2P4-1

マルチエージェント強化学習における主観的効用の進化過程に関する分析

Evolution Process of Subjective Utilities in a Multi-agent Reinforcement Learning Context

宮脇 昌哉 *1 Masaya Miyawaki　森山 甲一 *1 Koichi Moriyama　武藤 敦子 *1 Atsuko Mutoh　松井 藤五郎 *2 Tohgoroh Matsui　犬塚 信博 *1 Nobuhiro Inuzuka

*1名古屋工業大学 大学院工学研究科 情報工学専攻
Department of Computer Science, Graduate School of Engineering, Nagoya Institute of Technology

*2中部大学 生命健康科学部 臨床工学科
Department of Clinical Engineering, College of Life and Health Sciences, Chubu University

Utility-based reinforcement learning is a reinforcement learning method using subjective utilities derived from rewards. Every agent has its own utility derivation function and it is known that reward-based evolution brings it a function leading mutual cooperation in iterative prisoner's dilemma games. However, the evolution process itself has not yet been investigated in detail, e.g., how and why the functions evolved that way. This work investigates the evolution process itself from the viewpoints how the evolved function affects its owner's behavior and what determines the evolution direction. The result suggests that the direction can be explained by action preference depending only on one agent.

1. はじめに

チームでのスポーツのように、人間は集団内で行動する際に周囲の状況を判断して協調的な行動を選択することが可能である。一方、エージェントは強化学習を用いて自身の報酬の最大化を目的とした行動の学習を行うことが可能であるが、マルチエージェント環境では複数のエージェントが互いに干渉しあい、協調行動の学習が困難な場合が存在する。

そこで、人間の場合は相互協調自体に報酬を見出すという Rilling ら [1] の研究等をもとに、森山ら [2][3] は エージェント内部に情動機構の存在を仮定し、エージェント固有の情動機構によって客観的報酬から生成される主観的効用を Q 学習における報酬として用いる効用利用 Q 学習を提案した。そしてこの主観的効用を報酬に基づいて進化させることで、繰り返し囚人のジレンマゲームにおいて相互協調をもたらす主観的効用が得られることを確認した。

しかし、先行研究では相互協調をもたらす主観的効用が得られることを確認したにすぎず、主観的効用がどのように進化し、その要因が何であるかという点については検証されていない。よって、本研究では主観的効用の進化過程を明らかにし、主観的効用の性質について考察する。

2. 繰り返し囚人のジレンマ

繰り返し囚人のジレンマとは 2 人のプレイヤーがそれぞれ協調 (C) または裏切り (D) を同時に選択し、その組み合わせによって報酬 $r \in \{T, R, P, S\}$ を得るという一連の流れを複数回繰り返すゲームである。ただし $T > R > P > S$, $2R > T + S$ である。2 人のプレイヤーの行動組み合わせを (X, Y) とすると、行動組み合わせ (X, Y) と行動 X を選択したプレイヤーが得られる報酬との間の対応関係は (C, C) のとき R、(C, D) のとき S、(D, C) のとき T、(D, D) のとき P である。このことはどちらのプレイヤーも行動 D を選択したほうが相手の行動に関わらずより大きな報酬を得られるものの、結果として相互裏切り (D,D) が発生し両者ともに報酬 P を得ることになり、どちらも報酬 R を得られる相互協調 (C,C) の場合よりも、得られる報酬が小さくなってしまうことを意味している。

連絡先: 宮脇 昌哉，名古屋工業大学大学院工学研究科情報工学専攻，名古屋市昭和区御器所町，
m.miyawaki.474@nitech.jp

3. 効用利用 Q 学習と主観的効用の進化

効用利用 Q 学習 [2] では、エージェント固有の情動機構を表す関数である効用導出関数 $u(r)$ によって客観的報酬 r から生成される主観的効用 $u(\equiv u(r))$ を、従来の Q 学習における報酬の代わりとして用いる。森山らは主観的効用を報酬に基づき進化させ、客観的報酬から相互協調を導く主観的効用が得られるか実験を行った [3]。

