マルチエージェント強化学習における ソーシャルインタラクション情報の視覚化

Visualization of Social Interaction Information
in Multi-Agent Reinforcement Learning

◯ 張　坤　　前田　陽一郎　　高橋　泰岳
◯ Kun Zhang　Yoichiro Maeda　Yasutake Takahashi
福井大学大学院 工学研究科
Graduate School of Engineering, University of Fukui

Abstract:
In the multi-agent reinforcement learning, how to obtain an appropriate cooperative behavior from the mutual interaction is extremely important. In this research, we propose an interactive reinforcement learning system with the efficient cooperative ability through the social interaction among agents. On the other hand, we visualize the social interaction information of multi-agents in different conditions. Original position, visual range, moving speed effect the mutual interaction among agents, and can be analyzed through the visualization of interactive information.

1　緒言

複数のエージェントに協調動作を学習させるマルチエージェント強化学習では、エピソードや学習方策を共有することで、協調行動を効率的に学ぶことができる [1]。しかし、異質エージェントは一般に性質や能力が異なるため、相手に有効な方策を自身にそのまま適応できない場合が多い。そのため、相手の学習方策がどのような状況で、どの程度利用できるかなどを自律的に学習することが必要となる。

そこで、他エージェントの有効な学習経験を間接的に活かせれば、マルチエージェント全体の協調性が高まるものと考えられる。本研究では、マルチエージェント間の信頼度に基づくインタラクティブ強化学習が可能なシステムの構築を提案する [2]。

しかしながら、どのような要素が異質エージェント間のインタラクションに影響を与えるかを分析するのは困難であるため、インタラクションに影響を与える初期位置、視野範囲や移動速度などの条件を変えて、インタラクション情報を視覚化することにより、ソーシャルエージェントがどのようにグループ関係を構築するかを検証する。

2　信頼度を用いた強化学習システム

本研究では、各エージェントは自身の強化値のみではなく、知覚範囲内のエージェントの強化値 (例えば Q-learning の場合の Q 値) を利用することができ、これに基づいて、目標となる適切な行動を選択する。

信頼度を用いたインタラクティブ強化学習では、各エージェントが目標達成行動に携わりながら、それぞれのエージェントと一緒に目標を達成した時、環境から得た報酬により、相手との信頼度を構築する。報酬は高いほど信頼度が高まる。自身に適用できる強化値のみを取り入れるため、いろいろな強化値から自身に適用できる強化値のみを選択することが可能になる。

2.1　行動選択アルゴリズム

ここでは、試行錯誤でエージェント間で各グループの各個体への信頼度を生成し、更新する。その信頼度に基づいた行動選択戦略の獲得の流れを図 1 に示す。エージェント間の強化値の相互利用に基づいて、自己の利益のみで行動するだけではなく、他のエージェントとの共同利益も考慮して行動し、インタラクション機能によるマルチエージェントの協調行動を向上させることを目標とする。

学習前にはすべてのエージェント間には信頼度が存在しない。目標を達成した報酬を得たとき、携わったエージェントの所属グループの間にはグループ信頼度が生成される。そのグループ信頼度が同様にグループ内の各エージェント間の信頼度に転用される。さらに、途中で他のエージェントの強化値を利用した回数により、個体の信頼度を生成する。グループ信頼度と個体信頼度を構築することで、他エージェントの強化値は

