3D9-1

強化学習における報酬値探索への GA の適用

Applying GA to Searching Reward Value in Reinforcement Learning

井上 勇気[*1]
Yuhki Inoue

赤塚 洋介[*1]
Yousuke Akatsuka

佐藤 裕二[*2]
Yuji Sato

[*1] 法政大学 情報科学研究科
Graduate School of Computer and Information Science

[*2] 法政大学 情報科学部
Faculty of Computer and Information Science

Recently video game has the problem that is to spend much time on making algorithm of player in video game with growing complexity. One way to solve this problem, the hybrid decision making system (mix the traditional way of making algorism by handmade with classifier system) is proposed. On the other hand, reinforcement learning by classifier system has the problem that reward value setting affects converge speed and quality of learning. But, there is no indicator in the way of deciding the reward value and the reward value is decided by experience. For solution of the problem, this study presents that searching for optimum solution of reward value automatically with GA. By experiment, as a result of using GA for setting of reward value, this paper shows that converge speed become early and there is possibility of adapting reward value to agent role and algorithm of opposition on game.

1. はじめに

従来, テレビゲーム製作におけるゲーム内エージェントの意思決定は, 人手により製作されたアルゴリズムによって実現されてきた. 近年のゲームでは, プレイヤー数の増大, インターネットによるオンライン化などゲーム環境の複雑化に伴い, アルゴリズムの短寿命化が起こっている. アルゴリズムの短寿命化のため, アルゴリズムの開発時間の増大という問題を引き起こしている.

サッカーゲームにおける上記のプレイヤーの意思決定問題を解決する手段として, 強化学習のひとつであるクラシファイアシステム[Goldberg 89, 伊庭 02]を動的な環境に自動的に適応させることで, 問題解決を図ったハイブリッド型意思決定システム[Sato 05]が提案されている. また, クラシファイアシステムの特徴である環境からの報酬のフィードバックに注目し, サッカーゲームの FW, MF, DF の各ポジションを考慮した報酬値設定をおこなうことで, 学習の効率化を図る方法[Sato 06]も提案されている.

一方, クラシファイアシステムでは環境からの報酬に基づいて学習を行っているため, 報酬値の設定が適切でない場合, 効率のよい学習ができず, 学習に多くの時間を費やすことになる. また, クラシファイアシステムにおける報酬値の決定には指標がなく, 経験に基づく試行錯誤で決定しなくてはならないという問題がある.

上記問題を解決する手段として本研究では, クラシファイアシステムを基本とした学習における報酬値の決定に遺伝的アルゴリズム (Genetic Algorithm : GA) [Goldberg 89, 坂和 95]を用いることを提案する. また, GA による報酬値探索の結果, 環境への適応速度が早まる可能性を示す.

2. サッカーゲームの概要

2.1 プレイヤーの設計

ここで扱うサッカーゲームは 1 チーム 11 人で構成され, 対戦は 2 チーム, 計 22 人のプレイヤーによっておこなわれる. 各プレイヤーは認識器, 行動器, 意思決定部の 3 つから構成される. 図 1 にプレイヤーの構成の図を示す. プレイヤーはボールの位置, 各プレイヤーの位置など, 環境の状態を認識器から取得する. 認識器で取得した状態を意思決定部へ送信し, 送信された情報から意思決定部で行動を選択する. 行動器では意思決定部で選択された行動を環境に対して実行する. 行動は移動, パス, ドリブルなどに当たる. 本研究ではプレイヤーの意思決定部に, GA を用いた強化学習の一つであるクラシファイアシステムを用いた場合を考える.