3.1 実験の流れ

まず、集団内のあるエージェントの組が繰り返し囚人のジレンマゲームを行うことを考える。各ステップにおいて効用利用 Q 学習を用いてより多くの効用を得られるよう行動を選択し、環境から報酬を受け取る。これをゲーム終了までの全ステップに対して繰り返し行う。

次にこの 1 ゲームを集団内のエージェントの組全てに対して行い、各エージェントは自分以外の全エージェントとのゲームで得られた報酬の総和を保持しておく。

全てのゲームが終了すると、各エージェントの効用導出関数を染色体、累積報酬を適応度として遺伝的アルゴリズムによる進化を行い、次世代のエージェントの集団を生成し、これを指定した世代数繰り返す。

エージェント数は 100、世代数は 10000、繰り返し囚人のジレンマの繰り返し回数は 1000 回とする。効用導出関数 $u(r)$ は $u(r) \equiv ar^3 + br^2 + cr + d \quad (a, b, c, d \in [-10, 10])$ とし、係数 a, b, c, d を遺伝子とする。また、各遺伝子の初期値は $[-10, 10]$ の範囲でランダムとし、Q 学習の状態数は 1、つまり状態を考慮しないものとする。そして、Q 値の初期値 $Q_0(C) = Q_0(D) = 0$、ゲームの利得 $T = 5, R = 3, P = 1, S = 0$ とする。

3.2 実験結果

上記の実験を今回改めて 100 回行ったところ、図 1 のように係数 a, b が一定の傾向に収束した。実験 100 回のうち 79 回で、繰り返し囚人のジレンマにおける行動選択 1 回あたりの平均利得が 2.5 $(= (T + S)/2)$ を上回り、少なくとも 1 回は相互協調が発生することが確認された。

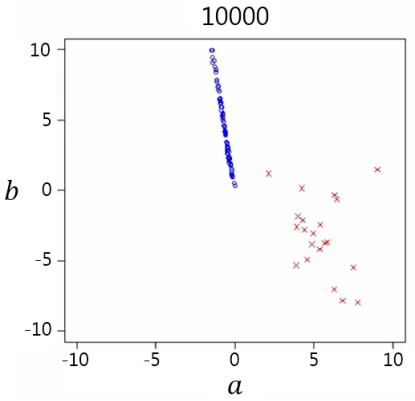

図 1: 各回における 10000 世代目の係数 a, b の平均値. （x 軸は係数 a、y 軸は係数 b を表す. 平均利得が 2.5 より大きい試行は青い丸印で、2.5 以下の試行は赤い×印でプロット）

4. 主観的効用の進化の方向

森山ら [3] は、繰り返し囚人のジレンマゲームにおいて客観的報酬から相互協調を導く主観的効用が得られることを確認したが、主観的効用がどのように進化し、また進化の方向を決める要因が何であるかについては検証していない。そこで、本研究では主観的効用の進化過程について分析する。

まず、主観的効用がどのように進化するかを調査する。効用導出関数の係数 a, b によって得られる平面を図 2 の通りセル単位で分割し、セルごとに 3.1 節と同様の実験をそれぞれ 100 回行うことで、各セルからどのように主観的効用が進化するかを確認した。ただし、今回は各セルからその近傍のセルへの進化を調べるため、世代数を 10000 ではなく 100 とした。実験の結果、図 3 の青矢印が示す主観的効用の大まかな進化の方向が得られた。おおむね $|a|$ が小さくなる方向へ進化することが分かる。

