人　工　知　能　99−11
（１９９５．　３．　７）

確率学習機構を有する遺伝的アルゴリズムの集団対戦型ゲームへの適用

冨川 裕樹　棟朝 雅晴　高井 昌彰　佐藤 義治
北海道大学工学部

本論文では、集団対集団で対戦を行う集団対戦型ゲームの戦略学習を議論する。集団対戦型のゲームにおいては、集団内の個々のエージェントがそれぞれ別々の戦略を取る場合に、それらを組み合わせた集団全体としての可能な戦略の組合せが非常に多くなる。

そこで、我々は確率学習による適合度評価を行う遺伝的アルゴリズム StGA (Stochastic Genetic Algorithm) を用いることで効率的な学習の実現を試みる。StGA では、可能な全戦略の中から少数の戦略をサンプリングし、それに対して確率学習および遺伝的操作を適用する。

シミュレーションによる比較実験を通して、StGA の集団対戦型ゲームにおける学習手法としての有効性を検証する。

An application of a stochastic genetic algorithm to strategy acquisition in games played by groups

Yuki Tomikawa, Masaharu Munetomo, Yoshiaki Takai, and Yoshiharu Sato
Faculty of Engineering, Hokkaido University

In this paper we discuss strategy acquisition in games played by the groups of agents. In such games, we have a huge number of possible strategies because each agent in a group could take a different strategy.

We employ StGA(a Stochastic Genetic Algorithm) which evaluates fitness values by using a stochastic learning automaton in order to realize effective learning in stochastic environments. The StGA samples a small number of strategies from all possible ones and applies stochastic learning and genetic operations to the sampled strategies.

Through simulation experiments, we show the effectiveness of the StGA in the strategy acquisition.

1 はじめに

本論文では、逐次的に適合度評価を行うことで確率的な環境に適応する遺伝的アルゴリズム StGA(Stochastic Genetic Algorithm)[1] を、集団対集団で対戦を行うゲームにおける戦略の学習に適用する。

強化学習手法の一つである確率学習オートマトン (Stochastic Learning Automata, SLA)[4] は、可能な戦略の数が非常に多い場合には収束が著しく遅くなる。StGA では状態空間を遺伝的アルゴリズム (Genetic Algorithms, 以下 GA と略す) の個体の形にコーディングし、状態空間からサンプリングを行い、その部分集団に対して遺伝的操作を適用することにより、問題に適応した形で状態空間の圧縮が行われる。

比較的単純なルールを持つ集団対集団の対戦ゲームを設定し、集団単位での戦略の学習に StGA を適用する。集団対集団で対戦を行うゲームにおいて、集団内の個々のエージェントがそれぞれ別の戦略を取る場合、それらを組み合わせた集団全体としての可能な戦略の組合せは非常に多くなる。このような戦略の組合せの数が非常に多いゲームを用いた実験を通して、StGA のゲームにおける学習手法としての有効性を検討する。

2 StGA の概要

StGA の概要を図 1 に示す。枠組として GA の個体集団と学習の対象である環境が与えられ、個体集団は環境への入力である行動 (Action) を文字列として符号化したものから構成されている。各個体に対して適合度値が与えられ、実際に行動がなされる時、その個体が集団から選択される確率として定義される。つまり、適合度の値を p_i とすると、$\sum_{i=1}^{r} p_i = 1$ が成立する (rは集団内の個体数)。

環境からの出力はその行動が成功したか、失敗したかの二値で示される。行動の結果を用いて、集団内の適合度値の評価がなされる。適合度の評価においては、SLA における linear reward-penalty scheme (L_{R-P})[4] を用いる。

$\vec{p}(n) = (p_1(n), p_2(n), \cdots, p_r(n))$ を時刻 n (この場合の時刻は、過去に行われた行動の数により決定される) における、集団内の適合度値からなるベクトルとする。選択された個体の番号が i であるときに、

