北海道大学工学部研究報告
第172号（平成7年）

Bulletin of the Faculty of Engineering
Hokkaido University, No. 172 (1995)

確率学習機構を有する遺伝的アルゴリズムの戦略獲得への応用

冨川裕樹　棟朝雅晴
高井昌彰　佐藤義治

（平成6年10月28日受理）

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

Yuki TOMIKAWA, Masaharu MUNETOMO, Yoshiaki TAKAI, and Yoshiharu SATO

(Received October 28, 1994)

Abstract

A stochastic genetic algorithm (StGA) realizes adaptive learning in stochastic environments. A stochastic learning automaton (SLA) is used for fitness evaluation of the genetic algorithm. The learning process of the StGA converges faster than that of the SLA because the genetic operators are applied to a sampled, relatively small sized population.

We apply the StGA to strategy acquisition of games defined by payoff matrices. Through simulation studies which compare a learning scheme by StGA and that by SLA, we show the effectiveness of our scheme.

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，利得行列で定義されたゲームの戦略獲得に応用する。

従来の遺伝的アルゴリズム(Genetic Algorithms，以下GAと略す)[3]では，正確な適合度値が必要なときに必要な数だけ求められることを暗黙の前提としている。しかし，ゲームの戦略獲得においては，相手の戦略に利得値が依存するため，得られた値そのものを正確な適合度値と見なすことはできず，ある確率分布に従った値であると考えるのが妥当である。また，利得値が連続であるとは限らず，極端な場合には戦略の成功・失敗の二値しか得られない場合もある。このような時には，利得値をそのまま適合度値として用いることはできず，過去の成功・失敗の割合を何らかの形で保存しなくては学習が行えない。

StGAではGAの適合度評価に確率学習オートマトン(Stochastic Learning Automata, SLA)[4]を用いることより，確率的な環境において適切な適合度の分布を個体集団内に生成する。本論文では，利得行列の形で表現されるゲームにおいて，StGAとSLAのそれぞれを用いた学習アルゴリズムを対戦させる比較実験を行い，取り得る戦略の数が多い場合におけるStGAの有効性ならびに環境変化への追従性を検証する。

情報工学科　情報図形科学講座

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。集団内における個体の重複は直接検出され，重複個体は除去される。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)。

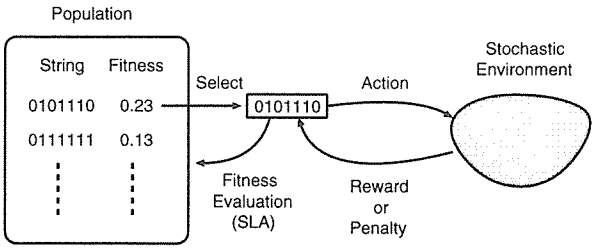

図1　StGA の概要

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。

$\vec{p}(n) = (p_1(n), p_2(n), \ldots, p_r(n))$ 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，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。

$$\begin{aligned} p_i(n+1) &= p_i(n) + \sum_{i \neq j}^{r} f_j(\vec{p}(n)), \\ p_j(n+1) &= p_j(n) - f_j(\vec{p}(n)) \quad (\forall j \neq i), \end{aligned}$$

(成功した場合)．　(1)

$$\begin{aligned} p_i(n+1) &= p_i(n) - \sum_{i \neq j}^{r} g_j(\vec{p}(n)), \\ p_j(n+1) &= p_j(n) + g_j(\vec{p}(n)) \quad (\forall j \neq i), \end{aligned}$$

(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。

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

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．対象とするゲームについて

本論文で対象とするゲームは将棋などの具体的なゲームをモデル化したものではなく，ゲーム理論における二人のプレイヤーによる行列ゲームである。ゲームのルールを以下のように定式化する[2)]。

・プレイヤー $p1$，$p2$の取り得る戦略の種類はそれぞれ n_{p1}，n_{p2} 通りある。
・$p1$が戦略 i，$p2$が戦略 j を選択した時，利得として $p1$へ $a^{p1}_{i,j}$，$p2$へ $a^{p2}_{i,j}$ を与える行列

$$A^{p1}=[a^{p1}_{i,j}],\quad A^{p2}=[a^{p2}_{i,j}] \qquad (0 \leqq i < n_{p1},\ 0 \leqq j < n_{p2})$$