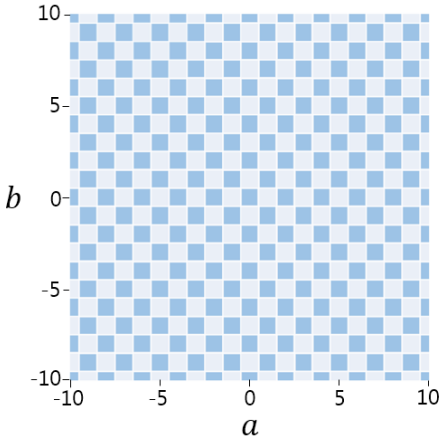

図 2: セル単位での分割図（x 軸は係数 a、y 軸は係数 b）

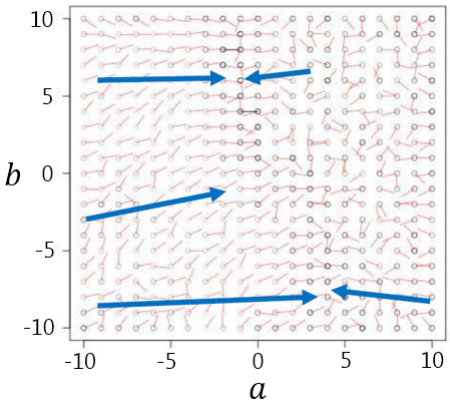

図 3: セル毎の主観的効用の進化の方向（丸印を起点に赤線の方向に進化する. 青矢印が大まかな進化傾向を表す. x 軸は係数 a、y 軸は係数 b）

5. 進化の方向を決める要因の分析

5.1 評価指標の導入

主観的効用がどのように進化するかについては大まかな傾向が得られたので、次にその要因について分析を行う。分析にあたり、主観的効用の進化によって各エージェントの行動選択に変化が生じることから、主観的効用と行動選択の関係に着目する。そして、自身のある行動によって得られる効用がその他の行動によって得られる効用よりも大きい場合、前者の行動がより多く選択されると考えられるので、まず自身の行動のみに依存する効用を定義し、そしてエージェントがどの行動を選好するかを表す指標である行動傾向（協調傾向および一致傾向）を自身の行動のみに依存する効用を用いて定義する。

5.1.1 協調傾向

まず、自身の協調行動や裏切り行動に対する効用 u_s と、協調行動に対する選好を表す協調傾向 $rate_s$ を以下の通り定義する。ただし自身の行動を $X \in \{C, D\}$ とする。

$$u_s(X) \equiv \begin{cases} \dfrac{u(R) + u(S)}{2} & (X = C) \\ \dfrac{u(T) + u(P)}{2} & (X = D) \end{cases}$$

$$rate_s \equiv \begin{cases} \dfrac{u_s(C)}{u_s(D)} & (u_s(D) \neq 0) \\ u_s(C) & (u_s(D) = 0) \end{cases}$$

u_s は相手が等確率で行動選択を行ったという仮定の下で、自身が協調行動を選択したときの効用と裏切り行動を選択したときの効用を表している。$rate_s$ は $u_s(C)$ と $u_s(D)$ の大小関係を表しており、$u_s(C) = u_s(D)$ となる閾値と $rate_s$ の値を比較することで、

$$u_s(D) > 0 \text{ のとき } rate_s \geq 1$$
$$u_s(D) = 0 \text{ のとき } rate_s \geq 0$$
$$u_s(D) < 0 \text{ のとき } rate_s \leq 1$$

を満たすならば協調選好、満たさないならば裏切り選好と判定する。また、$rate_s$ の値が閾値から遠いほど、より強い選好を

示す。

5.1.2 一致傾向

次に、相手と同じ行動や相手と異なる行動に対する効用 u_m と、それらへの選好を表す一致傾向 $rate_m$ を導入する。ただし、「相手と同じ行動」を $same$、「相手と異なる行動」を $different$ とし、$match \in \{same, different\}$ とする。

$$u_m(match) \equiv \begin{cases} \dfrac{u(R)+u(P)}{2} & (match = same) \\ \dfrac{u(T)+u(S)}{2} & (match = different) \end{cases}$$

$$rate_m \equiv \begin{cases} \dfrac{u_m(same)}{u_m(different)} & (u_m(different) \neq 0) \\ u_m(same) & (u_m(different) = 0) \end{cases}$$

