ブレンド交叉を用いたGAによる主観的効用の進化

Evolution of Subjective Utilities by GA with BLX-α

岡田 直也 *1 Naoya Okada　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 Q-learning, which uses subjective utilities as rewards of Q-learning, has been proposed and the utilities that derive mutual cooperation in a Prisoner's Dilemma game have been successfully evolved by real-coded genetic algorithm (RCGA). However, in that work, the genes were simply exchanged in the evolution process like a bit-string GA and the search space was not so wide as a result. This work investigates the evolution of the subjective utilities by RCGA with blend crossover (BLX-α) that has a powerful search ability by generating various chromosomes.

1. はじめに

集団内において人間は譲り合いなどの協調行動をするように，マルチエージェント環境において，人間の場合と同様の環境を想定すると，エージェントにも協調的な行動が求められる．適切な行動を選択するように学習するエージェントを考えるとき，強化学習 [1] により自身の報酬を最大化するような行動を学習するが，マルチエージェント環境ではエージェント間の相互作用により，全エージェントの報酬を最大化することが不可能な場合があり，個々が報酬を最大化することを目的とする強化学習では協調行動を学習することは困難である．

そこで人間について改めて考えると，人はそれぞれ違った価値観を持ち，客観的報酬をそのまま受け取るのではなく，自身の価値観を通して得られる主観的効用に基づいて学習し，行動を決定するのではないかと考えられる．森山ら [2] は，この考えからエージェント内部に何らかの情動機構の存在を仮定し，主観的効用を Q 学習 [3] の報酬として用いる効用利用 Q 学習を提案した．また，報酬をもとに主観的効用を進化させることで，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 [4]．

この研究では，進化の際の遺伝的アルゴリズム (GA) に，効用導出関数の係数を染色体の遺伝子とする実数値 GA を用いているが，その交叉手法に一様交叉を用いているため，実数値 GA として充分な探索が行われず，個体群も同一個体が多く生成されていた．そこで，遺伝子操作の交叉を実数値 GA に適するように提案されているブレンド交叉とすることで探索性能を向上させるとともに，多様性を維持した状態での主観的効用の進化の様子について考察する．

2. 準備

2.1 囚人のジレンマゲーム

囚人のジレンマゲーム [5] とはゲーム理論のゲームの一つで 2 人 2 行動ゲームである．プレイヤー A, B はそれぞれ行動 C(協調) と D(裏切り) から選択し，その組み合わせに応じた利得 $T, R, P, S \in r$ $(T > R > P > S)$ をそれぞれ得る．行動の組み合わせと利得の関係は表 1 のように表される．ただし，A が行，B が列から行動を選択し，A が左側，B が右側の利得を得る．

表 1: 囚人のジレンマゲームにおける利得表

$A \backslash B$	C	D
C	R, R	S, T
D	T, S	P, P

2.2 Q 学習

ある環境内でエージェントが現在の状態から最適と考えられる行動を選択するように学習する強化学習 [1] の代表例として，Q 学習 [3] がある．

エージェントは時刻 t において状態 $s_t \in S$ を知覚し，方策 π に基づく行動 $a_t \in A(s_t)$ を選択する．ここで，S はある環境において可能な状態の集合，$A(s_t)$ は状態 s_t において可能な行動の集合を表す．行動を選択後，エージェントは報酬 r_{t+1} を受け取り，新しい状態 s_{t+1} を知覚する．これらの情報より，Q 学習は行動価値 Q を最適方策 π^* における行動価値 Q^* に近づけるように以下の式により更新し，最適な行動を学習する．

$$Q_{t+1}(s,a) = \begin{cases} Q_t(s_t,a_t) + \beta\delta_t & if (s,a) = (s_t,a_t), \\ Q_t(s,a) & otherwise. \end{cases} \quad (1)$$

$$\delta_t \equiv r_{t+1} + \gamma \max_{a \in A(s_{t+1})} Q_t(s_{t+1},a) - Q_t(s_t,a_t). \quad (2)$$

上記で β は学習率，γ は割引率と呼ばれるパラメータで，それぞれ $0 < \beta \leq 1$，$0 \leq \gamma < 1$ の値を取る．

また，客観的報酬 r からエージェント固有の効用導出関数を適用して得られる主観的効用 $u(r)$ を Q 学習の報酬として用いる手法を効用利用 Q 学習 [2] という．