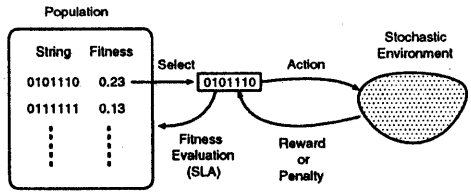


図 1: StGA モデル

それに基づいた行動の成功・失敗により $\vec{p}(n)$ から、$\vec{p}(n+1)$ が以下の式に従って求められる。

$$p_i(n+1) = p_i(n) + \sum_{i \neq j}^{r} f_j(\vec{p}(n)),$$
$$p_j(n+1) = p_j(n) - f_j(\vec{p}(n)) \quad (\forall j \neq i),$$
(成功した場合). (1)

$$p_i(n+1) = p_i(n) - \sum_{i \neq j}^{r} g_j(\vec{p}(n)),$$
$$p_j(n+1) = p_j(n) + g_j(\vec{p}(n)) \quad (\forall j \neq i),$$
(失敗した場合). (2)

L_{R-P}では、f_j と g_jが以下に示される線形関数となる。

$$f_j(\vec{p}) = ap_j, \quad g_j(\vec{p}) = b/(r-1) - bp_j, \ (j = 1, \cdots, r). \tag{3}$$

ここで a, b は、$0 < a < 1, 0 < b < 1$ を満たす定数である。

個体 i に基づく行動が成功した場合には、適合度 p_iの値を増加させ、その他の個体に対する p_j $(j \neq i)$ の値を一定割合 a だけ減少させる。また、失敗した場合には p_iの値を減少させ、その他の個体の適合度値 p_j $(j \neq i)$ を増加させるとともに平均化する ($b = 1$ の場合には、直ちに一様分布となる)。

行動が失敗した場合には、集団に対して遺伝的操作である交叉・突然変異を一定確率で適用することで状態空間内の探索を行い、行動の成功確率を向上させる。

StGA は以下の手順により実行される。

1. 集団の初期化を行う。初期個体は互いに重複の無いようにランダムに生成され、初期適合度値

は $p_i = 1/r$ (r : 集団内の個体数) として与えられる。

2. 集団内から一つの個体を適合度値に応じた確率で選択する。
3. 選択された個体に従った行動が環境に対して適用される。
4. 行動の結果が成功・失敗の形で環境から戻ってくる。
5. 行動の成否の結果を用いて、集団に含まれるすべての個体の適合度値を再評価する。
6. 行動が失敗した場合、一定確率で遺伝的操作である交叉・突然変異を集団に対して適用する。交叉・突然変異が適用された場合、以下の処理 (a), (b), (c) を行う。遺伝的操作の結果、個体の重複が発生した場合には、突然変異を繰り返すことで、重複を除去する。
   (a) 遺伝的操作により生成された個体を集団内で最も適合度値の低い個体と置き換える。
   (b) 新たに生成された個体の適合度値は、その親である個体の適合度値をそのまま継承する。
   (c) 適合度値の継承により、集団内における適合度値の総和が、$\sum_{i=1}^{r} p_i \neq 1$ となることがある。その場合には、$p_i \leftarrow p_i / \sum_{j=1}^{r} p_j$ による適合度値の修正を行い、総和が 1 となるようにする。

3 ゲームの設定

比較的単純なルールを持つ、集団対集団の対戦ゲームを設定し、集団単位での戦略の学習に StGA を適用する。

3.1 ゲームの枠組

対象とするゲームは、平面上をエージェントが移動しながら敵を見つけ、攻撃するゲームである。エージェントの動作は確率的であるが、その確率分布は採用する戦略に従う。エージェントは集団を構成し、集団対集団で対戦を行う。このゲームのルールは以下のように定義される。