を利得行列として定義する。
・利得行列 A^{p1}，A^{p2} の内容はプレイヤーからは不可視である。
・先手，後手は無し。
・繰り返し対戦を行い，利得の累積値を相手プレイヤーより多くすることを目的とする。

ここで，二人のプレイヤーが得る利得値の和がかならずしもゼロ（ゼロ和）である必要はないが，以下の実験ではゼロ和となる利得行列を用いた。

4．実験による比較

プレイヤー $p1$を StGA により学習するプレイヤー，プレイヤー $p2$を SLA により学習するプレイヤーとして対戦をする実験を行った。

4.1　実験条件

プレイヤーの取り得る戦略の種類を $n=n_{p1}=n_{p2}=1024$通りとした。StGA においては戦略の番号の二進表現(10ビット）による文字列を個体とし，集団内の個体数を20とした。遺伝的操作である交叉・突然変異は，ともに行動が失敗した時に20％の確率で行われるものとし，交叉としては一点交叉を用いた。式3のパラメータとして，StGA は $a=0.18$，$b=0.12$，SLA は $a=0.15$，$b=0.08$とした。双方のプレイヤーとも，一回の対戦において正の利得を得た場合にその時取った戦略は成功し，それ以外の場合には失敗したと判断する。

4.2 利得行列が固定の場合

利得行列の内容が変化しない場合について実験を行った。利得行列として $A^{p1}=-A^{p2}$ である図2のような行列を用いた。この行列は，$n/2$ 番目の戦略を取った時に勝つ（正の利得を得る）割合が最も大きくなっているが，この場合でも勝つ可能性は約50%になっている。このため，この $n/2$ 番目の戦略を取り続けたとしても勝ち続けることはできない。しかし，$n/2$ 番目付近の戦略をとることが有利であることは明らかであり，$n/2$ 番目付近の戦略を取りつつ相手プレイヤーの戦略の変化に速やかに対応するプレイヤーの方が対戦が進むにつれて有利となると予想される。

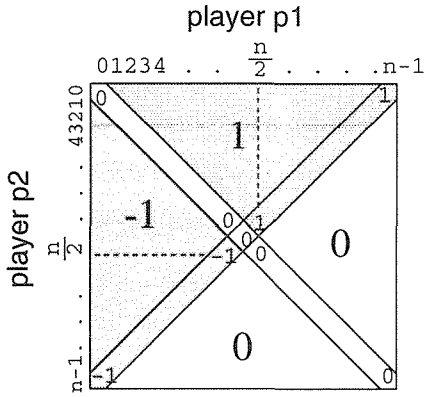

図2 利得行列 $A^{p1}=-A^{p2}$ の内容

本実験におけるプレイヤー $p1$(StGA)が得た利得の累積値の推移を図3に示す。連続10000回の対戦実験を10回行った結果である。グラフの横軸は対戦回数で，縦軸は累積利得値である。ここでは，利得行列を $A^{p1}=-A^{p2}$ としたためにゼロ和ゲームとなっているので，$p2$に関するグラフは $p1$に関するグラフと横軸に関して対称となる。

図3より，10回の実験の中で9回はStGAの累積利得値が増加しており，SLAに比べて有利となっている。勝っている9回を詳細に見ると，対戦の開始から単調に勝ち続けているものはほとん