$rate_s$ と同様に $rate_m$ は $u_m(same)$ と $u_m(different)$ の大小関係を表しており、$u_m(same) = u_m(different)$ となる閾値と $rate_m$ の値を比較することで、

$$u_m(different) > 0 \text{ のとき } rate_m \geq 1$$
$$u_m(different) = 0 \text{ のとき } rate_m \geq 0$$
$$u_m(different) < 0 \text{ のとき } rate_m \leq 1$$

を満たすならば一致選好、満たさないならば不一致選好と判定する。また、$rate_m$ の値が閾値から遠いほど、より強い選好を示す。

5.2 行動傾向の算出

上記の通り定義した行動傾向を図 2 の各セルに対して算出する。具体的には、セル内に存在する初期世代でのエージェント 100 個体の効用導出関数の係数 a, b, c, d の平均値に対して行動傾向を求め、これを全てのセルに対して行う。

行動傾向の算出の前に、まず $u_s(D), u_m(different)$ の値をセル毎に算出すると、図 4 の通りである。図 4 のピンク色の箇所が正の値、水色の箇所が負の値を表すので、$u_s(D), u_m(different)$ ともにおおむね $a < 0$ のセルでは負の値を、$a > 0$ のセルでは正の値を示している。

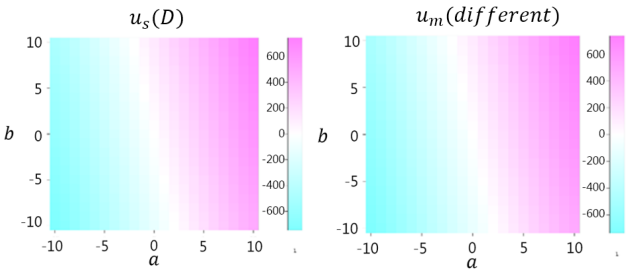

図 4: 裏切り行動や不一致行動に対する効用（左：$u_s(D)$、右：$u_m(different)$）

次に、行動傾向の値をセル毎に算出すると、図 5 の通りである。また、$\exp(rate_s)$ の最大値は 2.57、$\exp(rate_m)$ の最大値は 3.81 であった。これらのことから、ほぼ全てのセルにおいて協調傾向、一致傾向ともに閾値である 1（図 5 では $\exp(1) = 2.72$）を下回っていることが分かる。したがって、図 4 と図 5 から、行動傾向の判定条件をもとに各セルの行動傾向を判定した結果が図 6 である。おおむね $a < 0$ のセルは協調一致選好、$a > 0$ のセルは裏切り不一致選好と判定される。

