数理解析研究所講究録
第1990巻 2016年 136-145

レーティングシステムを利用した差分進化による コンピュータオセロプレイヤーの学習

広島修道大学商学部　阪井　節子 (Setsuko Sakai)
Faculty of Commercial Sciences, Hiroshima Shudo University
広島市立大学大学院 情報科学研究科　高濱　徹行 (Tetsuyuki Takahama)
Graduate School of Information Sciences, Hiroshima City University

1　はじめに

チェス，オセロのような完全情報2人ゼロ和ゲームは，ゲーム木で表現され，含まれる枝をすべて探索すれば理論上は最適解を確実に求めることができる．しかし，各節の平均分子数と木の平均深さによって推定されるゲーム木の大きさ (手数) は，チェスで約 3×10^{123} 手，オセロで約 10^{80} 手, 将棋で約 7×10^{218} 手であり，全件探索による最適解の決定は高速のコンピュータを用いても事実上不可能である．

近年，チェスやオセロなどの2人ゼロ和完全情報ゲームにおけるコンピュータプレイヤーに関する研究が活発に行われており，人間を上回る強さのプレイヤーの作成に成功している．これら多くのコンピュータプレイヤーには以下の3つの要素が含まれている．

- 定石: 序盤は定石に従って着手を選択する．
- 評価関数: 中盤は minimax 法や $\alpha\beta$ 法により着手を決定する．すなわち，合法手に基づき，ある深さまでゲーム木を構成し，葉となる盤面を評価関数で評価することにより次の手の評価値を決定し，最良手を選択する．
- 読み切り: 終盤は終局まで合法手を全て探索し，最適手を選択する．互いに最適手を選択すると仮定すると勝敗が確定することになる．

なお，囲碁のように評価関数の作成が困難なゲームでは，Monte Carlo 木探索を用いる方法もある．互いに終局まで合法手をランダムあるいは確率的に選択するプレイアウトを行うことにより，勝率の高い着手を選択するため，精密な評価関数が不要となる．

評価関数は非常に重要な要素であるが，評価関数の作成にはゲーム特有の知識が必要であり，試行錯誤的に評価関数を調整しなければならないという問題がある．この問題を解決する方法として，ゲームの勝敗を報酬として強化学習によりゲームの戦略を学習する方法，進化的計算などにより評価関数を学習する方法が提案されている．

本研究では，オセロゲームを対象とし，各プレイヤーがゲームのルール以外の事前知識を持たないという条件の下で，評価関数を学習する次の2種類の学習方法を提案する．

- 個人型学習：プレイヤーは1人のプレイヤーと対戦して戦略（評価関数）を学習する．まず，プレイヤーはランダムプレイヤーと対戦し戦略を学習する．次に，学習したプレイヤーと対戦してより良い戦略を学習する．
- グループ型学習：複数のプレイヤーがグループを形成し，グループ内で互いに対戦して戦略を学習する．初期グループは1人のランダムプレイヤーで構成される．プレイヤーはグループの全メンバーと対戦し学習する．プレイヤーがグループメンバーに対して一定の強さを持てば，グループメンバーとなる．

本研究では，オセロプレーヤーの評価関数を学習するために，進化的計算の一つである差分進化 (Differential Evolution; DE) により目的関数値を最適化する．学習プレイヤーが 1 人の対戦者プレイヤーと対戦して学習する個人型学習では，勝率を目的関数値とする．グループ内で互いに対戦して戦略を学習するグループ型学習では，グループ内でのレーティングを目的関数値とする．オセロプレイヤーは，学習した評価関数を用いて次の手番での行動を選択する．

本論文の構成は以下の通りである．2 節でオセロゲームと評価関数を説明する．3 節で差分進化によるオセロプレーヤの評価関数の進化アルゴリズムを提案する．4 節で個人型学習を提案し，実験結果を示す．5 節でグループ型学習を提案し，実験結果を示す．6 節はまとめである．

2 オセロ (リバーシ)

