エージェント間の報酬の受渡しを考慮した強化学習による役割の生成(2)

中原雅之　長名優子

東京工科大学大学院 バイオ・情報メディア研究科 コンピュータサイエンス専攻

1 はじめに

マルチエージェントシステムは複数のエージェントにより問題を解決するシステムであり，エージェント間の協調により，より効率的にタスクを達成できるという特徴を持っている．マルチエージェントシステムで解決できる問題の中にエージェントが役割分担をすることにより効率をあげることのできる問題が存在する．タスク全体を1つのエージェントが行うことが可能であれば役割分担を行うことの意義は小さいが，複数のサブタスクを経てはじめてタスク全体が達成される場合では，各エージェントがそれらのサブタスクの一部のみを担当することでより効率よく全体のタスクを行うことができると考えられる．

役割分担をエージェントの学習によって実現するためには，エージェント自身が学習することによって役割を得ることが望ましく，各エージェントの正しい行動は未知であるため強化学習のような学習が適している．文献[1]では，エージェント間で報酬の受け渡しを行い，タスク本来の目的以外の行動(サブタスク)を学習することを可能としている．しかし，設計者があらかじめ決定したタイミングで報酬の受け渡しを行っているため汎用性が低いという問題がある．

本研究では役割分担を実現するために，エージェント間での報酬の受け渡しを考慮した強化学習を提案する．エージェント間で報酬の受け渡しを行うことでタスク本来の目的以外の行動を学習し，役割の学習を可能とする．また，報酬の受け渡すタイミングの判断と受け渡す報酬の量をエージェント自身に決定させることで汎用性を高める．

2 エージェント間での報酬の受渡しを考慮した強化学習

本研究では，マルチエージェント環境における役割分担の学習の方法としてエージェント間での報酬の受け渡しを考慮した強化学習を提案する．

強化学習法には，非MDPの環境で有利性が示されているProfit Sharingを用いる．提案手法では，履歴として状態の変化が観測されたときのエージェントの行動とその時刻，またその時刻よりも前で最後にいずれかのエージェントが報酬を受け取った時刻の情報を保持しておき学習に利用する．以下に提案手法の流れを示す．

(1) 環境と履歴リストの初期化を行う．

(2) 各エージェントは状態 s_t を観測し，その状態に応じて行動を決定し実行する．行動は，ルーレット選択により確率的に決定される．

(3) 各エージェントが行動を実行することにより状態が s_t から s_{t+1} に遷移する．このとき，エージェントの行動に伴う状態の変化以外の環境の変化が起きたかどうかを観測する．そして，環境の変化が観測された場合には，履歴リストを更新する．

(4) 環境や他のエージェントから報酬を取得したエージェントは，他のエージェントに報酬の一部を渡すかどうかの判断を履歴リストを用いて行う．環境の変化が観測された時刻においてすべてのエージェントがとった行動のうち，過去に環境の変化が観測されたタイミングにおいても同じ行動がとられている回数が多いほど，その行動が環境の変化に影響を及ぼしている可能性が高いため，そのような行動をとったエージェントを報酬を渡す対象として選択する．

(5) 行動価値の更新を行う．報酬を取得した時刻より過去に遡り報酬を分配し行動価値を更新していく．行動価値 $Q^X(s_\tau, a_\tau^X)$ を以下のように更新する．

$$Q^X(s_\tau, a_\tau^X) \leftarrow Q^X(s_\tau, a_\tau^X) + r^X(\tau) \qquad (1)$$

報酬関数 $r^X(\tau)$ は

$$r^X(\tau) = \begin{cases} r_{t_r^X}^{\rightarrow X} - r_{t_r^X}^{X\rightarrow} & (\tau = t_r^X) \\ \dfrac{1}{M} r_{\tau+1}^X & (\tau \neq t_r^X) \end{cases} \qquad (2)$$

Generation of Role by Reinforcement Learning Considering Delivery of Reward among Agents (2)
Masayuki Nakahara and Yuko Osana (Tokyo University of Technology, nakahara@osn.cs.teu.ac.jp, osana@cc.teu.ac.jp)

