モンテカルロ大貧民プレイヤの自己対戦を用いた良好な棋譜データの抽出とシミュレーション方策の学習

○川岸成輝 岡田直也 永田祐一 小野典彦 （徳島大学）

Extraction of effective game record and learning of high quality simulation policy based on self-play by Monte Carlo Daihinmin players

∗N.Kawagishi N.Okada Y.Nagata and N.Ono (Tokushima University)

Abstract– In games with perfect information, computer players have been created using self-play that outperform human. In this paper, aimed at establishing a method for creating powerful computer players in imperfect-information games, we confirm how powerful computer players we can create using self-play and a Monte Carlo method in UEC Computer Daihinmin Contest. By extending the method proposed by Ohto and Tanaka, we propose its variant to extract effective game record and learn high quality simulation policy based on self-play by Monte Carlo Daihinmin players and discuss its performance through experiments.

Key Words: Imperfect-information games, self-paly, Monte Carlo methods, Computer Daihinmin Contest

1 はじめに

人工知能では，チェス，将棋，囲碁などの完全情報ゲームを対象として，人を凌駕するようなコンピュータプレイヤ（以下では特に断りのないかぎり，プレイヤと呼ぶ）の制作を目指した研究が展開されてきた．最近になるまで，これらの研究は対戦棋譜や局面の特徴量などのように人がお互いの対戦を通して獲得してきたゲーム固有の知識と教師付き学習との組合せに基礎をおく知識依存型の研究が主流であった．しかしながら，これらの知識は人と人との対戦から生まれた知識であり，人を凌駕するようなプレイヤを実装するための知識としては適切とはいえない．実際，DeepMind 社の AlphaGoZero はモンテカルロ法および深層学習に基礎をおくプレイヤ間の自己対戦を用いて制作され，ゲーム固有の知識を用いることなく，人の囲碁・世界王者を圧倒する性能を有している [1]．DeepMind 社は AlphaGoZero の枠組みを発展させ，囲碁だけでなく，チェスや将棋などの完全情報ゲームプレイヤの制作にも適用可能な枠組み AlphaZero を提案し，チェスおよび将棋にもおいても最強の知識依存型プレイヤに圧勝するプレイヤの制作に成功している．[2]

人を凌駕するようなプレイヤの制作を目指した研究は，完全情報ゲームだけでなく，ポーカー，大貧民および麻雀などの不完全情報ゲームに関しても進められている．しかし，これらの研究も知識依存型のものが主流であり，人が生み出したゲーム固有の知識を排したプレイヤの制作はほとんど試みられてこなかったというのが実情である．本研究の目的は，電気通信大学主催のコンピュータ大貧民大会（UECda）を対象として，ゲーム固有の知識や過去の優勝者プログラムの対戦棋譜などを可能なかぎり利用することなく，自己対戦とモンテカルロ法を用いて，どこまで強力なプレイヤが製作可能なのか実験的に確認することにある．

自己対戦とモンテカルロ法を用いた大貧民プレイヤの制作に関しては大渡らによる先駆的な研究 [3] がある．大渡らは，現実の対戦でプレイヤが採用しているとは考えにくいランダムなシミュレーション方策（以下では単に方策と呼ぶこともある）ではなく，現実の対戦棋譜に符合する方策を用いることで，強力なモンテカルロプレイヤが製作可能だと考え，そのような方策を自己対戦により学習可能な手法を提案している．しかし，一般にモンテカルロ法に基づくプレイヤの性能はその方策に依存し，強力な方策を有するプレイヤほど強力なプレイヤとなると考えられている．したがって，強力なプレイヤを製作するためには，単に現実の対戦棋譜に符合する方策ではなく，そうした棋譜の中でも特に良質なものだけに符合する方策を用いたモンテカルロプレイヤのほうがより良好な性能を示すものと期待できる．本研究では，以上の考えに基づき，先行研究となる大渡の手法を発展させた手法を提案し，両手法の比較実験を通じてその性能評価を行う．

