨大な時間を要することが予想される．

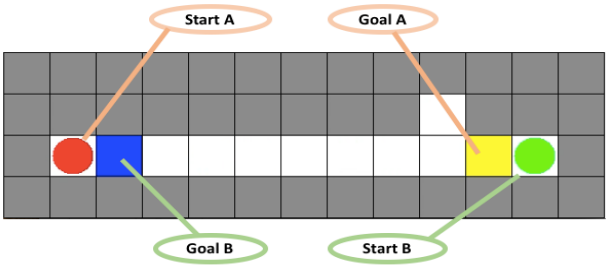

Fig.1 A narrow path problem.

一方，従来の強化学習に対し，複数の目的を設定して学習させることができる多目的強化学習（Multi-Objective Reinforcement Learning: MORL）[6),7)]が提案されている．多目的強化学習は，従来の強化学習が目的を達成したときにスカラ型の報酬を与えるのに対し，目的ごとに異なる報酬を設定し，ベクトル型の報酬として与えることができる．目的ごとに細かく報酬を設計することができるため，従来の強化学習をそのままマルチエージェント問題に適用するより，学習の収束性や速度の向上が期待できる．そこで，本研究では，MASに発生する競合状態を解消するために，各自律エージェントの意思決定機構としてMORLを導入する．そして，各自律エージェントが，近傍のエージェントとの相互作用を通して，競合回避行動を獲得できることを計算機実験により検証する．

次章では，多目的強化学習をMASに適用する利点について述べ，3章では使用する多目的強化学習について詳述する．4章では，提案手法を狭路すれ違い問題に適用し，計算機実験を通してその有効性を検証する．5章において，本稿のまとめと今後の課題について述べる．

2　多目的強化学習をMASに適用する利点

強化学習は，環境から得られる報酬を最大化する行動規則を学習する．そのため，強化学習をマルチエージェントシステムに適用する場合，個々のエージェントの最適化とシステム全体の最適化が必ずしも一致しない，という問題が発生する．例えば，協調問題の一つである追跡問題の場合，複数のエージェントが協力して獲物を包囲し，捕獲する必要がある．しかし，獲物を捕獲したエージェントのみに報酬を与えた場合，

第15回社会システム部会研究会，第59回システム工学部会研究会（2018年3月14日-16日・沖縄）

獲物を追い込むエージェントの学習が進まないという問題が発生する．一方，競合回避問題の一つである狭路すれ違い問題の場合，1章で述べた通り，各エージェントの目的とシステム全体の目的が一致しないという問題が発生する．

従来研究では，この問題を解決するために，以下のアプローチが提案されている．追跡問題において，保知ら[3]は，エージェント群が獲物を捕獲したとき，システム全体に与えられる報酬を各エージェントの貢献度に併せて分配する方法を提案している．また，張ら[4]は，獲物を捕獲したとき，各エージェントに与えられる報酬とシステム全体に与えられる報酬を，各エージェントの貢献度にあわせて分配する方法を提案している．一方，狭路すれ違い問題において，市川ら[5]は，エージェント間の学習進度が揃うよう各エージェントの学習パラメータ（学習率と割引率）を調整することで，競合回避行動を獲得している．この手法では，個々のエージェントが自らの目的を達成したときのみ報酬を与える．すなわち，システム全体の目的（2台のエージェントがすれ違う行動）を達成しても報酬を与えない．しかし，各エージェントの学習進度をエージェント間で共有し，どちらか一方の学習が収束し，他方の目的達成を阻害するとき，学習パラメータを調整することで，システム全体の目的を達成する確率を高めている．

このように，従来の強化学習をマルチエージェントシステムに適用する場合，個々のエージェントの最適化とシステム全体の最適化のバランスを取る仕組みが必要となる．その理由は，従来の強化学習が，個々のエージェントの報酬r_l，あるいは，システム全体の報酬r_g，または，その両方の報酬を一つの価値関数で管理することに起因する．例えば，エージェントがr_lまたはr_gのどちらか一方を受け取った場合，エージェントは一つの価値関数で報酬を管理しているため，与えられた報酬が個々のエージェント，あるいは，システム全体の獲得報酬を最大化する報酬なのか判断できない．また，両方の報酬を受け取った場合，報酬同士が打ち消し合う恐れがある．