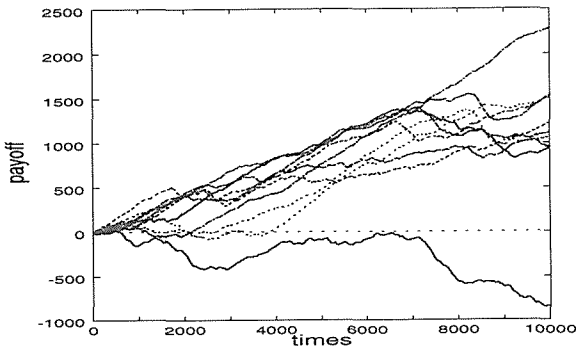

図3 StGAの利得の累積値の推移

ど見られず，初期の均衡状態から次第に勝ち続けるようになったり，勝ち続けている状態から均衡状態や負ける状態に遷移することが多い。これは今回用いた利得行列の性質上，唯一最適な戦略というものが存在しないことに起因する。すなわち，同じ戦略で勝ち続ける状況になっても，その間に相手プレイヤーがその戦略に勝る戦略を発見すると，同じ戦略を取り続けたのでは勝てなくなるためと考えられる。

次に，図3に示した10回の実験のうち $p1$ が勝っている1回に着目し，それぞれのプレイヤーが各対戦において選択した戦略の推移を，StGA については図4に，SLA については図5に示す。

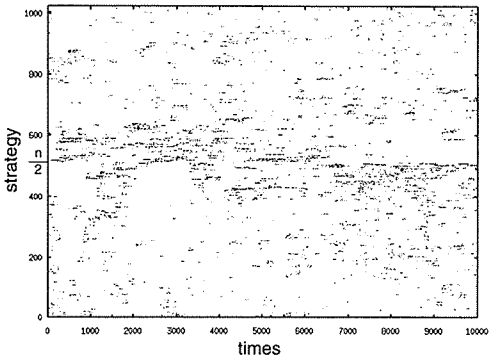

図4 StGA の戦略の推移（利得行列固定）

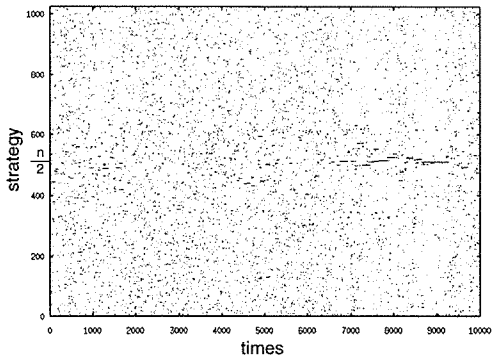

図5 SLA の戦略の推移（利得行列固定）

StGA では $n/2$ 番目に近い比較的成功率の高い戦略を多く選択している様子が確認できる。これは，StGA が可能な戦略からなる状態空間を個体集団の形でサンプリングしているため，失敗した場合でも適合度値が集団内の個体にしか分散されないためであると考えられる。一方 SLA の方は，戦略が全域に散らばっており，戦略にほとんど偏りが見られない。これは，SLA によるプレイヤーが StGA プレイヤーに敗けた場合，適合度値が状態空間全域に分散されてしまうためであると考えられる。

また，図5の7000～9000回目の対戦では，SLA が一定の戦略を取っている部分があり，SLA のプレイヤーが StGA プレイヤーに勝っている。上で述べたように，SLA では失敗した場合に適合度値が戦略全域に分散される。逆に勝ち続けている場合にはその戦略のみの適合度値が高くなるので一つの戦略が多く選択される。一方 StGA は同じ7000～9000回目の対戦において，$n/2$ 番目の戦略に近い戦略の内，$n/2$ 番目より小さい番号に偏った戦略をとっている。これは，戦略番号 $n/2 = 512 = (1000000000)_2$ と戦略番号 $n/2 - 1 = 511 = (0111111111)_2$ のハミング距離が非常に遠いためである。この問題は戦略番号をグレイコードで符号化することで解決できると考えられる。

4.3 利得行列が変化する場合

前節と同様の利得行列を用い，一定回数対戦を行うごとに行列の対角方向に沿って利得行列の各要素を平行移動する実験を行った。平行移動によってはみ出した部分は反対側にループして戻ってくるものとする。

利得行列の内容をこのように変化させた場合，はじめの実験で比較的良いと考えられる $n/2$ 番目の戦略の位置が，ゼロ和であるという性質を変えずに移動することになる。m 戦略分平行移動した場合の利得行列の内容を図6に示す。