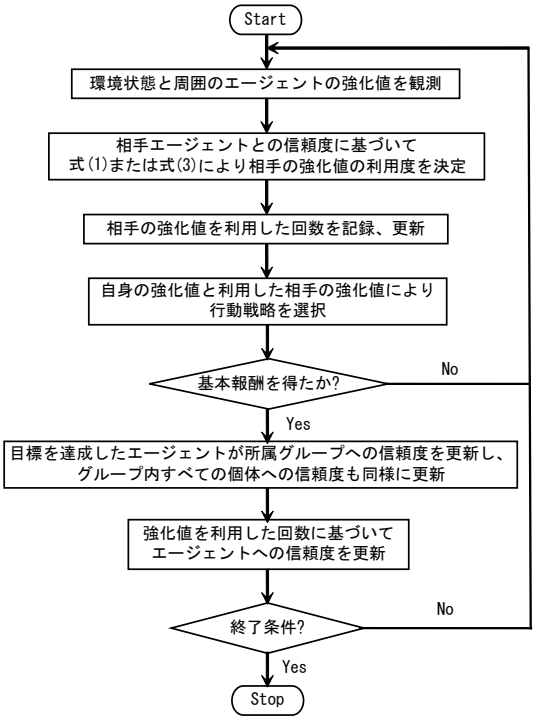

図 1: 信頼度に基づく行動選択アルゴリズム

自身に適応できないと信頼度の減少を引き起こし、適応できる他エージェントの強化値は信頼度の増加を促する。

2.2 信頼度を用いた強化値の計算

強化学習における二つの代表的な手法 (Q-learning と Profit Sharing) により、他エージェントとの信頼度と用いた強化値の計算方法について説明する。Q-learning の場合には、他エージェントの強化値をどの程度利用するかは式 (1) により計算される。エージェント間の信頼度は式 (2) のように更新される。

$$Q_t^*(s_t, a_t) = \sum_{o=1}^{n} (Q_t^o(s_t, a_t) \cdot \frac{C_t^o}{\sum_{i=1}^{n} C_t^i}) \qquad (1)$$

$$C_t^o = \sum_{i=1}^{t} (k_1 \cdot \frac{r_i^o}{R^*} + k_2 \cdot \frac{e_i^o}{E^*}) \qquad (2)$$

s_t: 時刻 t における状態； a_t: 時刻 t における行動；$Q_t^*(s_t, a_t)$: 他エージェントの強化値を利用する値；n: 知覚範囲内のエージェント数；C_t^o: 時刻 t におけるエージェント o の信頼度；$Q_t^o(S_t, a_t)$: 時刻 t におけるエージェント o の強化値；k_1 と k_2：重み係数；r_i^o: エージェント o の所属グループが得た報酬；R^*: すべてのエージェントグループが得た平均報酬；e_i^o: 時刻 i におけるエージェント o の強化値利用回数；E^*: すべてのエージェントの強化値利用回数；α: 学習率；γ: 割引率。

Profit Sharing 法を用いた信頼度に基づく強化値のインタラクティブ学習システムでは、信頼度 C_t^o の計算は式 (2) と同様である。他のエージェントの行動評価値 $w_t^*(s_t, a_t)$ は式 (3) のようになる。

$$w_t^*(s_t, a_t) = \sum_{o=1}^{n} w_t^o(s_t, a_t) \cdot \frac{C_t^o}{\sum_{i=1}^{n} C_t^i} \qquad (3)$$

$w_t(s_t, a_t)$: 時刻 t における自身の行動評価値;$w_t^o(s_t, a_t)$: 時刻 t におけるエージェント o の行動評価値;$w_t^*(s_t, a_t)$: 時刻 t における利用する行動評価値。

3 シミュレーション実験

提案した手法を検証するために、ダイナミックに変化する現象をリアルタイムで分析できるマルチエージェント・シミュレータ (構造計画研究所社製“ artisoc ”)[3] を用いて獲物追跡問題のシミュレーション実験を行なった。

3.1 シミュレーション条件

このシミュレーションでは、エージェントの基本特性の中でも特に重要な視野範囲と移動速度を 2 通りずつ設定し、表 1 のように A~D の 4 種類の異質ハンターエージェントを定義する。ここでは、視野範囲と移動速度の単位をセル (空間の最小単位) とし、100×100 セルの学習空間はループ空間として LOGO のタートルグラフィックスなどで用いられている座標系を用いる。ループ空間とは画面の上下、左右がつながっており、例えば画面右端へ消えたエージェントが再び左端に表示されるような空間である。獲物エージェントに隣接するハンターエージェントの数が 3 体以上になると、獲物の目標捕獲達成となる。ここでは、各ハンターエージェントを 10 体ずつ ($A_0 \sim D_9$) と獲物 10 体 ($P_0 \sim P_9$) を設定した。ハンターエージェントの目標行動はできるだけ早く獲物エージェントを捕獲することである。

表 1: 異質ハンターエージェントの設定条件

	視野 5×5	視野 10×10
速度 1 セル	A	B
速度 2 セル	C	D

4 種類のエージェントの異なる特性がソーシャルインタラクションにどのような影響を与えるかを分析するため、エージェントをランダムに初期配置した実験 (実験 1) と、初期位置がソーシャルインタラクションにどのような影響を与えるかを分析するため、特性の