そこで本研究では，上岡ら[6,7]が提案している多目的強化学習の一種であるMax-Min Actor-Critic (MMAC)をマルチエージェントシステムに適用する方法を提案する．MMACは，環境から与えられる報酬の種類だけ複数の状態価値関数を持つ．そして，拡張Max-Min法により最小の状態価値関数を選択し，そのTD誤差を用いて政策$\pi(s,a)$の行動価値関数を更新する．例えば，各エージェントが自らの目的を達成したときに正の報酬を与え，システム全体の目的を達成できなかったときに負の報酬を与えると，MMACはシステム全体の獲得報酬を下げる状態価値関数のTD誤差を用いて行動価値関数を更新する．そのため，システム全体の獲得報酬を下げる行動規則を抑制するよう学習することが期待される．また，MMACの学習則は従来のActor-Criticとほぼ同じであるため，従来研究の成果を適用することが容易であり，従来手法を実装することで学習性能の向上が期待できる．

3　アルゴリズム

本研究では，マルチエージェントシステムを構成する各エージェントの意思決定機構として，多目的強化学習の一つであるMax-Min Actor-Critic (MMAC)を採用する．MMACは，Actor-Criticを複数の目的が扱えるように拡張した学習アルゴリズムである．下記にActor-CriticとMMACの概要を説明する．

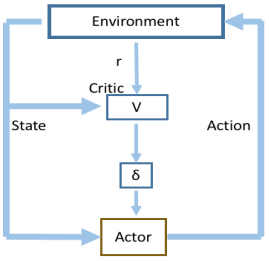

Fig.2 Actor-Critic

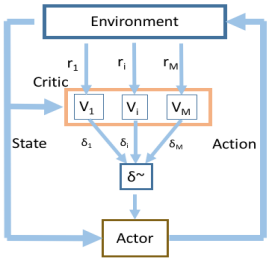

Fig.3 Max-Min Actor-Critic

3.1　Actor-Critic

Actor-Criticは，Fig. 2に示すように行動器（Actor）と評価器（Critic）から構成されている．行動器は，エージェントが観測した状態s_tにおいて行動a_tを実行する方策$\pi(s_t,a_t)$の行動価値関数$\theta(s_t,a_t)$を基に実行する行動を確率的に選択する．評価器は，環境から与えられる報酬を基に状態s_tにおける状態価値価関数$V_t(s_t)$を推定する．そして，状態価値のTD誤差によって，実行した方策の行動価値を更新する．評価器と行動器の学習は次のように行われる．

評価器は，時刻tにおける状態s_tの状態価値関数$V_t(s_t)$と環境から与えられる報酬rを基にTD誤差δ_tを次式により求める．

$$\delta_t = r_{t+1} + \gamma V_t(s_{t+1}) - V_t(s_t) \tag{1}$$

評価器の状態価値関数$V_t(s_t)$は，学習率を$\alpha_c(0 < \alpha_c < 1)$とすると，次式により更新される．

$$V_{t+1}(s_t) = V_t(s_t) + \alpha_c\delta_t \tag{2}$$

行動器の行動価値関数$\theta_t(s_t,a_t)$は，状態s_tで方策$\pi(s_t,a_t)$に従い行動a_tを選択し，次状態s_{t+1}に遷移したとき，学習率を$\alpha_a(0 < \alpha_a < 1)$とすると，次式により更新される．

$$\theta_{t+1}(s_t,a_t) = \theta_t(s_t,a_t) + \alpha_a\delta_t \tag{3}$$

3.2　Max-Min Actor-Critic (MMAC)

MMAC は, 複数の種類の報酬を受けることによって, 多目的な学習ができる．Fig. 3 に示すように MMAC の評価器は，環境から与えられる複数の種類の報酬$\boldsymbol{r} = (r_1, \ldots r_M)^T$の数だけ，状態価値関数$\boldsymbol{V} = (v_1, \ldots, v_M)^T$を持つ. MMAC は, 報酬$\boldsymbol{r} = (r_1, \ldots r_M)^T$が与えられると，評価器 i における状態価値の TD 誤差を次式により求める．

$$\delta^i = r_t^i + \gamma_i V_t^i(s_{t+1}) - V_t^i(s_t), \quad i = 1, \ldots, M \tag{4}$$

評価器の各状態価値関数は，学習率をα_cとすると，次式により更新される．

$$V_{t+1}^i(s_t) = V_t^i(s_t) + \alpha_c\delta^i \tag{5}$$

行動器では，まず，拡張Max-Min法を用いてベクトルのTD誤差を下記の(6)式によりスカラー化した$\tilde{\delta}$を

求める．そして，学習率をα_aとすると，行動価値を(7)式により更新する．

$$\tilde{\delta} = \delta^k,$$
$$k = \arg\min_i \{V^i + \zeta\delta^i\} \quad (6)$$
$$= \arg\min_i \left\{(1-\zeta)V^i(s_t) + \zeta\left(r^i + \gamma_i V^i(s_{t+1})\right)\right\}$$

$$\theta_{t+1}(s_t, a_t) = \theta_t(s_t, a_t) + \alpha_a\tilde{\delta} \quad (7)$$