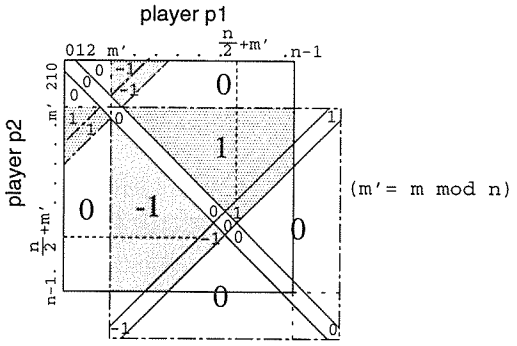

図6　m 戦略分平行移動した時の利得行列の内容

以下では利得行列の変化の仕方が異なる四つの実験を行った。

(I)　10回の対戦ごとに戦略一つ分だけ移動

(II-1) 10回の対戦ごとに $n/2$戦略分移動

(II-2) 100回の対戦ごとに $n/2$戦略分移動

(II-3) 1000回の対戦ごとに $n/2$戦略分移動

(I)は緩やかな変化が起きる場合，(II)は大きな変化が一定回数の対戦ごとに起きる場合を想定している。戦略の総数は $n=1024$である。

(I)　10回の対戦ごとに戦略一つ分だけ移動

連続10000回の対戦実験を行ったときのStGA, SLA それぞれのプレイヤーの戦略の推移を図7，図8に示す。図7より，StGAは利得行列の変化に十分追従した戦略をとっている。一方SLAは図8より，一部で短期間同一の戦略を取っているものの，ほとんど戦略に偏りが見られない。ただし，利得行列の内容が変化しない場合でもSLAは同様な傾向を示しているので，図8の結果のみではSLAの環境追従性の判断はできない。

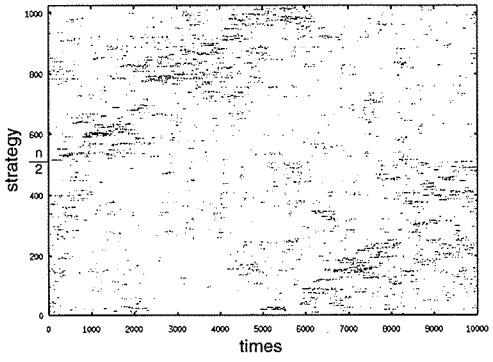

図7　StGA の戦略の推移（10回の対戦ごとに戦略一つ分だけ移動）

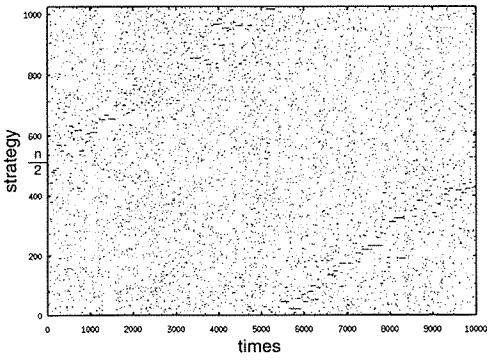

図8　SLA の戦略の推移（10回の対戦ごとに戦略一つ分だけ移動）

(II-1) 10回の対戦ごとに $n/2$戦略分移動

(II)では，実験結果として連続10000回の対戦実験を10回行った時のプレイヤー p1(StGA)の累積利得値の推移（図9，10，11）を考察する。

図9の結果は，学習アルゴリズムを用いずランダムな戦略をとるプレイヤー同士の対戦結果と同じ傾向を示している。この場合，StGA, SLA どちらのプレイヤーも学習が進む前に環境が大きく変化してしまうために学習の効果が現れず，結果的にほとんどランダムな戦略を取っていると考えられる。