オセロは，交互手番 2 人 0 和完全情報のボードゲームである．各プレイヤーは 8 × 8 の盤上に交互に石を置いてゆく．石の片面は黒，他方は白となっている．先手プレイヤーは黒面を上に，後手プレイヤーは白面を上にして用いる．

2.1 ルール

オセロ (リバーシ) は，次のルールに従って進められる．

1. 盤の中央の 4 セルの内，右上と左下のセルに黒面を，左上と右下のセルに白面を上にして石を置いた状態からゲームを開始する．
2. 奇数手番のプレイヤー (Black と表す) は石の黒面を上に，偶数手番のプレイヤー (White と表す) は白面を上にして交互に打つ．各プレイヤーは，縦・横・斜め方向に相手色の石を自分色の石で挟み，挟まれた石は裏返され自分色になる．各手番でプレイヤーがとりうる手 (合法手) は，相手の石を裏返せるセルに石を置くことである．もし，合法手がなければこのときに限りパスとなる．両プレイヤーが連続してパスすれば，ゲームは終了である．
3. 自分色の石の数が多いプレイヤーが勝ち，少ない方が負け，同じときは引き分けである．

図 1 は，オセロ盤の初期状態と各プレイヤーの第 1 手目の合法手を示している．左図の☆は第 1 手番プレイヤー (Black) の全ての合法手である．右図の☆は，第 1 手番で Black がセル (C,4) に石を置いたときの，第 2 手番における White の全ての合法手である．

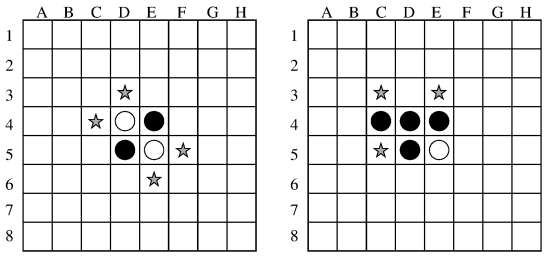

図 1: オセロ盤の初期状態と各プレイヤーの第 1 手目の合法手

2.2 オセロリーグ

人工的なオセロゲーム社会の例として，オセロ位置評価関数リーグ (Othello Position Evaluation Function League) [1] がある．このオセロリーグの目的は優れたプレイヤーを作ることではなく，オセロの (位置) 評価関数を評価することである．このリーグでは先読みとして 1 手先読みが採用されている．すなわち，現在の状態 (盤面上の石の配置) から次の状態を評価関数を使って評価し，最良評価値を与える合法手 (最良手) が次の手として選択される．最良手が複数ある場合は，その中から無作為に選択される．リーグでのプレイヤーの順位は，後述する標準 WPC(weighted price counter) を用いる標準プレイヤーと 1 手先読みで対戦を行った結果の得点で決定する．

オセロゲームとプレイヤーの行動は最良手が複数ある場合に無作為に選択することを除けば確定的であり，先手プレイヤーの決め方の違いによる 2 通りのゲームしかない．オセロリーグではよりよい性能測定を行うために，両プレイヤーが ε 貪欲戦略を用いることになってる．ε 貪欲戦略とは，合法手の中から確率 ε で無作為に手を選択し，確率 $1-\varepsilon$ で最良手を選択する戦略である．リーグでは $\varepsilon = 0.1$ としている．

2.3 Weighted piece counter(WPC)

The weighted piece counter (WPC) は，ゲーム盤の各セルに重みを割り当てる最も簡単な (位置) 評価関数である．x 行，y 列のセルを (x, y) と表し，セル (x,y) の状態を b_{xy}，重みを w_{xy}，と表す $(x,\ y = 1, 2, \cdots, 8)$．評価関数 $f_{\boldsymbol{w}}(\cdot)$ は，現在の状態 $\boldsymbol{b} = (b_{xy})$ から評価値への写像で次のように定義される．

$$f_{\boldsymbol{w}}(\boldsymbol{b}) = \sum_{y=1}^{8}\sum_{x=1}^{8} w_{xy} b_{xy} \tag{1}$$

$$b_{xy} = \begin{cases} 1, & \text{黒石が置かれている} \\ 0, & \text{空いている} \\ -1, & \text{白石が置かれている} \end{cases} \tag{2}$$

先手 Black は評価関数の最大化プレイヤーで，後手 White は最小化プレイヤーである．WPC に含まれる変数の数は 8×8=64 個であるが，オセロ盤の対称性を考慮すると，図 2 に示すように，10 変数 $(a \sim j)$ まで減らすことができる [3].