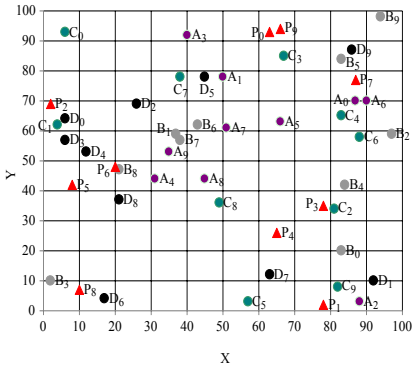

図 2: エージェントの初期配置 (実験 1)

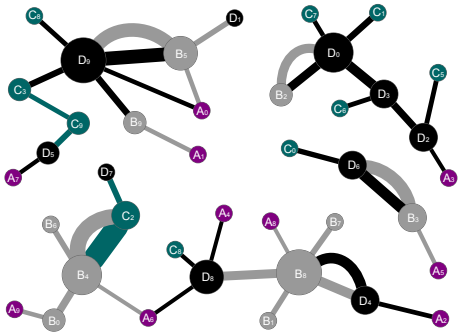

図 3: インタラクション頻度の可視化グラフ (実験 1)

同じエージェントを目標までの異なる初期位置に配置する実験 (実験 2) の 2 つを行なう。

3.2 実験 1(特性の異なるエージェントのランダム配置実験)

実験 1 では、表 1 のような 4 種類のエージェントを各 10 体ずつ図 2 ような初期位置にランダムに配置した。そのため、エージェントや獲物の初期位置のエージェント間のインタラクションへの影響はあまりないと考えられる。この図では、各エージェントの大きさを実際のサイズ (1 セル) の 3 倍に拡大して表記しており、○はハンターエージェント、△は獲物エージェントを示す。

学習途中で他エージェントの経験 (強化値) を何回利用したかについて、利用した回数の最も多いエージェント同士を可視化ツール Cytoscape[4] により関係の強さを線で示すと、図 3 のインタラクション頻度の可視化グラフを得られる。このグラフは他エージェントの強化値を 1 回を利用すると、1 回のインタラクションとし、よくインタラクションを行なったエージェントがこれにより明らかになる。この図のノードは A~D のそれぞれ 40 体エージェントを表し、各エージェントは自身とのインタラクションの回数が最も多いエージェントのみに線で繋いでいる。各ノード (エージェント) に繋がる線が多いほどこのノードのサイズが大きくなり、よくコミュニケーションを行なっているエージェントであることを示す。ノード間には強化値を利用した回数を線の幅で表し、線の色と同じ色のノードの強化値が他ノードに利用され、他ノードはこのノードの強化値を受け取ることを示している。

実験結果より、2 つ以上のノードに繋がるノードは、B と D 類のハンターエージェントが極めて多く、逆に A と C 類のエージェントが少ないことがわかる。すなわち、B と D 類のエージェントの強化値がよく他エージェントに利用され、他エージェントとのインタラクションが多いことがわかった。B と D 類のエージェントの共通特性は A、C 類のエージェントより視野が広いため、視野の広いエージェントはエージェント間のコミュニケーションを高める役割をはたしていると考えられる。

また、A 類のエージェントのノードの色と同じ線がないため、他エージェントとインタラクションを行なう能力が低いことが明らかである。同じ視野範囲である C 類のエージェントは A 類よりインタラクション能力がやや高く、エージェントの移動速度の速さはインタラクションの機会を増加させる可能性がある。しかし、B と D 類のエージェントは共に視野範囲が広いため、移動速度の影響はほとんどないこともわかった。

3.3 実験 2(同じ特性をもつエージェントの異なる初期配置実験)

実験 2 では、すべてのハンターエージェントは実験 1 の D 類と同じ特性 (10 × 10 の視野範囲と 2 セルの速度) に設定し、初期位置のみ異なる条件を与えた。ハンターエージェント (a、b、c、d が 10 体ずつ) は異なるエリア (a は左上、b は右上、c は左下、d は右下) にランダム配置され、図 4 のような初期位置とした。実験 1 と同様に、強化値の利用頻度を表すインタラクション情報を図 5 に示す。

実験結果より、図 4 の初期位置が近いエージェント b_2,b_3,b_5,b_6,b_7,b_8 は相互に強化値を利用し、図 5 のような 1 つのグループを作った。エージェント a_2,a_5,a_7 も初期位置が近いため、一緒に協調行動を行ないやすく、これらのエージェント間にはインタラクションの回数が多くなったことがわかった。