2 モンテカルロ大貧民プレイヤ

大渡らによる先行手法は，大貧民以外の不完全情報ゲームにも適用可能な枠組みであるがコンピュータ大貧民大会 (UECda)[4, 6] のプレイヤ制作への適用を通じて評価されている．

ここで，UECda とはトランプゲーム大貧民 (大富豪と呼ばれることがある) を対象として，2010 年から電気通信大学が主催しているプログラミングコンテストである．この大会にはヒューリスティックスな手法のみを用いるライト級および手法の制限がない無差別級の 2 種類の部門がある．大渡らの先行手法は無差別級のプレイヤ制作に適用され，過去の優勝プログラムに迫るプレイヤを自己対戦により制作することに成功している．

過去の優勝プログラムの多くはモンテカルロ法を用いており，大渡らの先行研究でもプレイヤはモンテカルロ法を用いて動作するが，以下の特色をもつ．

- UECda ではプレイヤが同一の相手プレイヤ達 4 人と一定の回数 (通常は 100 回) 試合を行い，その合計得点で勝敗を競う．その際，前試合の結果に応じてカード交換が行われ，次の試合に影響を与えることから，プレイヤにとっては役の提出だけでなくカード交換も合法手に含まれ，その選択もモンテカルロ探索の対象となる．

- プレイヤはその戦略に相当するシミュレーション方策にしたがってモンテカルロシミュレーションを行う．しかし，かならずしも試合終了までシミュレーションを行うわけではない．プレイヤが残り 2 人となった場合にはそれ以上のシミュレーションは行わず，探索により勝敗を調べる．また明らかな必勝手が見つかった場合にはシミュレーション方策にしたがうことなく，その手でシミュレーションを続ける．

- モンテカルロシミュレーション時の行動の割り振りは UCB1-Tuned[5] にいくつかの修正を施したものを用いて行っている．

- 状態 s における行動 s の選択確率 $\pi_\theta(s,a)$ を決定するシミュレーション方策は次の通りである．

$$\pi_\theta(s,a) = \frac{e^{\phi(s,a)\cdot\frac{\theta}{T}}}{\sum_{b\in A} e^{\phi(s,b)\cdot\frac{\theta}{T}}} \tag{1}$$

ここで $\phi(s,a)$ は状態･行動対の特徴ベクトル，θ は各特徴の重みベクトル，A は状態 s における行動集合関数，T(=1) は温度パラメータである．特徴ベクトルの要素については大渡らの論文 [3] を参照されたい．プレイヤの戦略に相当するシミュレーション方策の質は重みベクトル θ で決まることに注意されたい．

3 先行手法

一般にモンテカルロ法の性能はシミュレーション方策の質によって決まる．上記のモンテカルロ大貧民プレイヤの場合は，そのシミュレーション方策を決定する重みベクトル θ によって，その性能が決まる．したがって，強力なモンテカルロ大貧民プレイヤを制作するためには，任意の相手プレイヤ達との対戦においても得られる合計得点の期待値が最大となるような良好な重みベクトル θ を学習すればよい．

大渡らは，(i) 同一の重みベクトル θ をもつプレイヤ達を一定期間，繰り返し自己対戦させて対戦棋譜を生成し，(ii) それらにプレイヤ達のシミュレーション方策が符合するように θ を更新するという 2 ステップを繰り返すことで，良好な θ を学習する手法を提案している．ただし，AlphaGoZero や AlphaZero などの自己対戦手法とは異なり，教師の方策が決定論的であるものとして学習する．

この手法の詳細は以下の通りである．

1. 現在の世代 $g = 0$ とする
2. 重みベクトル θ を零ベクトルに初期化する
3. $g = g + 1$
4. 現在の θ をもつプレイヤどうしの対戦棋譜を生成する
5. ステップ 2 で生成した対戦棋譜 (教師データ) に対して，シミュレーション方策が符合するように θ を学習する
6. ステップ 3 に戻る

ただし，ステップ 5 ではシミュレーション方策が決定論的となるように温度パラメータ T=0 として学習を行い，ステップ 4 では T の値をもとの値 1 に戻して自己対戦を行うものとする