	A	B	C	D	E	F	G	H
1	a	b	c	d	d	c	b	a
2	b	e	f	g	g	f	e	b
3	c	f	h	i	i	h	f	c
4	d	g	i	j	j	i	g	d
5	d	g	i	j	j	i	g	d
6	c	f	h	i	i	h	f	c
7	b	e	f	g	g	f	e	b
8	a	b	c	d	d	c	b	a

図 2: WPC を決定する $a \sim j$ の 10 変数

	A	B	C	D	E	F	G	H
1	1	-0.25	0.1	0.05	0.05	0.1	-0.25	1
2	-0.25	-0.25	0.01	0.01	0.01	0.01	-0.25	-0.25
3	0.1	0.01	0.05	0.02	0.02	0.05	0.01	0.1
4	0.05	0.01	0.02	0.01	0.01	0.02	0.01	0.05
5	0.05	0.01	0.02	0.01	0.01	0.02	0.01	0.05
6	0.1	0.01	0.05	0.02	0.02	0.05	0.01	0.1
7	-0.25	-0.25	0.01	0.01	0.01	0.01	-0.25	-0.25
8	1	-0.25	0.1	0.05	0.05	0.1	-0.25	1

図 3: 標準 WPC

図 3 は，Yoshioka [2] 等が手作業で作成した標準 WPC (以後，標準 WPC と呼ぶ) で，オセロ研究で対戦者としてよく使われる．

3 差分進化を用いたWPCの学習

各プレイヤーの戦略がWPCであるオセロで戦略を学習する問題は，WPCを決定ベクトルとする最適化問題として定式化でき，その問題は差分進化 (Differential Evolution, DE) を用いて最適化できる．

3.1 WPCの学習

例えば，あるプレイヤーが他のプレイヤーと N 回対戦してWPCを学習する場合を考える．このとき，WPCの学習は，この対戦で得られる得点 (勝率など) の最大化問題として次のように定義できる．

$$\begin{aligned} &\text{maximize} \quad f(\boldsymbol{x}) \\ &\qquad f(\boldsymbol{x}) = (Win + 0.5Draw)/N, \quad \boldsymbol{x} = (a,b,c,d,e,f,g,h,i,j) \end{aligned} \tag{3}$$

ここで，Win は勝ちゲーム数，$Draw$ は引き分けゲーム数である．この種の得点は，実際のオセロトーナメントでも用いられている．

3.2 差分進化 (Differential Evolution)

差分進化 (Differential evolution; DE) はStorn and Price[4, 5] によって提案された進化的アルゴリズムである．DEは確率的な直接探索法であり，解集団を用いた多点探索を行う．DEは非線形問題，微分不可能な問題，非凸問題，多峰性問題などの様々な最適化問題に適用されてきており，これらの問題に対して高速で頑健なアルゴリズムであることが示されてきている．

DEでは，探索空間中にランダムに初期個体を生成し初期集団を構成する．各個体は決定ベクトルに対応し，n 個の決定変数を遺伝子として持つ．各世代において，全ての個体を親として選択する．各親に対して，次のような処理が行われる．選択された親を除く個体群から互いに異なる $1+2\,num$ 個の個体を選択する．最初の個体が基本ベクトルとなり，残りの個体対が差分ベクトルとなる．差分ベクトルに F(scaling factor) が乗算され基本ベクトルに加えられ，変異ベクトル (mutant vector) が生成される．変異ベクトルと親が交叉し，CR(crossover factor) により指定された確率で親の遺伝子をベクトルの要素で置換することにより，子のベクトル (trial vector) が生成される．最後に，生存者選択として，子が親よりも良ければ，親を子で置換する．