2.2 クラシファイアシステムとは

クラシファイアシステムは, IF (条件部) ―THEN (帰結部) のクラシファイアと呼ばれるルールの集合からなる, ルールベース型意思決定システムの一種である. クラシファイアシステムは, 環境からの情報を認識として取得し, その情報を条件部と照合した後, 対応するクラシファイアの帰結部を環境に対して実行する. 実行した行動の成否が環境から報酬としてフィードバックされ, クラシファイアの信頼度を増減することで強化学習をおこなう. 図 2 にクラシファイアシステムを用いたプレイヤーと環境の関係の概要, 図 3 にクラシファイアシステムを示す. クラシファイアシステムでは, バケツリレー方式[Goldberg 89, 伊庭 02]と呼ばれる学習により, 時系列的な動作の流れを学習することが可能である.

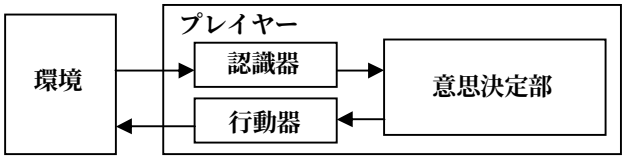

図 1 プレイヤーと環境の関係

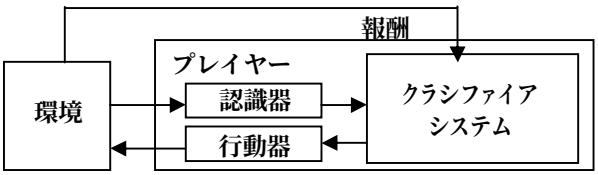

図 2 環境とクラシファイアシステムの関係

連絡先:井上 勇気, 法政大学, 東京都小金井市梶野町 3-7-2, 042-387-6023, yuhki.inoue.z2@gs-cis.hosei.ac.jp

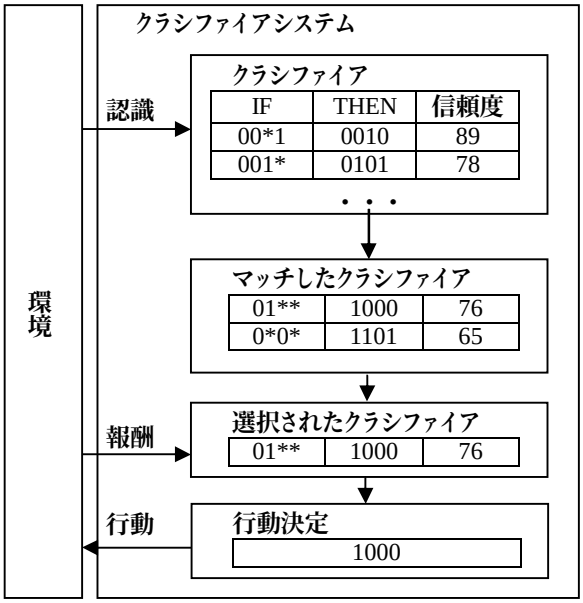

図 3 クラシファイアシステム

2.3 クラシファイアシステムの問題点

クラシファイアシステムの学習は環境からの報酬に頼っているため，環境からの報酬値が適切に設定されていない場合，効率の良い学習ができず，学習に多くの時間がかかる．一方，環境から得られる報酬値の設定には決定指標がなく，学習の目的，エージェントの置かれた環境，エージェントの役割など複数の要素を考慮した上で，学習に有効な値を経験に基づいて試行錯誤で設定する必要があるという問題点が存在する．

3. GA による報酬値の探索の提案

3.1 提案の基本的考え方

前述の問題の対策として，具体的な教師データがなくても，とりあえず強化学習が行えるクラシファイアシステムの特徴に着目して，ここでは報酬値の探索に GA を用いることを提案する．報酬値の決定には決定指標がないため，サッカーゲームの勝率を適応度として GA を実行する．GA を用いて自動的に最適な報酬値を設定できれば，経験に基づく報酬値設定という非効率な作業を省くことができると考えるためである．

3.2 GA の設計

本研究では GA を用いてクラシファイアシステムの報酬値を自動的に探索する．そのシステム構成を図 4 に示す．ゲームに勝つことが最大の目標であるため，高い勝率を示す報酬値を GA で選択，低い勝率を示す報酬値を淘汰することで，より勝率を高くする報酬を探索する．