教師の方策 π_* が決定論的な場合，状態 s での教師の行動が x のときの π_* と π_θ の誤差関数は

$$L(\pi_*, \pi_\theta) = -ln\pi_\theta(s,x) \tag{2}$$

で与えられ，重みベクトル θ は最急降下法を用いて，つぎの更新式で学習する．

$$\theta \leftarrow \theta + \frac{\alpha}{T}[\phi(s,x) - \sum_{b\in A}\pi_\theta(s,b)\phi(s,b)] \tag{3}$$

ここで α は学習率である．

4 提案手法

4.1 先行手法に関する疑問点

前節で説明した先行手法は単純でありながら少ない世代の学習だけで UECda の過去の優勝プログラムに迫る性能のプレイヤの制作に成功している．しかしながら，このような結果が得られる原理について大渡らはほとんど説明していない．

筆者らの理解は以下の通りである．一般にモンテカルロ法に基づくプレイヤは，そのシミュレーション方策 π_θ が強力であるほど強力であり，しかも π_θ そのものよりも強力である (実際，大渡らはこれを示唆する実験結果を示している)．したがって，π_θ を自分自身の対戦棋譜に符合するように更新することで，自分自身がさらに強力となる．この原理が働くため，自己対戦によりプレイヤの性能を強化できたのであろう．

先行手法に関しては対戦棋譜の利用方法に関しても大きな疑問点がある．これは現在の重みベクトル θ をもつプレイヤ達の対戦棋譜全体を教師データとして学習を行っている点である．対戦棋譜は 5 人のプレイヤが経験するエピソードの集合体であり，それらのエピソードには教師データとしてふさわしいものとそうではないものがあるはずであるが，先行手法ではそれらをすべて教師データとして採用している．

これでは，前の試合で大富豪だったにも関わらず，その順位を維持できないばかりか大貧民になってしまったときのエピソードまで教師データとして利用していることになる．もちろん大貧民というゲームでは，プレイヤが最善の手を取り続けていてもそのようなエピソードを回避できないこともあるため，それが教師データとしてふさわしくないとは一概にいえないのであるが，それを教師データから除外することで，より良好な結果が得られる可能性があると考えられる．

4.2 提案手法の概要

以上の疑問点を踏まえ，以下のように棋譜データの利用方法にのみ修正を施した先行手法の変形版を提案する．

1. 現在の世代 $g = 0$ とする
2. 重みベクトル θ を零ベクトルに初期化する
3. $g = g + 1$

4. 現在の θ をもつプレイヤどうしの対戦棋譜を生成する

5. ステップ 2 で生成した対戦棋譜から良好なエピソードだけ（教師データ）を抽出し，それらに対してシミュレーション方策が符合するように θ を学習する

6. ステップ 3 に戻る

ただし，ステップ 5 ではシミュレーション方策が決定論的となるように温度パラメータ T=0 として学習を行い，ステップ 4 では T の値をもとの値 1 に戻して自己対戦を行うものとする．

なお，良好なエピソードとは対戦棋譜に含まれる以下の選別基準に合致するものだけである．

[1] 大富豪以外のプレイヤがその地位を上げることに成功したもの

[2] 大貧民以外のプレイヤがその地位を維持することに成功したもの

大貧民の場合には，いかなるエピソードが良好なのかは自明ではなく，他にも様々な選別基準が考えられるが，明らかに問題のある基準もある．たとえば，自己対戦の結果，合計得点が第 1 位となったプレイヤの全エピソードのみを良好と考えることも可能であるが，この基準では最終的に第 1 位にはなれなかったものの僅かな差で第 2 位となったプレイヤや下位の地位にありながら頻繁にその地位を上げることに成功していたプレイヤのエピソードを教師データとして学習を行うことができない．高い地位にあっても低い地位にあってもたえず良好な行動が選択可能なプレイヤを制作したいのであれば，上記の選別基準が適当なものの 1 つとなる．

5 実験

ここでは以下の 2 種類の実験を行い，先行手法と提案手法の性能比較を行った．

(1) 両手法を用いてシミュレーション方策の学習を行い，その過程でシミュレーション方策そのものおよびそれに基づくモンテカルロプレイヤの性能の推移を比較する．これを実験 1 とする．