DEには幾つかの形式が提案されており，DE/best/1/binやDE/rand/1/expなどがよく知られている．これらは，DE/*base*/*num*/*cross* という記法で表現される．"*base*" は基本ベクトルとなる親の選択方法を指定する．例えば，DE/randは基本ベクトルのための親を集団からランダムに選択する．"*num*" は基本ベクトルを変異させるための差分ベクトルの個数を指定する．例えば，DE/rand/1は，各親 $\boldsymbol{x}_i$ に対して，3個体 $\boldsymbol{x}^{r1}, \boldsymbol{x}^{r2}, \boldsymbol{x}^{r3}$ を $\boldsymbol{x}_i$ および互いに重複しないようにランダムに選択する．基本ベクトル $\boldsymbol{x}^{r1}$ および差分ベクトル $\boldsymbol{x}^{r2} - \boldsymbol{x}^{r3}$ から変異ベクトル $\boldsymbol{m}$ を以下のように生成する．

$$\boldsymbol{m} = \boldsymbol{x}^{r1} + F(\boldsymbol{x}^{r2} - \boldsymbol{x}^{r3}) \tag{4}$$

ここで，F はスケーリングパラメータである．"*cross*" は子を生成するために使用する交叉方法を指定する．例えば，'bin' は一定の確率で遺伝子を交換する二項交叉 (binomial crossover) を表し，'exp' は，指数関数的に減少する確率で遺伝子を交換する指数交叉 (exponential crossover) を表す．変異ベクトル $\boldsymbol{m}$ と親 $\boldsymbol{x}_i$ を交叉率 CR で交叉し，子ベクトル (trial vector) $\boldsymbol{x}_i^{\text{child}}$ を生成する．

図4に，DE/rand/1/binの擬似コードを示す．

```
DE/rand/1/bin()
{
// Initialize a population
 P=NP individuals generated randomly in S;
 for(t=1; t ≤ T_max; t++) {
  for(i=1; i ≤ NP; i++) {
// DE/rand/1/bin operation
     x^r1=Randomly selected from P(r1 ≠ i);
     x^r2=Randomly selected from P(r2 ∉ {i,r1});
     x^r3=Randomly selected from P(r3 ∉ {i,r1,r2});
       m=x^r1+F(x^r2 − x^r3);
     x^child=trial vector is generated from
     x^i and m by the binomial crossover operation;
// Survivor selection
     if(f(x^child) > f(x^i)) z^i=x^child;
     else                    z^i=x^i;
  }
  P={z^i, i = 1,2,···,NP};
 }
}
```

図 4: DE の擬似コード．ここで，S は探索空間，$T_{\max}$ は最大世代数である．

4 個人型学習

個人型学習では，学習プレイヤーは 1 人の対戦者プレイヤーと対戦し WPC を学習する．

4.1 適合度

本研究では，プレイヤーはオセロリーグと同様に，0.1-貪欲戦略を用いる．また，ゲームの得点が非常に不安定なため，得点の下限値を最大化する．先手後手プレイヤーを入れ替えて N 回対戦を行いこれを 1 試合として S 試合行い，1 試合毎の平均得点を記録する．目的関数値（適合度）は，総平均得点と平均得点の標準偏差を用いて定義する．個人型学習による WPC の学習は，次のような最大化問題として定義される．

$$\begin{aligned} \text{maximize} \quad & f(\boldsymbol{x}) = \text{score}(\boldsymbol{x}) - \alpha\,\sigma(\boldsymbol{x}) \\ & \text{score}(\boldsymbol{x}) = \frac{1}{S}\sum_{i=1}^{S}\text{score}_i(\boldsymbol{x}), \quad \sigma(\boldsymbol{x}) = \sqrt{\frac{1}{S}\sum_{i=1}^{S}(\text{score}_i(\boldsymbol{x}) - \text{score}(\boldsymbol{x}))^2} \\ & \text{score}_i(\boldsymbol{x}) = (Win_i + 0.5Draw_i)/N, i = 1,2,\cdots,S \end{aligned} \tag{5}$$