GA における個体の定義は FW，MF，DF の PASS と DRIBBLE の値をセットとして 1 つの個体とみなす．次の世代の子個体を生成するための親個体の選択にはエリート選択を用いて，交叉方法は一様交叉を用いる．交叉率は 0.6 とした．実験に要する時間を考慮すると，個体数が多い場合非常に多くの時間を必要とするため，個体数を少なくする必要がある．従って，多様性を保つために突然変異率は通常より高く設定する．突然変異率は，予備実験を行い，予備実験の結果から，0.2 に設定する．

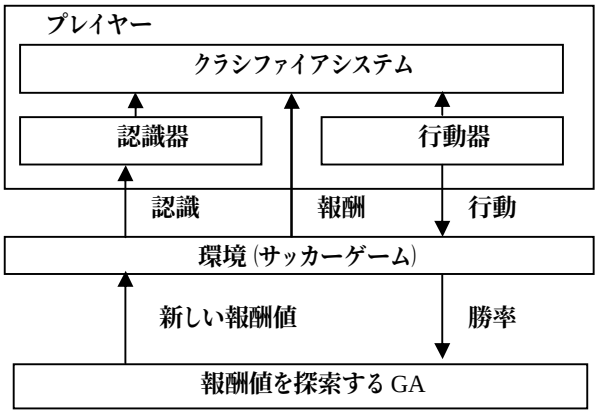

図 4 提案手法の構成図

表 1 報酬セット

FW	MF	DF
PASS の値	PASS の値	PASS の値
DRIBBLE の値	DRIBBLE の値	DRIBBLE の値

3.3 実験方法

FW，MF，DF それぞれの PASS と DRIBBLE の 6 種類の報酬値を探索対象とする．この 6 種類の報酬値を持った報酬セットを 4 種類用意する．表 1 に報酬セットを示す．それぞれの報酬セットの報酬値をランダムに設定する．各報酬セットが個体となる．報酬セット1つにつき 200 試合を 3 回くり返し，その平均勝率を適応度として採用する．

GA の実行に用いる適応度として，平均勝率を用いる．結果として採用する実データは，10 世代ごとに 200×3 回の平均勝率が一番よかった報酬セットを使い，200 試合を 10 回繰り返し，取得したデータの平均を採用する．

対戦アルゴリズムとして，攻守のバランスが取れたアルゴリズム ALG_A，攻撃に重点を置いたアルゴリズム ALG_B，守備に重点を置いたアルゴリズム ALG_C の三種類を用意する．学習させるプレイヤーの意思決定部にはハイブリッド型意思決定システムを用いる．

4. 実験結果

4.1 アルゴリズム A との対戦結果

攻撃と守備のバランスをとったアルゴリズムである，アルゴリズム A（ALG_A）と対戦した結果を図 5 に示す．初期世代は 40 試合程度で 70%の勝率に収束しており，その後の勝率の上昇は見られない．10 世代，20 世代の勝率は 80%弱を示している．30 世代を見てみると，40 試合程度で 80%近くに達しており，その後も安定して 80%程度の勝率を示している．20 試合から，50 試合の間を見れば，30 世代の勝率が，早い段階で収束している．

4.2 アルゴリズム B との対戦結果

攻撃に重点を置いたアルゴリズムである，アルゴリズム B（ALG_B）と対戦した結果を図 6 に示す．初期世代の勝率推移を見ると，30 試合程度までは勝率が 70%程度まで伸びて，その後下がっていき，63%程度に収束している．10 世代の勝率の推移を見てみると，初期世代よりは高い勝率を示したものの，50 試合程度まで安定していない．20 世代，30 世代の勝率推移では，