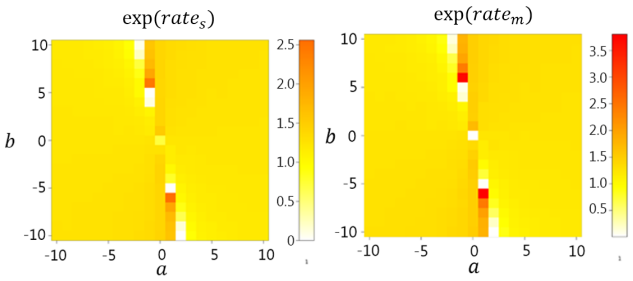

図 5: セル毎の行動傾向の値（左:$\exp(rate_s)$、右:$\exp(rate_m)$）

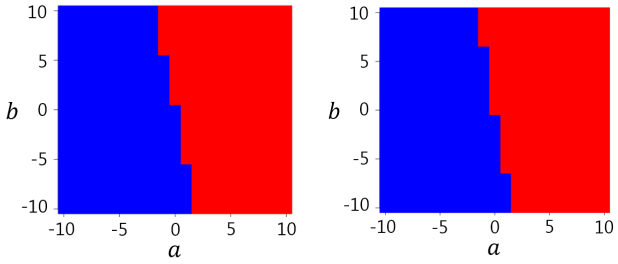

図 6: セル毎の行動傾向の判定結果（左：セル別協調傾向（赤：裏切り選好、青：協調選好）、右：セル別一致傾向（赤：不一致選好、青：一致選好））

また、行動傾向の値が閾値から遠いほど、すなわちここでは小さいほどより強い選好を示すので、図 7 左のように協調傾向と一致傾向のうち、値がより小さい、つまりより強い選好を表す支配的な行動傾向を求め、図 7 右のように支配的な行動傾向によって各セルを裏切り優先エリア（赤）、不一致優先エリア（橙）、協調優先エリア（青）、一致優先エリア（水色）に分類する。

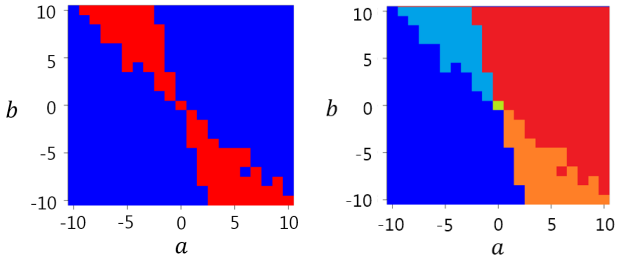

図 7: セル毎の支配的な行動傾向とそれによるエリア分類（左：セル別支配的傾向（赤：一致傾向、青：協調傾向）、右：エリア分類図（赤：裏切り優先、橙：不一致優先、青：協調優先、水色：一致優先））

5.3 行動傾向による分析

次に、前節で分類した各エリアにおける行動傾向を用いて、主観的効用の進化の方向を決める要因について分析する。なお、分析のため、主観的効用の進化による行動傾向の変化が分かりやすいよう図 5 のカラースケールを調整した図 8 を用いる。また、主観的効用の進化の大まかな方向を表す図 3 の矢印を図 7 右に反映したものが図 9 である。図 9 は、各エリアにおいて主観的効用がどの方向に進化するかを表している。

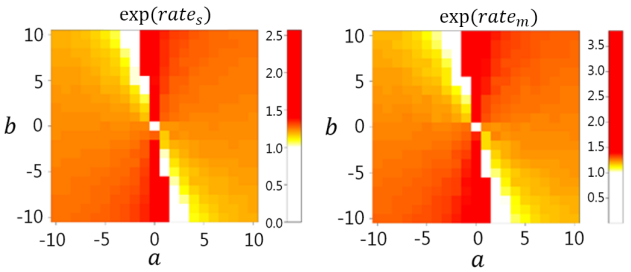

図 8: セル毎の行動傾向の値（左:exp($rate_s$)、右:exp($rate_m$)）

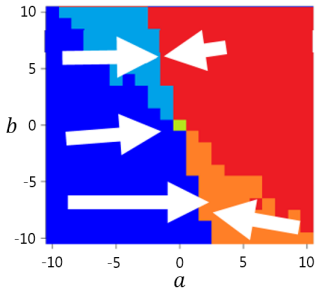

図 9: エリアごとの大まかな進化の方向（白矢印の方向に進化する. x 軸は係数 a、y 軸は係数 b を表す. 赤：裏切り優先、橙：不一致優先、青：協調優先、水色：一致優先）

5.3.1 一致優先エリア（水色）

図 8 から、一致優先エリアでは裏切り優先エリアに近づくにつれて各行動傾向の値が小さくなる。つまり協調選好や一致選好が強くなるので、相互協調が増加する。また、このエリアは一致優先なので、裏切り優先エリアに近づくにつれて一致選好がより支配的になり、相手の裏切りに対して自身が裏切りを選択し、一方的に裏切られるリスクを回避する可能性が高くなる。したがって、裏切り優先エリアに近づくにつれてより多くの利得が得られるようになり、図 9 の左上の矢印のように進化する。