ここで，α は得点と標準偏差の間の重みであり，Win_i と $Draw_i$ はそれぞれ第 i 試合の勝ち数と引き分け数である．$\alpha = 1.3$ のとき $f(\boldsymbol{x}) > 0.5$ ならば，信頼度 90%で対戦者に対する学習プレイヤーの平均勝率が 50%超えている（すなわち，対戦者より強い）と判断できる．

4.2 アルゴリズム

個人型学習のアルゴリズムは，次のようになる．

1. 対戦者としてランダムプレイヤーを用意する．反復回数 k=1.
2. 学習者として NP 個の個体 (WPC) を生成し，初期集団を構成する．
3. 学習者と対戦者が対戦し，学習者の適合度 f を求める．世代数 $t = 1$.
4. 全ての個体を親として順次選択し，DE/rand/1/bin を用いて子を生成する．子は対戦者と $S \times N$ 回対戦し適合度 f' を評価する．$f' > f$ ならば子が生存者となり，そうでなければ親が生存者となる．
5. 集団を生存者集団で置換する．
6. t=t+1. t が最大世代数 $T_{\max}$ の半分を超えていなければ, 4) に戻る．

7. 最良個体が判定条件を満足すれば，最良個体を対戦者と置換し 2) に戻る．そうでなければ，置換は行わず次に進む．判定条件は，最良個体の適合度が 0.5 を超えることである．

8. t が最大世代数 $T_{\max}$ を超えていなければ, 4) に戻る. そうでなければ，学習は失敗で次に進む．

9. k=k+1. k が最大反復回数 $K_{\max}$ を超えていなければ，2) に戻る．そうでなければ，終了する．

4.3 実験

表 1 に実験条件を示す．学習によって得られた戦略 (WPC) の性能を調べるために, 対戦者を更新したとき新しい最良 WPC と標準 WPC を 1000 回対戦させ勝率を求める．

表 1: 個人型学習の実験条件

最大反復回数	K_{max}	200
最大世代数	T_{max}	100
集団サイズ	NP	20
標準偏差の重み	α	1.3
試合数	S	10
1 試合当たりの対戦数	N	50
x の各成分の初期値		[-1, 1]
スケーリングファクタ	F	0.5
交叉率	CR	0.9

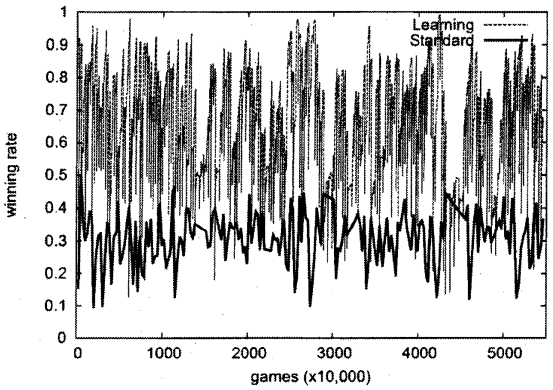

図 5: 個人型学習の実験結果

図 5 は，実験結果である．細線は，最良 WPC の対戦者 WPC に対する勝率の推移を表す．太線は，対戦者 WPC が更新されたときの最良 WPC の標準 WPC に対する勝率の推移を示すグラフである．なお，横軸は対戦ゲーム数 (単位 : 10,000 ゲーム) である．この実験で 200 (K_{max}) 個の最良 WPC が生成される．1 個の WPC を評価するために 500 ($S \times N$) 回対戦し，20(N) 個の WPC を評価するには 10,000 回対戦する．したがって，初期集団の評価と 100($T_{\max}$) 世代の評価のためには，あわせて 1,010,000 回の対戦が必要である．細線のグラフからわかるように，新しい最良 WPC の古い最良 WPC(対戦者 WPC) に対する勝率は 70%から 80%となっており，安定的に 50%を超えている．このことから，個別型学習によって古い戦略よりも強い戦略を学習できたと考えられる．しかし，太線のグラフからわかるように，新しい最良 WPC の標準 WPC に対する勝率は 30%程度であり，安定して標準 WPC より強い戦略を得られていない．このことから，個別型学習ではよい戦略が学習できるが，更新されたときよい戦略がすぐに失われてしまっていると考えられる．