連絡先: 岡田 直也，名古屋工業大学大学院工学研究科情報工学専攻，愛知県名古屋市昭和区御器所町，n.okada.862@nitech.jp

2.3 遺伝的アルゴリズム

遺伝的アルゴリズム (Genetic Algorithm:GA) [6] とは，生物の進化の過程を模した最適化アルゴリズムである．複数の遺伝子からなる染色体と呼ばれる個体を複数用い，その個体群に選択，交叉，突然変異などの遺伝的操作を繰り返し行うことで，問題に応じた目的関数によって各個体に与えられる適応度を最大化する個体を求める．最も基本的な GA の一連の流れは以下の図 1 のように行われる．

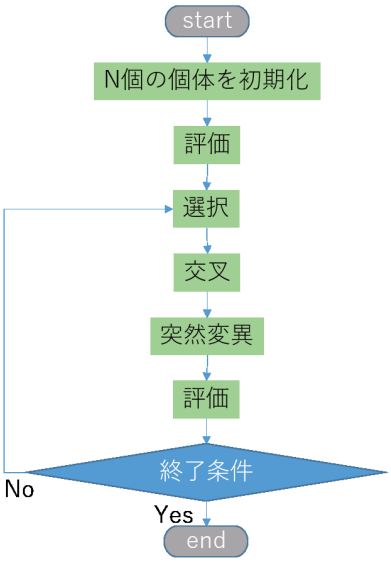

図 1: 遺伝的アルゴリズムのフローチャート

また，染色体の遺伝子として実数ベクトルを用いるものを実数値 GA といい，連続な探索空間で関数を最適化する問題において高い性能を示す．

3. 先行研究

森山ら [4] は囚人のジレンマゲームにおける主観的効用を客観的報酬に基づき進化させ，相互協調を導く主観的効用が得られるか検証した．効用導出関数 $u(r)$ を $u(r) = ar^3 + br^2 + cr + d$ とし，係数 a, b, c, d を遺伝子とする実数値 GA により相互協調を起こすような進化が得られた．また宮脇ら [7] は，効用導出関数の係数 a, b を座標軸とする平面領域において，主観的効用はおおむね $\|a\|$ が小さくなる方向へ進化することを確認し，進化の方向を決める要因として ab 平面を行動傾向によって分類できることを示した．

しかし，これらの研究では交叉手法として一様交叉が用いられていたため，親個体の遺伝子の実数値を入れ換えているだけであり，後期世代になるにつれて同一個体が多く生成され，探索性能としては低いと考えられる．

4. 提案手法

遺伝的操作の交叉において，各遺伝子の値を入れ換えるのではなく，親個体に基づいて新たな値を生成するブレンド交叉 (Blend Crossover:BLX-α)[8] を適用する．この操作を用いることでできるだけ同一個体が生成されるのを防ぎ，個体群に多様性を持たせる．また，従来手法では探索できていないであろう探索点にも進化させることを可能にし，それによる主観的効用の進化の様子について観察する．

4.1 ブレンド交叉 (BLX-α)

ブレンド交叉 (以下，BLX-α とする)[8] は，Esheleman によって考案された交叉手法である．親個体の実数ベクトルの各変数の区間 d_i を両側に αd_i だけ拡張した区間から一様乱数に従ってランダムに子個体を生成する．すなわち，親個体の周辺の各辺が軸に平行な超直方体の領域が子個体の生成領域となる．以下では遺伝子長=2 の簡単な例を示す．

親個体のベクトル (2 次元) をそれぞれ x 座標，y 座標で示し，(x_1, y_1)，(x_2, y_2) とすると，子個体の x 座標と y 座標の発生範囲を以下のように決める．

- $dx = \|x_1 - x_2\|$，$dy = \|y_1 - y_2\|$，
- $min\_x = min(x_1, x_2)$，$min\_y = min(y_1, y_2)$，
- $max\_x = max(x_1, x_2)$，$max\_y = max(y_1, y_2)$ とおく．

ここで，$min(a, b)$ は a, b のうち小さな方の値，$max(a, b)$ は a, b のうち大きな方の値を取ることとする．また，

- 次世代の x 座標の最小値を min_cx とし，$min\_cx = min\_x - \alpha dx$ とする．
- 次世代の x 座標の最大値を max_cx とし，$max\_cx = max\_x + \alpha dx$ とする．
- 次世代の y 座標の範囲も同様にして，$min\_cy = min\_y - \alpha dy$，$max\_cy = max\_y + \alpha dy$ とする．