で与えられる．ここで，$r_{t_r^X}^{\rightarrow X}$ はエージェント X が報酬を取得した時刻 t_r^X においてエージェント X が取得した報酬の量，$r_{t_r^X}^{X\rightarrow}$ はエージェント X が時刻 t_r^X に取得した報酬のうち他エージェントに渡す報酬の量，M はエージェントのとることのできる行動の数を表している．

3 計算機実験

提案手法の有効性を検証するために経路探索問題のタスクを用いて計算機実験を行った．

図1のような5×5のマップにA，B2体のエージェントを配置し，ゴールGを目指す経路探索問題を考える．ここでは，ゴール地点に到達するには，ゴール地点とは別の位置にあるスイッチSを押してからゴールに向かう必要がある．また図1において×はスイッチを押すことで通過することが可能になる障害物を示している．

3.1 報酬の受渡しの有効性

ここでは提案手法と報酬の受渡しを行わない手法の比較を行った．報酬の受渡しを行わない場合はゴールしたエージェントにのみ報酬を与えるものとした．提案手法と報酬の受渡しを行わない手法のゴールに到達するまでのステップ数の変化を図2に示す．実験結果より，提案手法により報酬の受渡しを行うことで効率の良い学習が行われていることがわかる．また，提案手法ではスイッチを押すことをゴールしたエージェントから他のエージェントに渡される報酬により学習できることが確認できた．また，提案手法において充分に学習が行われた後の行動を見ると，ゴール前の分岐点に最初に到達したエージェントがスイッチを押し，後から分岐点に到達したエージェントがゴールに向かうといった役割分担が見られることがわかった．

3.2 報酬を受け渡す判断の有効性

ここでは提案手法と報酬を受け渡す条件をあらかじめ設定した場合との比較を行った．エージェントはゴール地点に到達することで環境から報酬を取得し，ゴール前にある障害物が通過可能となったときに環境の変化が起きたことを観測できるものとする．報酬を受け渡す条件を設定した場合では，ゴールしたエージェントはスイッチを押したエージェントに対して取得した報酬の半分の報酬を渡すように設定した．それぞれのゴールに到達するまでのステップ数の変化を図3に示す．

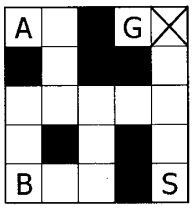

図 1: 経路探索問題 (エージェント 2 体)

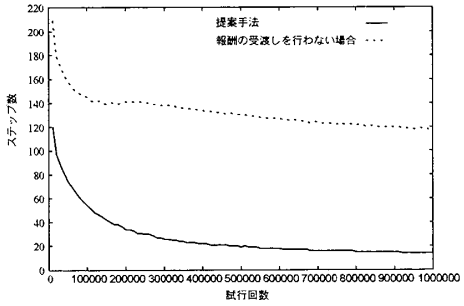

図 2: 学習によるステップ数の変化 (1)

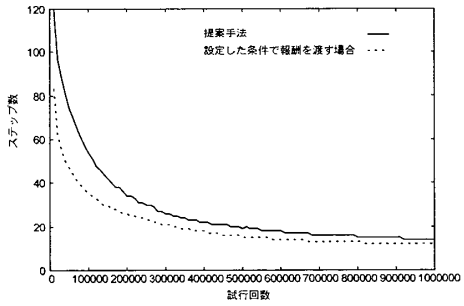

図 3: 学習によるステップ数の変化 (2)

実験結果をみると，提案手法で学習を行った場合と報酬を受け渡す条件と報酬量をあらかじめ設定した場合では，学習の初期ではゴール状態に到達するまでのステップ数に差があるが，試行回数が増えるごとに差が縮まっていることがわかる．また，試行回数1000000回の時点では，ほぼ同程度のステップ数でゴール状態に到達できることがわかる．実験結果より提案手法において設計者があらかじめ報酬を渡す条件と報酬量を決定した場合とほぼ同程度の性能で学習ができていることがわかる．

参考文献

[1] M. Saitoh and Y. Oyama: "Economy-like reward distribution for division of labor," Proceedings of the 24th IASTED International Multi-Conference on Artificial Intelligence and Applications, Innsbruck, pp.499–506, 2006.