5 グループ型学習

個別型学習では，安定した学習結果を得ることが困難であった．本節では，グループで学習を行うグループ型学習を提案する．ここでは，適切なプレイヤーによるグループを作るために，レーティングシステムを用いる．新しいプレイヤーがグループの一員になるには，グループのメンバー全員と対戦しレーティングを決定する．レーティングがよければ，ゲームのメンバーとなる．メンバー数が上限を超えれば，最低レーティングのプレイヤーはグループを退会する．

5.1 レーティングと適合度

レーティングシステムは，プレイヤーの強さを評価するために用いられる．レーティングは，プレイヤーが他のプレイヤーと対戦した結果の勝敗数によって求められプレイヤーの強さを表す尺度である．色々なレーティングシステムがあるが，本研究ではよく知られているイロレーティングシステム [6] を用いる．

イロレーティングにおける 1 ゲームの勝点は，勝者は 1, 敗者は 0, 引き分けのときは両者とも 0.5 である．例えば，プレイヤー A と B が 1 ゲームを行ったとき，A が勝てば A の勝点は 1, B は 0 となる．A と B のレーティングをそれぞれ R_A と R_B，勝点をそれぞれ S_A と S_B とする．このとき，A と B の期待勝点 E_A と E_B は，それぞれ次式で与えられる．

$$E_A = \frac{1}{1+10^{(R_B-R_A)/400}}, \quad E_B = \frac{1}{1+10^{(R_A-R_B)/400}} \tag{6}$$

ここで，$E_A + E_B = 1$ となる．このとき，A と B の新しいレーティングは，次式で与えられる．

$$R_A = R_A + K(S_A - E_A), \quad R_B = R_B + K(S_B - E_B) \tag{7}$$

ここで，K はパラメータである．K=32 とレーティングの既定値として 1,500 がよく用いられる．

グループ型学習における WPC の学習問題は，次のように定義される．

$$\text{maximize} \qquad f(\boldsymbol{x}) = rating(\boldsymbol{x}, G) \tag{8}$$

ここで，$rating(\boldsymbol{x}, G)$ は，WPC $\boldsymbol{x}$ を持つプレイヤーのレーティングである．レーティングは，グループ G の各メンバーと N 回ずつ対戦して求める．レーティングの決定に必要な総ゲーム数は $N \times |G|$ である．$|G|$ はグループ G のメンバー数である．

5.2 アルゴリズム

グループ型学習のアルゴリズムは次のようになる．

1. レーティングが 1,500 のランダムプレイヤー 1 人だけのグループ G を用意する．反復回数 k=1.
2. 学習者として NP 個の個体 (WPC) を生成し，初期集団を構成する．各 WPC は G の全メンバーとそれぞれ N 回対戦しレーティングを求める．世代数 $t = 1$.
3. 集団の全ての個体を親として選択し，DE/rand/1/bin を用いて子を生成する．子は G の各メンバーとそれぞれ N 回対戦しレーティングを求める．親と子のレーティングの高い方が生存者となる．
4. 集団を生存者集団で置換する．
5. t=t+1. t が最大世代数 $T_{\max}$ の半分を超えていなければ, 3) に戻る．
6. 最良個体が判定条件を満足すれば，最良個体と G のメンバー全員がそれぞれ N 回対戦するリーグ戦を行ってレーティングを決定する．$|G|$+1 がグループ G の定員 $G_{\max}$ を超えていれば，最低レーティングの個体を G メンバーから外し，2) に戻る．判定条件を満足していなければ，G のメンバーの更新は行わず次に進む．判定条件は，最良個体のレーティングが 1600 を超えることである．
7. t が最大世代数 $T_{\max}$ を超えていなければ，3) に戻る．そうでなければ，学習は失敗で次に進む．
8. k=t+1. If k が最大反復回数 $K_{\max}$ を超えていなければ, 3) に戻る．そうでなければ，終了である．

5.3 実験

表 2 に実験条件を示す.