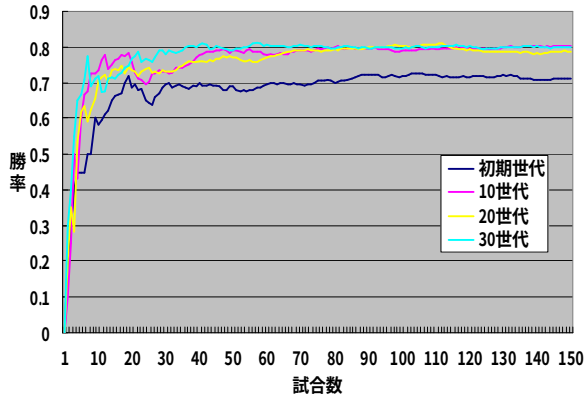

図 5 対 ALG_A における世代と勝率の関係

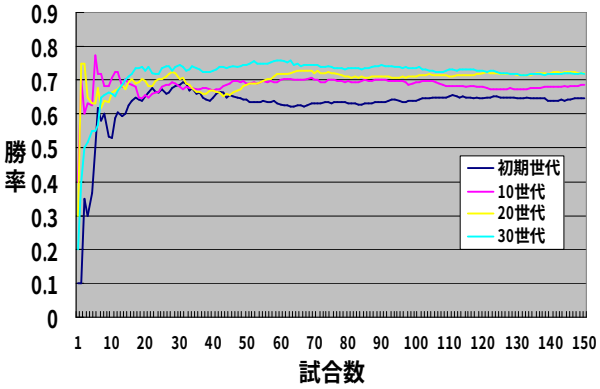

図 6 対 ALG_B における世代と勝率の関係

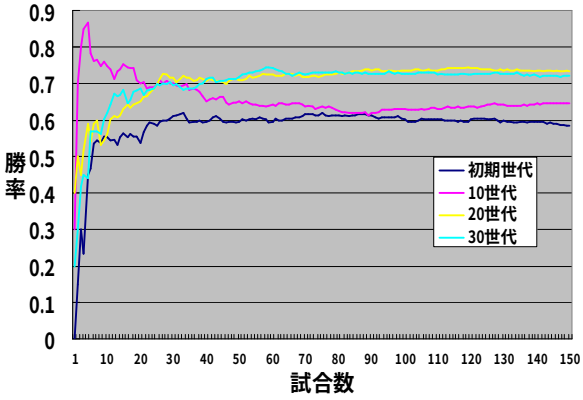

図 7 対 ALG_C における世代と勝率の関係

最終的な勝率は 20 世代, 30 世代は同等の勝率を示しているが, 20 世代は 70 試合程度まで上昇, 下降をしており, 安定していない. 一方で, 30 世代を見てみると, 20 試合前後の早い段階で 70%を超えており, 早い段階での勝率の収束が見られる. その後も, 勝率の緩やかな下降が見られるものの, 安定した勝率を示している.

4.3 アルゴリズム C との対戦結果

守備に重点を置いたアルゴリズム, アルゴリズム C (ALG_C) と対戦した結果を図 7 に示す. 初期世代を見てみると, 勝率は 60%をわずかに超える程度である. 10 世代では, 最初の数試合は非常に良い勝率を示したものの, これは誤差の範囲と考えられる. 試合を重ねていくと, 勝率は下降し, 最終的な勝率は 65%程度で収束しており, 安定した勝率を得られていない. 20 世代, 30 世代の勝率の推移を見てみると, 大きく差はないものの 20 世代, 30 世代ともに 50 試合程度の試合数で, 70%強程度の安定した勝率を示しており, 早い段階での勝率の収束が見られる.

5. 考察

5.1 対 ALG_A の結果に対しての考察

初期世代から 70%程度の勝率を示しているのは, ハイブリッド型意思決定システムで用いているアルゴリズムが ALG_A であるためと考えられる. 30 世代での勝率が早い段階で高い値に安定していることから, 報酬値が GA を用いることによって改善されているものと考えられる.