- ゲームが行われるのは 40 × 40 の格子状に区切られた平面上である。平面の外周は壁になっていて、エージェントはその範囲から出ることはできない。また、エージェントは方位として東西南北 (E, W, S, N) とそれらの中間の方向 (NE, NW, SE, SW) をとり得る。
- エージェント a_i (以下、戦車と呼ぶ) は 5 台でチーム T を構成する。

$$T = \{a_i | 1 \leq i \leq 5\}$$

  対戦は 2 つのチームの間で行われる。
- 戦車 a_i の戦略 $S(a_i)$ は 4 つのパラメータ $\{at, m1, m2, m3\}$ で表現される。at は攻撃に関するパラメータで、$m1, m2, m3$ は移動に関するパラメータである。これらについては 3.3節で説明する。
- チーム T の戦略 $S(T)$ は、チーム T 内の各戦車の戦略からなる集合である。

$$S(T) = \{S(a_i) | a_i \in T\}$$

- 戦車は各ステップにおいて、まず移動に関するパラメータに従って確率的に移動する。次に攻撃のパラメータに従って条件を満たした場合は確率的に攻撃を行う。
- 戦車は攻撃が 3 回命中すると消滅する。
- どちらかのチームの戦車が全滅するか、あるいは一定ステップ数が経過するまでを 1 回の対戦とする。両チームとも同時に全滅した場合は引き分けとする。

3.2 戦車の能力

戦車の能力を次のように設定した。以下、味方とは自分と同じチームに属する戦車を意味し、自分と異なるチームの戦車を敵と呼ぶ。

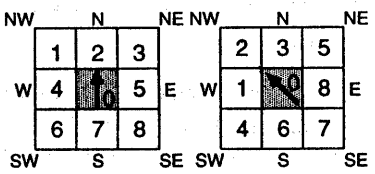


図 2: 戦車の向きと移動方向

3.2.1 移動に関する能力

戦車は現在地の隣接 9 マス (現在いるマスも含む) のいずれかへ移動できる (図 2)。移動に関する戦略は、「基準方向」(現在の戦車の向き、あるいは他の戦車のいる方向) と移動先から定められる「確率分布」に従って実際の行動を決定するものである。図 2 の左の図は、現在の基準方向が 北 (N) 向きの場合を表していて、移動方向は図のように 0 ～ 8 まで番号付けられる。この移動方向の番号は基準方向に対して相対的に定められる。基準方向としては 8 方位のみ考え、それより細かい方向は 8 方位の中でより近いものに帰着させることとする。

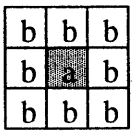
可能な移動戦略は、移動先への確率分布としてあらかじめ表 1 のように定められている。例えば移動の戦略として p1 をとった時、「3」 の方向 (基準方向の右前方) へ移動する確率は 10 % である。そして「3」へ移動した場合、移動後の戦車の向きは北東 (NE) となる。

表 1: 移動方向の確率分布

	0	1	2	3	4	5	6	7	8
p1	0 (%)	10	70	10	0	0	5	0	5
p2	0	10	50	10	5	5	10	0	10
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

3.2.2 攻撃に関する能力

攻撃は、攻撃目標となるマス目に対して大砲を撃つことにより行われる。大砲の撃ち込まれたマス目にいる戦車は、敵・味方に関わりなくダメージを受ける。大砲の命中率は、攻撃目標との距離により異なる。図 3 で a が攻撃目標であり、距離が遠くなるほど命中率が悪くなり、攻撃がそれてしまう (b のマス目に当たる) 確率が高くなる。



距離	a	b
0 ～ 2	90 (%)	1.25 (%)
3 ～ 4	50	6.25
5 ～ 6	18	10.25
7 ～ 8	12	11

図 3: 距離と大砲の命中する確率

3.2.3 センサーの能力

戦車は近くに敵や味方がいるかどうかを知るためのセンサーを持っている。センサーで感知できる範囲は、現在いるマス目から距離 8 マス以内の領域である。各戦車はセンサーで感知した結果により、自分が持つ移動に関する 3 つのパラメータ $\{m1, m2, m3\}$ のうちどれに従って行動するかを決定し、また攻撃に関するパラメータ at に従ってそのステップで攻撃するかどうかを決定する。

可能な攻撃戦略は、ある距離にいる敵を攻撃する確率の分布として表 2のように定められている。例えば攻撃の戦略として $p1$ をとった時、距離 3 ～ 4 に位置する敵を攻撃する確率は 20 % である。