表 2: グループ型学習の実験条件

最大反復回数	$K_{\max}$	200
最大世代数	$T_{\max}$	100
集団サイズ	NP	20
グループの定員	$G_{\max}$	10
各グループメンバーとの対戦数	N	50
レーティングパラメータ	K	32
レーティングの既定値		1,500
$\boldsymbol{x}$ の各成分の初期値		[-1, 1]
スケーリングファクタ	F	0.5
交叉率	CR	0.9

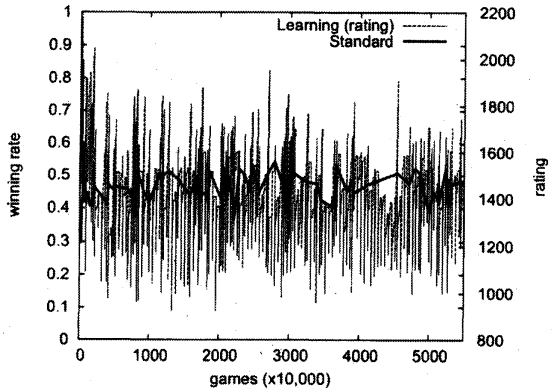

図 6: グループ型学習の実験結果

図 6 は，実験結果のグラフである．細線は，集団の最良 WPC のレーティングの推移を表す．太線は，対戦者グループの一更新が行われたとき，グループの最良 WPC と標準 WPC が 1,000 回対戦した結果の最良 WPC の勝率である．グラフの横軸はゲーム回数 (単位:10,000 回) である．1 回の実験で 200 ($K_{\max}$) 個の最良 WPC が生成される．対戦者グループのメンバー数は 1〜10($G_{\max}$) であり，1 個の WPC はレーティングを求めるために各グループメンバーと 50(N) 回ゲームを行うので，1 個の WPC のレーティングを求めるには 50〜500($|G| \times N$) 回ゲームをすることになる．よって 20(NP) 個の WPC を評価するには 1,000〜10,000 回のゲームをする．集団の初期化と 100($T_{\max}$) 世代の評価に，101,000〜1,010,000 回のゲームを行う．すなわち，1 個の最良 WPC は，高々1,010,000 回のゲームで生成される．さらに，グループメンバーの更新には 50〜2,250 ($N \times_{|G|} C_2$) 回ゲームが行われる．

表 3 は，約 5500 万回ゲームを行った時点での対戦者グループメンバーの WPC をまとめたものである．最良 WPC である No.6 の平均勝率は 0.483, 標準偏差=0.0697 である．

表 3: 5500 万試合あたりの対戦者グループの WPC

No.	Rating	a	b	c	d	e	f	g	h	i	j
1	1341	1.85	0.2	-0.62	1.42	0	-0.92	0.29	0.31	0.95	0
2	1601	1.94	0.07	0.72	0.01	-0.78	-0.66	0.01	0.38	0.16	-0.13
3	1516	1.08	-0.11	0.69	0.04	-0.33	-0.34	-0.07	-0.01	0.21	-0.03
4	1591	2.07	-0.7	1.52	-0.28	-1.64	-0.47	-0.1	0.15	0	-0.3
5	1604	2.34	-0.53	1.18	-0.03	-0.73	-0.1	-0.22	-0.06	0.15	-0.2
6	**1710**	**1.86**	**-0.2**	**1.53**	**-0.21**	**-1.54**	**-1.07**	**-0.22**	**0.49**	**0.42**	**-0.37**
7	1505	11	-1.58	0.6	0.4	-1.69	-0.53	-0.33	0.76	0.86	-0.67
8	1644	1.84	-0.29	0.83	-0.3 2	-1.18	-0.26	0.05	0.22	-0.02	-0.04
9	1294	0.49	0.4	-0.99	0.81	-0.39	-0.98	0	0.93	-0.24	0.35
10	1194	1.58	0.26	-0.67	1.06	-0.03	-0.83	0.97	0.66	0.28	-0.15