30 世代で最も良い勝率を示した報酬値を表 2 に示す. 表 2 からを見てみると, PASS の値では FW, MF, DF のそれぞれの値で大きく差がついている. このことからポジションごとの役割分担がなされているものと考えられる.

5.2 対 ALG_B の結果に対しての考察

30 世代では 20 試合程度で安定した勝率を示していることから, 効率よく学習できているものと考えられる. 10 世代, 20 世代の勝率は 30 世代と比べて安定するまでに時間がかかっていることから, 学習に時間がかかっていると考えられる.

30 世代で最も良い勝率を示した報酬値を表 3 に示す. 表 3 から, 報酬値に大きな差はないものの, DF の行動に比較的高い報酬が割り当てられている. この理由として, ALG_B が攻撃的なアルゴリズムであるため, 守備を役割とした DF の行動に重点が置かれた結果であると考えられる.

5.3 対 ALG_C の結果に対しての考察

勝率を比べてみると, 30 世代, 20 世代ともに安定した勝率を示しているが, 初期世代, 10 世代と比べてはるかに良い勝率を示していることから, GA を用いることによって質の良い学習を行える報酬値へと改善されているものと考えられる.

30 世代で最もよい勝率を示した報酬値を表 4 に示す. 表 4 より, PASS の値で FW, DF に大きな報酬が割り当てられている. また, MF の PASS の値は FW, DF と比べ非常に小さい値となっている. DF の DRIBBLE の値を見てみると, FW, MF の値に比べて小さな報酬が割り当てられている. この理由は, 対戦アルゴリズム ALG_C が守備的なアルゴリズムであることから, 学習の際に, 攻撃に重点を置いた報酬値表に進化しているためと考えられる.