以下の図 2 に遺伝子長=2 (2 次元) における BLX-α による子個体生成の例を示す．

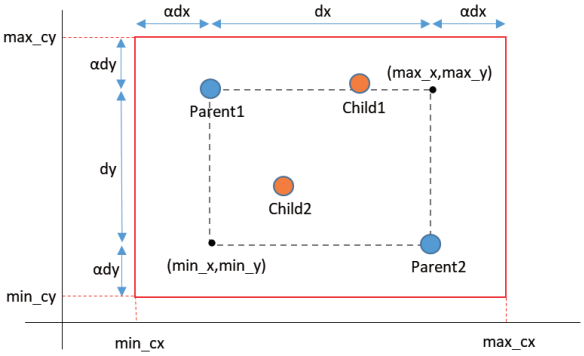

図 2: BLX-α の概要 (2 次元)

4.2 提案への実装

本研究では，囚人のジレンマゲームにおける主観的効用の進化に用いる遺伝的アルゴリズムの遺伝的操作として，選択にはルーレット選択，交叉には BLX-α を適用し，突然変異は正規分布に基づいたガウス突然変異を行う．ただし，BLX-α のパラメータ α の値は実験に応じて適宜設定することとする．また，BLX-α は突然変異を併用せずとも多様な解を生成可能な交叉として考案された [8] ことから，遺伝的操作として突然変異を用いない場合についても同時に検証を行う．以下の図 3 には，先行研究での実験に新たに交叉として BLX-α を適用した遺伝的アルゴリズムの一連の流れについて示す．ここで，N は個体数，G は世代数の値である．

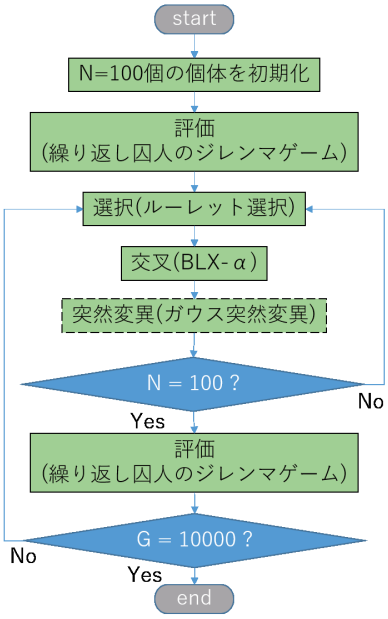

図 3: BLX-α を用いた遺伝的アルゴリズムの一連の流れ

5. 実験

5.1 実験手法

森山ら [4] の実験と同様に，3 次の効用導出関数の係数を実数値 GA で扱う染色体の遺伝子として生成し，その個体群に総当たりで繰り返し囚人のジレンマゲームを行わせる．ゲームで得られる合計利得の和を適応度とし，それに基づいて効用導出関数を実数値 GA によって進化させる．遺伝的操作における交叉手法として BLX-α を適用し，主観的効用の進化について観察する．ただし，交叉以外の設定は先行研究のものを用いる．そして，BLX-α のランダム性より，本実験では突然変異を行う場合と行わない場合についてそれぞれ $\alpha = 0.3, 0.4, 0.5$ として検証した．実験の主な流れは前章の図 3 で示したようになる．また，ゲームの利得から $(T + S)/2 = 2.5$ となるため，平均利得が 2.5 以上となるとき相互協調が起こっていることを表す．

5.2 実験結果

(突然変異あり)

$\alpha = 0.3, 0.4$ のとき，相互協調を起こすような進化が見られた．しかし，どちらも相互協調が起こるまでの進化が従来手法に比べて遅くなる傾向となった．図 4 は $\alpha = 0.4$ のある試行における各世代の 100 個体の 1 ゲームあたり平均利得を示している．ここでは 2500 世代前後で相転移が起こり，その後 2.5 以上となっていることから相互協調を維持するように進化していることが分かる．また，図 5 は同じ試行における各世代の効用導出関数の係数 a, b それぞれの平均値を示し，図 6 は 10000 世代時点の 100 個体の効用導出関数の係数 a, b の値を先行研究での値とともに示す．図 5 では，文献 [7] で示されたような方向へ進化する様子が確認できた．しかし，相互協調へ進化するのが遅く世代数がかかった試行では，相互裏切りを起こす領域内を転々とし，その後相互協調へ進化が進むという結果も得られた．図 6 からは従来手法に比べて個体群が分散するように生成されていることが分かる．