表 4 は，表 3 の 10 個の WPC と標準 WPC が 10 万試合 (50 試合を 2000 回) 行った結果である．

表 4: グループ WPC と標準 WPC の対戦結果

No.	平均	標準偏差
1	0.37482	0.0682486
2	0.48117	0.0711487
3	0.429065	0.0682567
4	0.44901	0.0695372
5	0.42012	0.0675928
6	0.475255	0.0697942
7	0.446095	0.0693073
8	0.51983	0.0697006
9	0.4275	0.0674741
10	0.288475	0.0628866

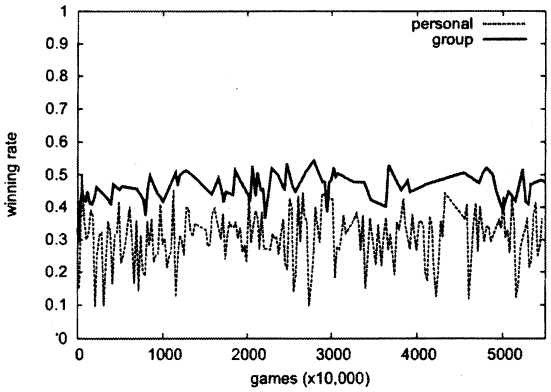

図 7: 個別学習とグループ型学習の結果比較

表 3 でレーティングが最大となった No.6 の WPC が標準 WPC との対戦勝率では第 3 位となっており，2 位であった No.8 が 1 位となっている．

図??は，標準 WPC に対する個人型学習とグループ型学習による最良 WPC の勝率を示したものである．グループ学習による WPC の方が安定的に高い勝率を示しており，複数対戦者に勝る戦略を学習するグループ型学習によって，比較的安定的な学習に成功したと考えられる．

6 おわりに

本研究では，2 つの学習モデル，個別型学習とグループ型学習を提案した．個別型学習はしばしばよい戦略を学習できるが，すぐに失われ安定した学習ができないことを示した．レーティングを用いたグループ型学習もよい戦略を学習できる．グループ型学習では，よい戦略は失われず，グループ内に保持されることを示した．さらに重要なことは両方のモデルとも標準プレイヤーに関する事前知識を用いずに標準プレイヤーよりもよいプレイヤーを生成することができるということである．

今後は，グループ型学習においてグループ全体を差分進化の個体集団に対応さえる方法について検討する予定である．

謝辞

本研究の一部は，JSPS 科研費 (C)(No.24500177, 26350443) の援助を受けた．

参考文献

[1] K. Krawiec and M. G. Szubert, "Learning N-tuple networks for Othello by coevolutionary gradient search," in *Proc. of the 13th annual conference on Genetic and evolutionary computation.* ACM, 2011, pp. 355–362.

[2] T. Yoshioka, S. Ishii, and M. Ito, "Strategy acquisition for the game Othello based on reinforcement learning," *IEICE Transactions on Information and Systems*, vol. 82, no. 12, pp. 1618–1626, 1999.

[3] P. Hingston and M. Masek, "Experiments with Monte Carlo Othello," in *IEEE Congress on Evolutionary Computation 2007*, Sep. 2007, pp. 4059–4064.

[4] R. Storn and K. Price, "Minimizing the real functions of the ICEC'96 contest by differential evolution," in *Proc. of the International Conference on Evolutionary Computation*, 1996, pp. 842–844.

[5] ——, "Differential evolution – A simple and efficient heuristic for global optimization over continuous spaces," *Journal of Global Optimization*, vol. 11, pp. 341–359, 1997.

[6] A. E. Elo, *The rating of chessplayers, past and present.* Batsford London, 1978.