表 2: 攻撃の確率分布

	0 ～ 2	3 ～ 4	5 ～ 6	7 ～ 8
p1	70 (%)	20	10	0
p2	10	65	20	5
⋮	⋮	⋮	⋮	⋮

3.3 戦略の表現形式

戦車の戦略を表す 4 つのパラメータ $\{at, m1, m2, m3\}$ についてここでまとめておく。

1. 攻撃 (at) : センサーの範囲で一番近い距離にいる戦車が敵である場合に、その敵との距離によって攻撃するかどうかを判断する確率分布を決定する。

2. 移動 1 ($m1$) : センサーの範囲に敵も味方もいない場合における、移動方向の確率分布を決定する。この時、基準方向は戦車の現在の方向。

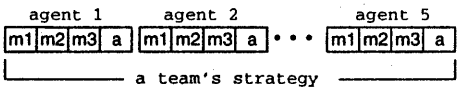


図 4: チームの戦略

3. 移動 2 ($m2$) : センサーの範囲で一番近い距離にいる戦車が敵である場合における、移動方向の確率分布を決定する。基準方向はその敵のいる方向。

4. 移動 3 ($m3$) : センサーの範囲で一番近い距離にいる戦車が味方である場合における、移動方向の確率分布を決定する。基準方向はその味方のいる方向。ただし、戦車にはチーム内で一意に順序を定める ID 番号が付けられており、一番近くにいる味方の ID 番号が自分より小さい場合には $m3$ には従わず、$m1$ に従って移動する。

今回の実験では、各パラメータともに可能な戦略の種類を 16 とした。これらはあらかじめ与えられており、個々の分布そのものは変化しない。このとき、戦車の可能な戦略の組合せは $2^{16} = 65536$ 通りである。

チームの戦略は、そのチーム内の各戦車の戦略の集合である（図 4）。1 チームは戦車 5 台なので、チームの戦略の可能な組合せは $2^{16\times5} = 2^{80}$ 通りである。チーム（戦車）は、その中から最適な戦略の組合せを学習する。

4 StGA による学習

StGA では、チームの戦略は二進表現のストリングで内部表現される。可能な全戦略の中から一部はサンプリングされ、それぞれのストリングに適合度値が与えられる。

1 回の対戦の結果により、チームが勝利した場合を成功、チームが負けた場合を失敗とし、(1)(2) 式に従って適合度値の更新を行う。引き分けの場合は、(3) 式のパラメータ a の値として勝利した場合の半分の値を用いる。

繰り返し対戦を行う中で、勝利する可能性の高い戦略を得ることが学習の目的である。

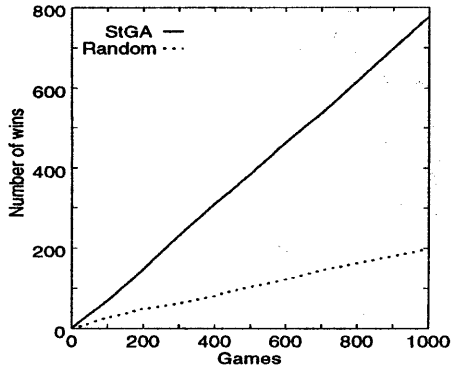


図 5: StGA と 100 回対戦ごとにランダムに戦略を変えるチームとの対戦結果

5 実験

敵チームとして、(1) 100 回対戦するごとにランダムに戦略を変えるチーム、(2) SLA により学習を行うチームを構成し、StGA による学習を行うチームと対戦する実験を行った。

5.1 実験条件

StGA で学習を行うチームについて、式 (3) のパラメータ a, b の値をそれぞれ 0.15, 0.12 とした。また可能な全戦略からサンプリングする戦略数を 20 とした。遺伝的操作である交叉・突然変異は対戦に負けた場合にともに 20 %の確率で行われるものとした。