$\alpha = 0.5$ のときは，進化するにつれて絶対値の大きい値が生成されていまい，途中世代で測定不可になるような値に発散する結果となった．

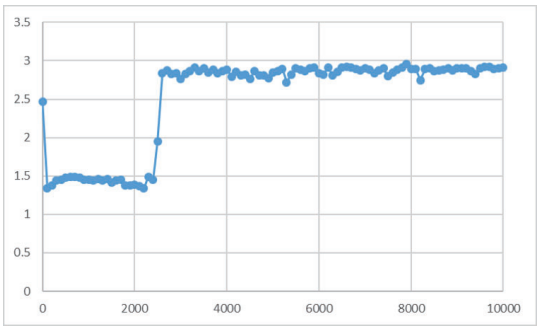

図 4: 突然変異あり，$\alpha = 0.4$ のある試行での 100 世代ごとの 1 ゲームあたり平均利得 (横軸は世代数，縦軸は平均利得を表す．)

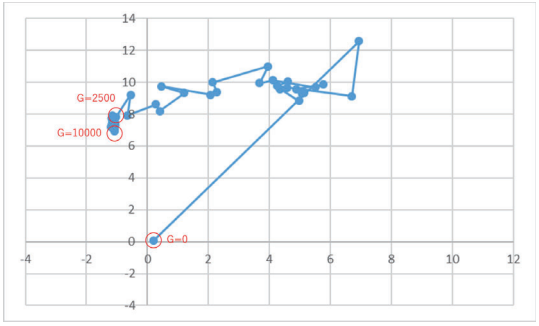

図 5: 突然変異あり，$\alpha = 0.4$ のある試行での 100 世代ごとの係数 a, b の平均値 (横軸は係数 a，縦軸は係数 b を表す．)

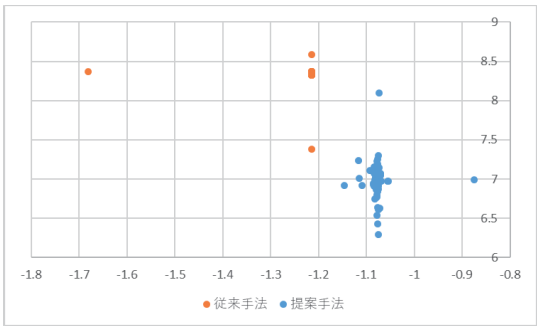

図 6: 突然変異あり，$\alpha = 0.4$(青) と従来手法 (橙) のある試行での 10000 世代目における個体群の効用導出関数の係数 a, b の値 (横軸は係数 a，縦軸は係数 b を表す．)

(突然変異なし)

$\alpha = 0.3, 0.4$ のとき，100 個体は早期世代で値が収束してしまい，それ以降の進化が止まってしまう結果となった．図 7 は $\alpha = 0.4$ のある試行における各世代の効用導出関数の係数 a, b それぞれの平均値を示し，値が収束してしまう進化の様子が読み取れる．また，図 8 は各世代の 100 個体の 1 ゲームあたり平均利得を示す．平均利得は約 1.4 の値が続き，これは相互裏切りが続いていることを意味している．

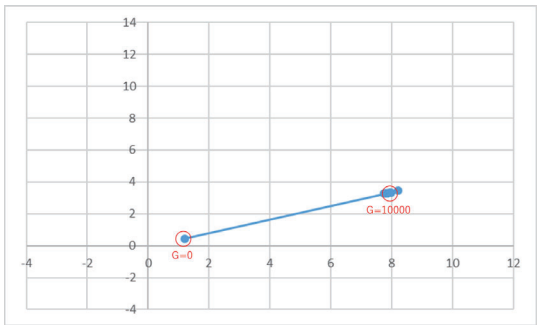

図 7: 突然変異なし，$\alpha = 0.4$ のある試行での 100 世代ごとの係数 a, b の平均値 (横軸は係数 a，縦軸は係数 b を表す．)

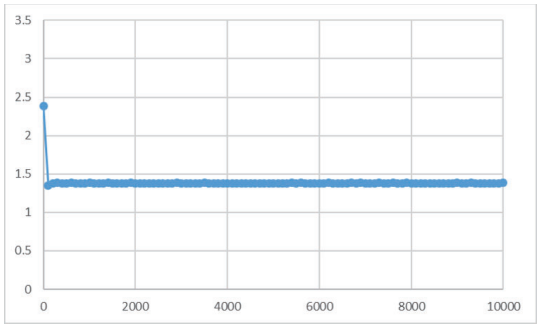

図 8: 突然変異なし，$\alpha = 0.4$ のある試行での 100 世代ごとの 1 ゲームあたり平均利得 (横軸は世代数，縦軸は平均利得を表す．)