表 2 対 ALG_A 30 世代で勝率が最高だった報酬値表

	FW	MF	DF
PASS	3	437	47
DRIBBLE	1	10	13

表 3 対 ALG_B 30 世代で勝率が最高だった報酬値表

	FW	MF	DF
PASS	19	26	30
DRIBBLE	10	15	36

表 4 対 ALG_C 30 世代で勝率が最高だった報酬値表

	FW	MF	DF
PASS	36	5	38
DRIBBLE	10	19	6

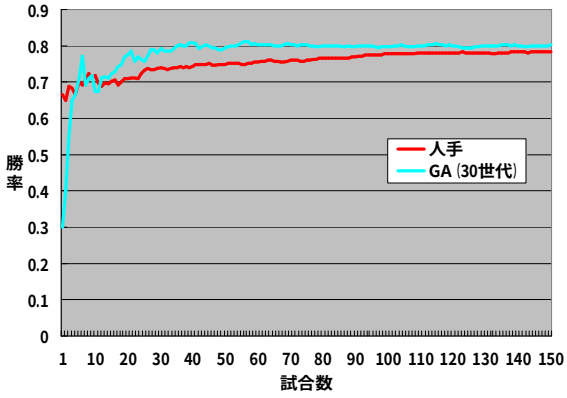

図 8 対 ALG_A 人手による設定と GA による設定の比較

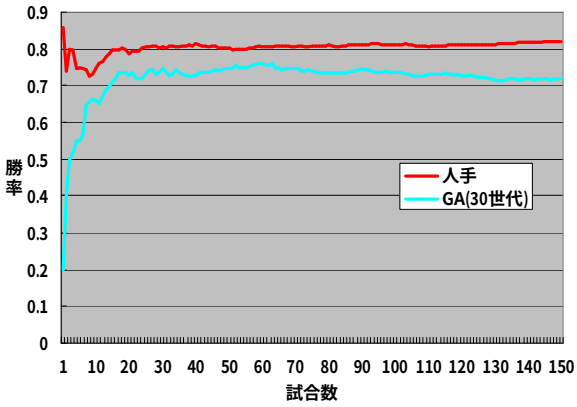

図 9 対 ALG_B 人手による設定と GA による設定の比較

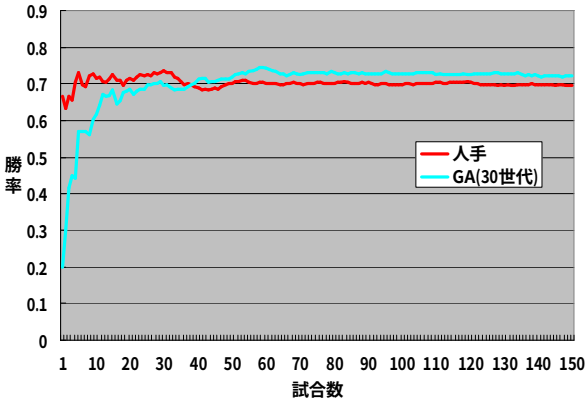

図 10 対 ALG_C 人手による設定と GA による設定の比較

5.4 人手による報酬値との比較

アルゴリズム A（ALG_A）における，人手による報酬設定と GA による報酬設定の勝率の比較を図 8 に示す．図 8 より，GA による設定は 50 試合程度で勝率が 80%に達している．一方で，人手による設定は 50 試合程度で 75%程度である．また，最終的な勝率を比較してみると，人手による設定が 80%弱であるのに対して，GA による設定では 80%と人手による設定よりも高い勝率を示している．以上のことから，ALG_A において，GA による設定は，人手による設定よりも，効率よく学習が行われているものと考えられる．

図 9 にアルゴリズム B（ALG_B）における，人手による報酬値設定と GA による報酬値設定の勝率の比較を示す．図 9 より，GA による報酬値設定の勝率は，人手による設定の勝率を下回っている．人手による設定と同等，もしくはそれ以上の勝率を得られなかった原因として考えられるのは，30 世代では最適な報酬値設定が求まらなかったためと考えられる．

図 10 にアルゴリズム C（ALG_C）における，人手による報酬値設定と GA による報酬値設定の勝率の比較を示す．図 10 より，GA による報酬値設定が，人手による報酬値設定の勝率を 40 試合から 50 試合程度で上回り，試合数を重ねていった後も，GA による報酬値設定が人手による報酬値設定よりも高い勝率を保っている．GA による報酬値設定が効率よく学習している結果と考えられる．従って，図 8～10 より，GA による報酬値設定は，人手による報酬値設定とほぼ同等の勝率を得られると考えられる．

6. まとめ

本研究では，サッカーゲームにおけるプレイヤーの意思決定にクラシファイアシステムを用いた場合，プレイヤーの行動に対する報酬値の設定に GA を用いることを提案し，その有効性をサッカーゲームによる実験で評価した．その結果，ゲームにおける勝率は上昇し，学習の効率の上昇が確かめられた．また，得られた勝率は，経験に基づいて試行錯誤で設定した報酬値の場合とほぼ同一の結果を得た．従って，報酬値設定という面倒な作業を GA により自動化できる見通しを得た．

参考文献

[Goldberg 89] Goldberg, D. E.: GENETIC ALGORITHMS in Search, Optimization, and Machine Learning, Addison-Wesley, 1989.

[伊庭 02] 伊庭斉志: 遺伝的アルゴリズム, 医学出版, 2002.

[Sato 05] Y. Sato and R. Kanno, “Event-driven Hybrid Learning Classifier Systems for Online Soccer Games”, Proc. of the 2005 IEEE Congress on Evolutionary Computation, IEEE Press, pp. 2091-2098 (2005).

[Sato 06] Y. Sato, Y. Akatsuka, and T. Nishizono, “Reward Allotment in an Event-driven Hybrid Learning Classifier System for Online Soccer Games”, Proc of the 2006 Genetic and Evolutionary Computation Conference, ACM Press, pp. 1753-1760 (2006).

[坂和 95] 坂和正敏, 田中雅博: 遺伝的アルゴリズム, 朝倉書店 1995