5.1.1 ランダムな戦略をとるチームとの対戦

100 回対戦するごとにランダムに戦略を変えるチームと StGA による学習を行うチームとの対戦結果を図 5 に示す。横軸は対戦回数、縦軸は累積勝利数であり、値は 1000 回対戦を行う実験を 10 回行った結果の平均である。

対戦の早い段階から StGA により学習を行うチームの勝利数が相手チームを大きく上回っており、100 回ごとに相手がランダムに戦略を変化させてもそれに即座に対応していると考えられる。

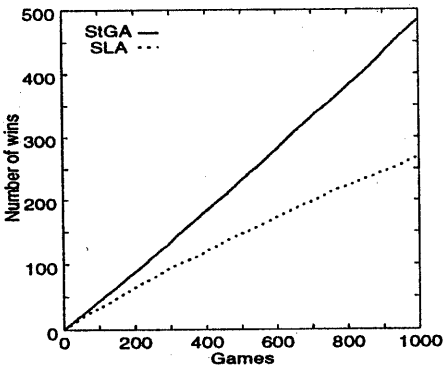


図 6: StGA と SLA の対戦結果

5.1.2 SLA で学習を行うチームとの対戦

敵チームとして SLA により学習を行うチームを用い、StGA による学習を行うチームとの対戦結果を図 6 に示す。

SLA を今回のゲームに適用する場合、単純に 2^{80} 通りある可能な全戦略に適合度値を与えるというインプリメントを行うと、適合度の初期値が $1/2^{80}$ と小さい値になるため学習に非常に時間がかかると予想される。また 1 回の対戦を行うごとに 2^{80} 個のすべての適合度値を更新しなければならないために、更新に要する計算量は膨大である。

そこでチーム全体の戦略に適合度値を与えるのではなく、各戦車ごとで適合度を与えることとし、戦車が対戦において最後まで残ったかどうかではなく、自分の属するチームが勝利したかどうかで適合度値の更新を行うこととした。

条件をそろえるために、StGA で学習を行うチームについても同様に戦車単位で適合度値の更新を行うように設定した。

図 6 より、最初の 100 回目くらいの対戦までは両チームの勝利数にそれほど差はなく、ともにまだ戦略が定まらない状態であると考えられる。その後徐々に StGA チームの勝利数が上回るようになり、少しずつ差が広がっている。

6 おわりに

確率学習による適合度評価機構を有する遺伝的アルゴリズム StGA の応用として、集団対集団で対戦を行うゲームにおける戦略獲得への適用を試みた。可能な戦略の数が非常に多い場合に収束が非常に遅くなるという SLA の問題点を、StGA は可能な全戦略の中からサンプリングを行い、その部分集団に対して SLA を適用し、それらに対して遺伝的操作を行うことで解決を試みている。

本論文では、集団対集団で対戦を行う比較的単純なルールを持つゲームを設定し、集団全体としての戦略の学習に StGA を適用した。敵チームとして一定回数対戦するごとにランダムに戦略を変えるチームと SLA により学習を行うチームを構成し、それらとの対戦を行うシミュレーション実験を通して、StGA が戦略の学習手段として有効であることを示した。

今後対戦相手として SLA 以外の学習アルゴリズムを用いた場合との比較実験を予定している。

参考文献

[1] 棟朝雅晴, 高井昌彰, 佐藤義治: "確率学習による適合度評価機構を有する遺伝的アルゴリズム (1) – 基本モデル –", 情報処理学会第 49 回全国大会講演論文集 (2), pp.231–232 (1994).

[2] 冨川裕樹, 棟朝雅晴, 高井昌彰, 佐藤義治: "確率学習による適合度評価機構を有する遺伝的アルゴリズム (2) – 戦略獲得への応用 –", 情報処理学会第 49 回全国大会講演論文集 (2), pp.233–234 (1994).

[3] D. E. Goldberg: *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison Wesley (1989).

[4] K. S. Narendra and M. A. L. Thathachar: "Learning automata – a survey", *IEEE Transactions on System, Man, and Cybernetics*, Vol. 4, No. 4, pp.323-334 (1974).