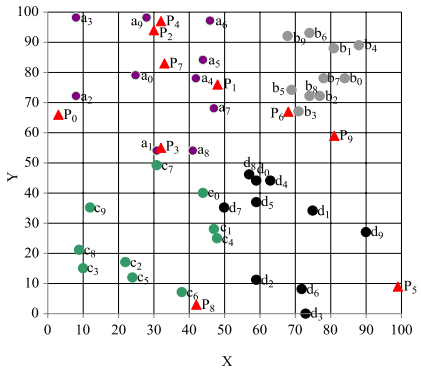

図 4: エージェントの初期配置 (実験 2)

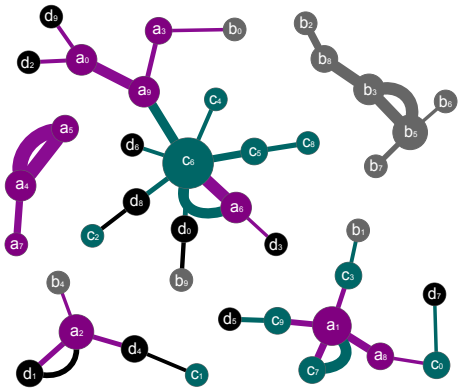

図 5: インタラクション頻度の可視化グラフ (実験 2)

また、視野範囲以内で獲物を観測できるエージェント (a_0,a_1,a_2,a_4,a_5,a_8,b_3,b_5,b_8,c_6 など) は、他エージェントとのインタラクションも多いことがわかる。これらのハンターエージェントの周辺には獲物がいるため、他エージェントとのインタラクション機会が増加したと考えられる。図 4 ではエージェント c_6 の周囲には 3 体の獲物がおり (空間の上下がつながっているため)、周辺に他ハンターエージェントがあまりいないが。エージェント c_6 の強化値インタラクションにより多くの他エージェントを呼び集め、協調して集中している獲物エージェントを捕獲するリーダーのような役割を果たしたものと考えられる。

さらに、左上エリアには 6 体の獲物がいるため, エージェント $a_0 \sim a_9$ は他エージェントとのインタラクションが多くなり、右下エリアのエージェント $d_0 \sim d_9$ は 1 体のみの獲物までの距離が遠く、他エージェントとのソーシャルインタラクション回数が少なくなった。

4 結 言

本研究ではエージェント間の強化値インタラクションを通じて、優れた協調能力をもつインタラクティブ強化学習システムにおいて、インタラクション情報を可視化する手法を提案した。異質エージェント間のインタラクションに影響を与える要素を分析するため、エージェント特性と初期配置に関して異なる条件を設定し、異質エージェント間のインタラクション情報を可視化する実験を行なった。

インタラクションに影響を与える特性では、視野の広いエージェントは他エージェントとのインタラクションが多く、協調を促進する役割を果たすことがわかった。また、初期位置の近くに配置されるハンターエージェント間のインタラクションが多くなるため、ハンターエージェントの周辺に獲物がいる場合には、他エージェントとのインタラクションの機会が増加することもわかった。

今後の課題として、エージェント間の強化値の利用回数のみを可視化するのではなく、より多くの協調に関するインタラクション情報を可視化することで、協調行動を促進できる要素を見出し、複雑な環境にも自律的に適応できるマルチエージェントインタラクティブシステムを構築する必要があると考えられる。

参考文献

[1] M.Tan, “ Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents, ”*Proc. of the 10th International Conf. on Machine Learning*, pp.330-337, 1993.

[2] 張坤, 前田陽一郎, 高橋泰岳, “ 異質のマルチエージェント間のインタラクションを考慮した学習モデル, ”日本知能情報ファジィ学会誌, Vol.24, No.5, pp.1002-1011, 2012.

[3] MAS コミュニティ, http://mas.kke.co.jp/modules/tinyd0/index.php?id=9

[4] Cytoscape, http://www.cytoscape.org

連絡先

〒 910-8507　福井県福井市文京 3-9-1
福井大学 大学院工学研究科 システム設計工学専攻
張 坤 (進化ロボット研究室)
Tel & Fax: 0776-27-8050
E-mail: kzhang@ir.his.u-fukui.ac.jp