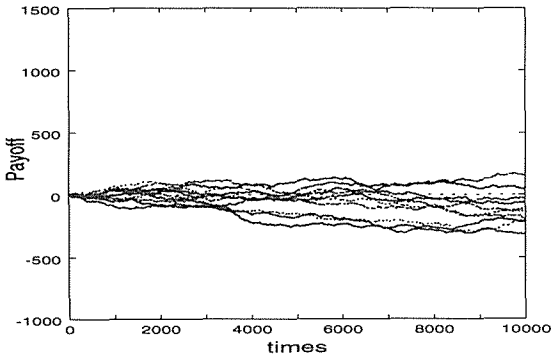

図9　StGA の利得の累積値の推移（利得行列を10回の対戦ごとに $n/2$戦略分移動）

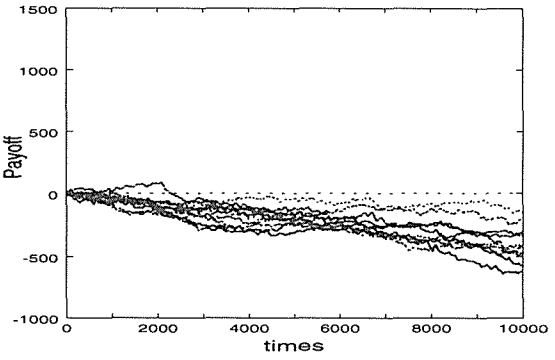

図10　StGA の利得の累積値の推移（利得行列を100回の対戦ごとに $n/2$戦略分移動）

(II-2) 100回の対戦ごとに $n/2$戦略分移動

図10に示された結果より，10000回の対戦後の累積利得値の大きさは，最大のものでも500程度であることがわかる。この場合も100回という比較的短い対戦間隔で大きな環境変化が起きるために両プレイヤーともランダムに近い戦略を取っていると考えられる。しかし全体的に見てみると，SLA の方が StGA よりわずかではあるが勝っている。これは，StGA の場合サンプリングされた個体群が状態空間を広く探索する前に環境が大きく変化してしまうために変化に対応しきれず，全戦略を対等に扱う SLA の方が有利となっているためと考えられる。つまり，大きな環境変化が短い間隔で連続的に起きる状況においては，StGA の追従性が SLA よりも悪くなる場合もあり得ることをこの実験結果は示している。

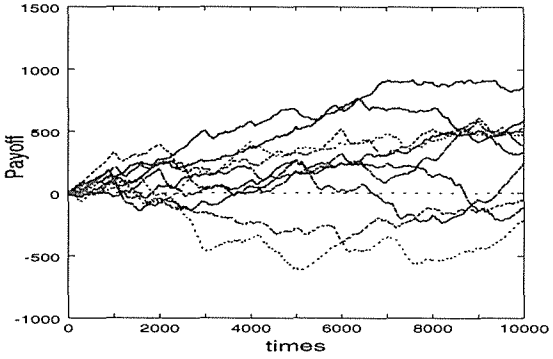

図11　StGA の利得の累積値の推移（利得行列を1000回の対戦ごとに $n/2$戦略分移動）

(II-3) 1000回の対戦ごとに $n/2$戦略分移動

図11の結果から，大きな環境変化が一定回数の対戦ごとに起きる状況でも，状態空間を十分探索するだけ対戦が行われる場合にはStGAの学習効果が有効となり，StGAのプレイヤーが勝つことが多いことがわかる。StGAのプレイヤーは1000回ごとの大きな変化の直後で一時的に累積利得が減少することもあるが，状態空間の探索が行われて学習が進むにつれ，利得を回復している。

5．おわりに

本論文では，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，ゲームにおける戦略獲得への適用を試みた。StGAは問題に適応する形で状態空間をサンプリングするため，可能な戦略の数が多いゲームにおける戦略の獲得において有効であることをシミュレーション実験を通して確認した。

また，利得行列の内容が変化するゲームを用いて状態空間の変化に対する追従性を調べた結果，状態空間が緩やかに変化する場合にはStGAはその変化に対して良く追従することがわかった。しかし大きな変化が短い間隔で連続して起こる場合では，StGAの追従性はSLAよりも劣ることがわかった。

今後の課題としては，利得行列の内容が対戦の繰り返しの過程でプレイヤーの戦略に応じて次々と変化するような，より複雑なゲームに対する適用や，SLA以外の学習アルゴリズムとStGAとの比較実験などがあげられる。

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，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，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).