ただし，ζは$[0,1]$の正の定数である．なお，$V^i + \zeta\delta^i$を最小にするiが複数存在する場合はその中から等確率でランダムに選択する．

4　計算機実験

4.1　実験設定

本実験では，提案手法の有効性を検証するために，提案手法を Fig. 1 に示す狭路すれ違い問題に適用する．各エージェントの目的は，スタートからゴールに到達することである．この問題では，両方のエージェントは最短 16 ステップでゴールに到達することができる．しかし，各エージェントが利己的に行動すると互いに目的が達成できないよう設定されている．

実験では，エージェントが 500 ステップ行動するか，両方のエージェントがゴールしたとき，エピソードを更新する．そして，1000 エピソード 100 トライアル行い，提案手法の有効性を検証する．従来研究 [5] では，報酬として，エージェントがゴールしたとき正の報酬 r_{goal} を与え，それ以外のとき負の報酬 r_{nom} を与える．本実験では，さらに 500 ステップ行動しても両方のエージェントがゴールできなかったとき負の報酬 r_{nogoal} を与える．各エピソードにおいて，ゴールに到達したエージェントは，エピソードが更新されるまでゴールに留まり，行動選択や状態価値や行動価値の更新はしない．ただし，両方のエージェントが 500 ステップ以内にゴールできなかった場合，エピソード更新時に双方のエージェントに r_{nogoal} を与える．

エージェントは，状態入力として自分と相手のエージェントが存在する座標を完全知覚する．また，行動として上下左右と停止のいずれかを選択して実行する．実行する行動方策は Soft-max 法 [3]により選択し，逆温度パラメータを 0.1 とする．行動器と評価器の学習率をそれぞれ$\alpha_a = 0.2$，$\alpha_c = 0.2$とし，割引率を$\gamma = 0.9$とする．また，状態価値Vと行動価値θの初期値をそれぞれ$V_0 = 5.0$，$\theta_0 = 5.0$とする．

4.2　実験 1：MMAC の有効性の検証

実験 1 では，MMAC の有効性を検証するために，次の 3 種類の報酬の与え方をした場合の競合回避行動の獲得率を比較する．

① Actor-Critic（2 種類の報酬）
② Actor-Critic（3 種類の報酬）
③ MMAC

Fig.4 Learning history

Table.1 The acquiring rate of conflict avoidance behaviors

	Success rate	The rate of learning of shortest pass
Actor-Critic (two rewards)	5%	1%
Actor-Critic (three rewards)	100%	75%
Max-Min Actor-Critic	100%	100%

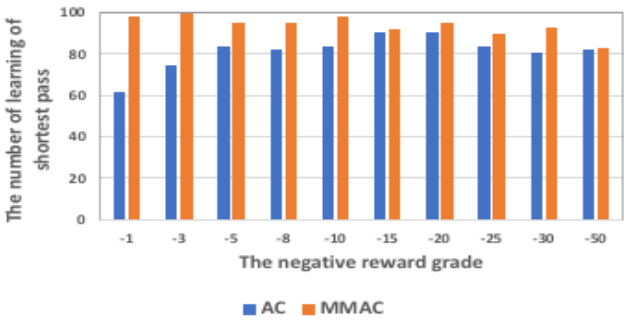

Fig.5 The number of learning of shortest pass at each reward

なお，設定②と③の違いは，②の Actor-Critic が一つの状態価値関数で 3 種類の報酬を管理するのに対し，③の MMAC は 3 つの状態価値関数で 3 種類の報酬を別々に管理する点である．

■ 実験設定

報酬として，Actor-Critic（2 種類の報酬）では$r_{goal} = 5.0$，$r_{nom} = -0.01$，Actor-Critic（3 種類の報酬）では$r_{goal} = 5.0$，$r_{nom} = -0.01$，$r_{nogoal} = -3.0$，MMAC では$r_{goal} = 5.0$，$r_{nom} = -0.01$，$r_{nogoal} = -3.0$をベクトルで与える．なお，$r_{goal} = 5.0$と$r_{nom} = -0.01$は，市川ら [1]のパラメータ設定を踏襲している．ただし，拡張 Max-Min 法のパラメータζを 0.1 とする．