$\alpha = 0.5$ のときは突然変異ありの時と同様に，早期世代でそれぞれの遺伝子の値が膨大な値へ進化していく結果となった．図 9 は $\alpha = 0.5$ のある試行での 100 世代ごとの 1 ゲームあたり平均利得を示し，平均利得は 2.5 前後となるように進化するものの，相互協調を維持するような結果は得られなかった．

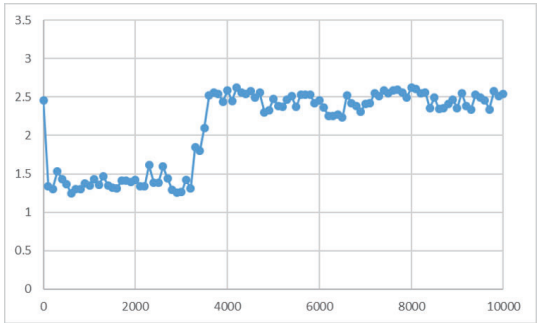

図 9: 突然変異なし，$\alpha = 0.5$ のある試行での 100 世代ごとの 1 ゲームあたり平均利得 (横軸は世代数，縦軸は平均利得を表す．)

6. 考察

相互協調へと進化した突然変異ありの $\alpha = 0.3, 0.4$ では，個体群は早期世代でまず相互裏切りを起こすようなある領域に集まり，そこから相互協調を起こすほうへ進化していくことが分かる．従来手法に比べて相互協調への進化が遅くなったのは，個体群が分散して生成されることで個体同士が集まるのに世代数がかかったことが要因の一つだと考えられる．

また，進化の方向として，初期値から相互裏切りを起こす方へ進化した後，相互協調を起こす方へ進化する場合の他に，相互裏切りを起こす領域内を転々としてから相互協調へと進化していく場合や，一つの場所へ収束することはなく相互裏切りを起こす領域内を転々とし続ける場合が結果として得られたこと，そして突然変異を適用しないと相互裏切りで値が収束してしまうことから，効用導出関数の係数は，相互協調を起こす一定方向に進化するとは限らず，相互裏切りから抜け出す条件が存在することが考えられる．また，それには突然変異が関係していると考えられる．

7. まとめ

実数値 GA による交叉手法に BLX-α を用いることで，従来に比べ個体が分散して生成されるようになり，充分な探索を行うことができた．また，それにより相互協調へ進化するまでの世代数は増加したものの，進化の方向をより詳しく見ることができた．さらに，突然変異が相互協調へ進化するために重要な役割を持っている可能性があるということと，BLX-α にも突然変異を併用しなければ局所解に陥ってしまう問題が存在することが分かった．

今後の課題として，相互協調へ進化するために突然変異が行っている役割についての議論や，相互協調を起こした後の進化の方向についての検証などが挙げられる．

参考文献

[1] R.S. Sutton and A.G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, 1998 (三上貞芳, 皆川雅章訳, 強化学習, 森北出版, 2000).

[2] K. Moriyama, Utility based Q-learning to facilitate cooperation in Prisoner's Dilemma games, *Web Intelligence and Agent Systems*, **7**(3):233–242, 2009.

[3] C.J.C.H. Watkins and P. Dayan. Technical Note: Q-Learning. *Machine Learning*, **8**:279–292, 1992.

[4] K. Moriyama et al., Evolving Subjective Utilities: Prisoner's Dilemma Game Examples, *Proc. 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 233–240, 2011.

[5] W. Poundstone, Prisoner's Dilemma, Doubleday, 1992. (松浦俊輔他訳, 囚人のジレンマ, 青土社, 1995.)

[6] 川上浩司 他: 進化技術ハンドブック 第 I 巻 基礎編, 近代科学社, 2010.

[7] M. Miyawaki et al., Evolution Direction of Reward Appraisal in Reinforcement Learning Agents. *Proc. 12th KES International Conference on Agent and Multiagent Systems: Technologies and Applications (KES-AMSTA)*, pp. 13–22, 2018.

[8] L.J. Eshleman and J.D. Schaffer, Real-Coded Genetic Algorithms and Interval-Schemata, *Foundations of Genetic Algorithms*, **2**:187–202, 1993.