(2) 両手法を用いて学習済みのプレイヤおよび過去の UECda 優勝プレイヤのリーグ戦を行い，それらの性能を比較する．これを実験 2 とする．

5.1 実験 1

この実験では，先行手法および提案手法でシミュレーション方策の学習を行いながら，各世代の学習が終了する度にシミュレーション方策そのものならびにそれを用いたモンテカルロプレイヤを共通の相手と対戦させ，それらの性能の推移を比較した．対戦相手は，Talbe 1 に示す過去の UECda 決勝進出プログラムとした．

Table 1: List of opponents

Blauweregen	Glicine	Fujigokoro	
paoon	jn16	wisteria	kowl

いずれの手法においても毎世代 5000 試合分の対戦棋譜を生成し，先行研究ではそのすべてを，提案手法では前章の選別基準 [1] および [2] に合致するエピソードだけを棋譜から抽出してそれらをそれぞれ教師データとして学習した．

そして，毎世代の学習が終了する度にシミュレーション方策そのものおよびそれを用いるモンテカルロプレイヤをそれぞれ上記の対戦相手と対戦させ平均得点の推移を調べた．あらゆる組合せで 100 試合を 1 セットして 5 セットずつ対戦を行った．

Fig. 1 がシミュレーション方策そのものの平均得点，Fig. 2 がその方策を用いたモンテカルロプレイヤの平均得点の推移である．いずれの結果においても提案手法が先行研究をやや上回る性能を示すことが多いが，明らかな差は認められなかった．また，概ねどの世代においてもシミュレーション方策そのものよりもそれを用いたプレイヤのほうが強力であることも確認できた．

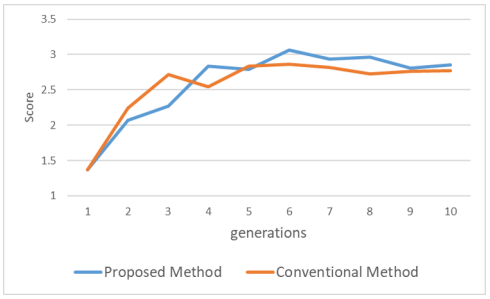

Fig. 1: Performance by simulation policy

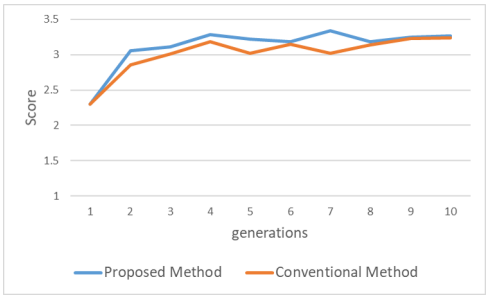

Fig. 2: Performance by Player

5.2 実験 2

ここでは先行手法および提案手法でそれぞれ 10 世代だけ学習済みのプレイヤに実験 1 でも用いた過去の UECda プレイヤ Blauweregen，Glicine，paoon，wisteria，Fujigokoro の 5 人を加えた合計 7 人による各対戦 100 試合ずつのリーグ戦を合計 2 回行い，総得点数を集計して性能比較を行った

各プレイヤの 4200 試合での総獲得ポイントの結果を Fig.3 に示す．

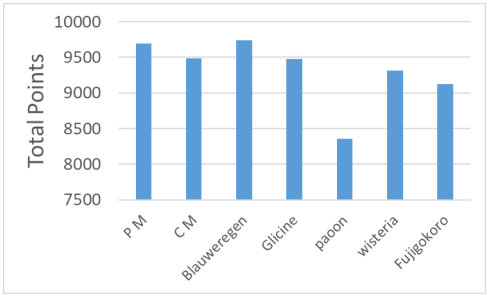

Fig. 3: Results against the champions.