5.3.2 協調優先エリア（青色）

図 8 から、協調優先エリアでは不一致優先エリアに近づくにつれて各行動傾向の値が大きくなる。つまり一致選好や協調選好が弱まるので裏切りを選択する回数が増加する。ただしこのエリアは協調優先なので協調選好以上に一致選好が弱まり、一方的な裏切りを選択する可能性が高くなる。そのため不一致優先エリアに近づくにつれてより多くの利得が得られ、図 9 の左下の矢印のように進化すると考えられる。

5.3.3 裏切り優先エリア（赤色）

裏切り優先エリアでは一致優先エリアに近づくにつれて各行動傾向の値が大きくなる。つまり裏切り選好や不一致選好が弱まり、協調を選択する回数が増加する。またこのエリアは裏切り優先なので、裏切り選好以上に不一致選好が弱まり、相互協調の回数が増加する。よって一致優先エリアに近づくにつれてより多くの利得が得られるようになり、図 9 の右上の矢印の方向へと進化する。

5.3.4 不一致優先エリア（橙色）

不一致優先エリアでは協調優先エリアに近づくにつれて各行動傾向の値が小さくなる。またこのエリアは不一致優先なので、協調優先エリアに近づくにつれて不一致選好がより支配的になり、相互協調や相互裏切りを回避して一方的な裏切りを選択する可能性が高くなる。したがって、協調優先エリアに近づくにつれてより多くの利得が得られるようになり、図 9 の右下の矢印の方向に進化が行われる。

5.3.5 エリア間

裏切り不一致選好エリアと協調一致選好エリアの境界を含むセルではセル内部の全エージェントの主観的効用の値によってどちらか一方のエリアに分類されるが、実際には両方のエリアの個体群が混在し、これらの個体群によってゲームが行われる。一致優先エリアの個体群と裏切り優先エリアの個体群とのゲームでは相互裏切りが繰り返し発生し、裏切り優先エリアの個体同士のゲームで相互裏切りが多く発生する一方で一致優先エリアの個体同士では相互協調が数多く発生する。したがって一致優先エリアの個体群が裏切り優先エリアの個体群よりも多くの利得を得ることになり、進化的に優位に立つ。

不一致優先エリアの個体群は協調優先エリアの個体群をほぼ一方的に裏切ることができる。協調優先エリアの個体群同士では相互協調が多く発生するが、不一致優先エリアの個体群同士では一方的な裏切りが行われる。したがって不一致優先エリアの個体群が協調優先エリアの個体群よりも多くの利得を得ることになり、進化的に優位に立つ。

以上のことから、各エージェントの行動傾向の違いによって得られる累積利得に差が生じ、図 9 の矢印の方向に進化が行われたと考えられる。

6. まとめ

主観的効用の進化過程について主観的効用から算出される行動傾向を用いて分析を行うことで、行動選択に一定の傾向が生じていることが判明した。さらに、固有の行動傾向をもったエージェント間のゲームによって行動組み合わせの発生回数及び累積報酬に差が生じることが、累積報酬を適応度とする主観的効用の進化に影響を与えているということが分かった。

今後の課題としては、今回導入した行動傾向の妥当性を検証することが挙げられる。

謝辞

本研究の一部は、JSPS 科研費 JP16K00302、栢森情報科学振興財団、および堀科学芸術振興財団の助成を受けて行われた。

参考文献

[1] J. K. Rilling, D. A. Gutman, T. R. Zeh, G. Pagnoni, G. S. Berns, and C. D. Kilts: *A Neural Basis for Social Cooperation*, Neuron, Vol. 35, pp. 395–405, 2002.

[2] K. Moriyama: Utility based Q-learning to facilitate cooperation in Prisoner's Dilemma games, *Web Intelligence and Agent Systems*, Vol. 7, No. 3, pp. 233–242, 2009.

[3] K. Moriyama, S. Kurihara, and M. Numao: Evolving Subjective Utilities: Prisoner's Dilemma Game Examples. *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 233–240, 2011.