■ 実験結果

2 台のエージェントが同時にゴールするまでに要したステップ数の遷移を Fig. 4 に示す．Fig. 4 から Actor-Critic（3 種類の報酬）と MMAC では学習が進むとステップ数が減少し，安定していることが分かる．そのため，競合回避行動を学習できたことがわかる．次に，競合回避行動の学習確率の表を Table. 1 に示す．Table. 1 から Actor-Critic（3 種類の報酬）と MMAC では競合

回避行動の学習率が 100%となっている．最短経路 16 ステップでゴールすることができた成功率は，Actor-Critic（3 種類の報酬）が 75%に対し，MMAC は 100%であった．

4.3 実験 2：報酬の大きさによる学習への影響

強化学習では，複数の種類の報酬を与えた場合，報酬の組合せによっては，相反する場合や報酬を打ち消し合う可能性がある．そこで，本実験では，両方のエージェントがゴールしない場合に与える報酬 r_{nogoal} の大きさを変更した場合，r_{nogoal} が学習結果に与える影響を検証する．

■ 実験設定

実験 2 では，学習の成功率が高かった Actor-Critic（3 種類の報酬）と MMAC において，両方のエージェントがゴールできなかった場合に与える報酬r_{nogoal}が-1.0, -3.0, -5.0, -8.0, -10.0, -15.0, -20.0, -25.0, -30.0, -50.0 のときの学習結果を検証する．ただし，拡張 Max-Min 法のパラメータζを 0.1 とする．

■ 実験結果

r_{nogoal}の値を変更したときの最短経路学習回数を Fig. 5 に示す．MMAC は r_{nogoal} = −1.0〜10.0 間のとき最短経路の学習成功回数が多いことから，設計者が適切な報酬の大きさを試行錯誤的に設定しなければならないという負担を軽減できるものと期待できる. 次に，Actor-Critic（3 種類の報酬）と MMAC を比較した場合，MMAC の方が, 最短経路の学習成功回数が多いことがわかる. MMAC が最短経路の学習成功回数が多い理由として，MMAC は拡張 Max-Min 法により最小の状態価値を高めるように学習する．そのため，MMAC は，両方のエージェントがゴールに到達できないときに与えられる負の報酬 r_{nogoal} を受ける行動，すなわち，他者のゴール到達を妨害する行動の行動価値を下げるように学習する．その結果，両方のエージェントがゴールする行動が獲得され，Actor-Critic（3 種類の報酬）より最短経路の学習成功回数が多くなったと考えられる．

5 おわりに

本研究では，マルチエージェント強化学習（Multi-Agent Reinforcement Learning: MARL）を実現するために，各エージェントの意思決定機構として多目的強化学習の一手法である Max-Min Actor-Critic (MMAC）を搭載し，個々のエージェントのパフォーマンスとシステム全体のパフォーマンスを評価する複数の報酬を与えることで，競合回避行動を獲得できることを確認した．また，2 台のエージェントが狭路ですれ違う問題において Actor-Critic と MMAC を比較した結果，最少ステップ数でタスクを達成する割合が Actor-Critic より MMAC の方が高いことを確認した．

今後の課題として，多数のエージェントからなるマルチエージェント環境において，多目的強化学習により協調行動や競合回避行動が獲得できるか検証する予定である．

謝辞

本研究は，東洋大学井上円了記念研究助成を受けたものです．

参考文献

1) 高玉圭樹：マルチエージェント学習 − 相互作用の謎に迫る −，コロナ社，(2003)
2) R. S. Sutton and A. G. Barto, Reinforcement Learning: An introduction, A Bradford Book, (1998)
3) 保知良暢, 松井藤五郎, 犬塚信博, 世木博久：マルチエージェント強化学習における報酬発生条件に基づく貢献度判別と報酬分配, 人工知能学会全国大会論文集, Vol.16, 2D3-02, (2002)
4) 張坤, 前田陽一郎, 高橋泰岳：貢献度評価に基づくマルチエージェント強化学習の報酬分配, 日本知能情報ファジィ学会ファジィシステムシンポジウム講演論文集, Vol.26, 246/251, (2010)
5) 市川嘉裕，高玉圭樹：学習進度に基づくマルチエージェント Q 学習における競合回避，計測自動制御学会論文集，Vol.48, No.11, 764/772, (2012)
6) 上岡拓未, 内部英治, 銅谷賢治：複数の価値関数を用いた多目的強化学習, 電子情報通信学会技術研究報告：信学技報, Vol.105, No.658, 127/132, (2006)
7) 上岡拓未, 内部英治, 銅谷賢治：Max-Min Actor-Critic による複数報酬課題の強化学習, 電子情報通信学会論文誌. D, 情報・システム, Vol.90, No.9, 2510/2521, (2007)