Fig.3 より提案手法プレイヤは歴代 UECda 優勝プ

Table 2: Each Player's Best/Worst Score and the number of Over 300 games

Player	Proposed Method	Convertion Method	Blauweregen	Glicine	paoon	wisteria	Fujigokoro
Best	394	362	354	372	323	347	320
Worst	266	254	251	248	210	221	253
Over 300	20	20	23	18	3	14	13

レイヤの中でも上位の強さを有していることが確認できる．

6 考察

提案手法に関してはさらなる実験を通じた検証が必要であるが，興味深い実験結果がいくつか確認できた．

実験 1 では，先行手法と同様，提案手法でも，多様な過去の UECda 優勝プログラムなどとの対戦を経ることなく，それらに対応可能なプレイヤのシミュレーション方策を自己対戦だけで学習可能であることが確認できた．提案手法では，先行手法よりも少ない教師データで先行手法とほぼ同等の速度でほぼ同等もしくはそれ以上の性能のシミュレーション方策が学習可能であることも確認できた．

実験 2 の先行手法および提案手法で学習済みのプレイヤに歴代 UECda プレイヤを交えて行ったリーグ戦の結果は，自己対戦でそれらの中で上位の成績をおさめるプレイヤの制作の可能性を示唆しており，さらに興味深い．これをうかがわせるデータもある．Table 2 はリーグ戦参加プレイヤの 1 セットあたりの総得点数の最良値，最悪値および 300 点越え (1 セットの全 100 試合でたえず平民でいられれば 300 点となる) の回数を示したものであるが，提案手法で学習済みのプレイヤは，300 点越えの回数こそ Blauweregen に劣っているものの，300 点に到達することができなかった場合でも僅かに 300 点に届かなかったことが多く，総得点数の最悪値も全プレイヤの中で一番高くなっていた．

この結果が示唆しているように提案手法は，堅実なプレイヤのシミュレーション方策を学習する傾向がある．これは，たとえ下位のプレイヤであってもその順位を維持もしくは上昇させるものも良好なエピソードとする選別基準をとっているためだと考えられる．

7 おわりに

近年，チェス，将棋，囲碁などの完全情報ゲームを主たる対象として，自己対戦とモンテカルロ法を組み合せ，人を凌駕するような強力なコンピュータプレイヤの制作を目指した研究が展開されている．こうした研究はポーカー，大貧民，麻雀などの不完全情報ゲームに関しても進められており，先駆的な研究の 1 つに，UEC コンピュータ大貧民大会 (UECda) を対象とした大渡らの試みがある．これはモンテカルロ法に基づく大貧民プレイヤの自己対戦棋譜を教師データとして，プレイヤの戦略に相当するシミュレーション方策を学習するものである．

本研究では，自己対戦棋譜そのものではなく，それに含まれる良好なエピソードを教師データとすることでより効果的な学習が可能になると考え，大渡らの手法の変形版を提案し，その学習性能に関する実験的考察を行った．実験の試行回数が十分とはいえず，現時点では両手法の優劣を論ずることはできないが，提案手法で学習済みのプレイヤは，先行手法による学習済みのプレイヤおよび過去の UECda 優勝プログラムとのリーグ戦で好成績をおさめるなどの興味深い実験結果を得ている．

参考文献

1) David Silver, Julian Schrittwieser, Karen Simonyan, ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas baker, Matthew Lai, Adrian bolton, Yutian chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel , Demis Hassabis:Mastering the game of Go without human knowledge Nature,Vol..550 (2017)

2) David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, Demis Hassabis:A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play Science 07 Dec 2018 Vol. 362, Issue 6419, pp. 1140-1144

3) 大渡勝己，田中哲郎：方策勾配を用いた教師あり学習によるコンピュータ大貧民の方策関数の学習とモンテカルロシミュレーションへの利用, 情報処理学会研究報告 (2016)

4) 西野哲郎:第一回 UEC コンピュータ大貧民大会 (UECda-2006) の実施報告，情報処理学会誌，Vol. 48, No.8, pp884-888(2007)

5) P.Auer, N. Casa-Bianchi , and P. Fischer ：Finite-time Anaalysis of the Multiarmed Bandit Problem. Machine Laerning, Vol,47 pp. 235-236 (2002)

6) 電気通信大学．UEC コンピュータ大貧民大会，http://www.tnlab.inf.uec.ac.jp/daihinmin/2018/ (2018.2.1 閲覧)