# コンピュータ大貧民

西野　哲朗

チェスや将棋などの完全情報 2 人ゲームにおいては，コンピュータプログラムが人間のチャンピオンに勝利するような状況になっている．一方，不完全情報多人数ゲームに関しては，いまだ強いコンピュータプログラムが実装されたとは言いがたい状況にある．例えば，カードゲームの大貧民では，相手がどのカードを持っているのかわからないし，相手が多人数であるため利害関係が複雑になる．筆者らは，モンテカルロ法を，TD 学習に基づく差分学習法に拡張して実装し，UEC コンピュータ大貧民大会において優勝した．本稿では，そのプログラムで採用したアルゴリズムについて紹介する．

キーワード：ゲーム情報学，大貧民，モンテカルロ法

## 1. はじめに

コンピュータゲーム研究は，近年，めざましい発展を遂げている．1994 年にチェッカーのコンピュータプログラムが世界チャンピオンに勝利した．さらに，1997 年には，コンピュータがオセロで世界チャンピオンを破り，チェスでは，IBM のスーパーコンピュータ Deep Blue が，当時世界チャンピオンのカスパロフに勝利した．さらに，2010 年には，将棋ソフト「あから 2010」が女流王将を破り，2012 年には，囲碁ソフト「Zen」が，4 子で男子プロに勝利した．

このようなゲームプログラムの発展においては，モンテカルロ木探索が高い効果をあげてきた [3]．完全情報 2 人ゲームの 1 つである囲碁では，Crazy Stone の登場によりプレイヤープログラムの強さが飛躍的に向上した．Crazy Stone は探索木にモンテカルロ法を用いたプレイヤープログラムである．Crazy Stone と同じように，探索木にモンテカルロ法を用いる手法を採用しているプレイヤープログラム MoGo は，囲碁のプロ棋士に勝利した．

その一方で，不完全情報ゲームにおいてもモンテカルロ木探索が有効であることがわかってきているが，その理論的な説明はなされていない．しかし，最近になって完全情報を仮定したモンテカルロ木探索 PIMC の不完全情報ゲームへの応用について，その性質の検討をするための，対象ゲームを分類する指標の提案などが行われるようになった [6]．この文献の中で不完全情報ゲームを，ゲームの進展に従って捨て札などの情報が増えることで状態の未知情報の推定ができるトリック型ゲームと，ポーカーのように最後まで明示的な情報が増えないものに分類している．

トリック型ゲームでは着手決定において，未知状態の推定を用いれば，より有効な意思決定が可能となるように思われる．しかしながら，実験的には必ずしも有効と言えないゲームもあることが報告されている [4]．モンテカルロ法は，ほかのゲームのプレイヤープログラムにも幅広く応用されている．その結果，プレイヤープログラムは人間に完全情報 2 人ゲームで勝つことができるようになったが，不完全情報多人数ゲームでは，いまだ人間に勝つようなプレイヤープログラムは開発されていない．

電気通信大学では，UEC コンピュータ大貧民大会（UECda）を，毎年 11 月末に開催している [5]．本大会ではプログラム同士の高速対戦を行うため，配布されたカードの善し悪しに左右されない，プレイのアルゴリズム本来の優劣を競うことができる．この大会を契機として，不完全情報多人数ゲームの 1 つである大貧民の研究が盛んになった．初期の大会では全探索などが行われていたが，その後，大貧民においても，乱数シミュレーションを用いたプレイヤープログラムが優秀な成績を収めてきた [2]．実際，須藤らは，モンテカルロ法と，その制御に UCB1-TUNED と呼ばれるアルゴリズムを用いた．そして，第 4 回 UEC コンピュータ大貧民大会 (UECda-2009) において優勝を収めている．須藤らが用いた UCB1-TUNED とは多腕バンディット問題 [1] を解決するためのアルゴリズムである．

一方，不完全情報多人数ゲームに関しては，強いプレイヤープログラムの実装が達成されたとは言いがた

にしの　てつろう
電気通信大学 総合情報学科
〒 182–8585 東京都調布市調布ヶ丘 1–5–1

  

い．ゲームの状態とは，自身の手札や場に出ているカードなど，ゲームが行われている状況を示す．大貧民では，相手がどのカードを持っているのかわからないし，相手が多人数であるため利害関係が複雑になる．筆者らは，一連のゲーム状態の遷移に従って手札提出を行わせるために，モンテカルロ法を，TD 学習に基づく差分学習法に拡張して実装し，UECda-2011 に参加した．決勝戦では，対戦相手の 4 つすべてがモンテカルロ法を用いたプレイヤープログラムであったが，提案手法を実装したプレイヤープログラムは UECda-2011 で優勝することができた．以下では，そのプログラムで採用したアルゴリズムについて紹介する．

## 2. コンピュータ大貧民とは

大貧民はトランプで遊ぶカードゲームの 1 つであり，「ど貧民」，「大富豪」，「階級闘争」などとも呼ばれる．カードを参加者にすべて配り，手持ちのカードを順番に場に出して早く手札をなくすことを競うゲームである．1 ゲームでの順位が次ゲーム開始時の有利不利に影響する点が特徴で，勝者をより有利にするゲーム性から大富豪との名称がついた．

地方ルールが数多く存在することも大きな特徴である．地方ルールには，一度負け出すとなかなか逆転できないという欠点を補正する方向に働くものが多い．順位は，手持ちのカードのなくなった順に，大富豪，富豪，平民，貧民，大貧民（ど貧民）となる（平民は複数存在しうるが，存在しない場合もある）．第 2 ゲーム以降は，カードを配ったあとのゲーム開始時までに，大貧民は大富豪に 2 枚，貧民は富豪に 1 枚，手持ちの最も強いカードを差し出さなければならない．このカード交換を「税金」または「献上」という．

トランプの大貧民は，日本発祥のゲームである．ルールがシンプルで多くの日本人が知っているゲームだが，その割に，奥が深く，地方ルールなどもたくさんある．おそらく必勝手がなく，名人やグランド・マスターもいないという特殊なゲームである．

UEC コンピュータ大貧民大会で採用している大貧民のルールは，以下のとおりである．

- ゲームの開始：ゲームはダイヤの 3 を持っている人から始まる．必ずしもダイヤの 3 を出さなくてもよい．
- パスについて：場のカードと手札の関係上，カードを出せない場合はパスとなる．カードが出せる場合でも戦略上パスすることができるが，いったんパスすると，場が流れるまで自分に順番が回ってくることはない．
- スペードの 3：スペードの 3 はジョーカーよりも強い．ジョーカーが 1 枚で出された場合，スペードの 3 で切ることができる．
- 場の流れ方：全員がパスしたら場が流れ，最後にカードを出した人が 場にカードがない状態からカードを出すことができる．仮に自分以外がパスしたとき，自分がカードを出すことができれば 連続してカードを出すことができる．
- 8 切り：8 を含んだ手を出した場合，場のカードがクリアされ カードを出した人が任意のカードを出すことができる（権利をとることができる）．
- 革命：同じ番号のカードを 4 枚，もしくはジョーカーを含んだ 5 枚をセットで出すと，革命が起こる．革命後はカードの強さが逆転する．
- 階段（シークエンス）：同一マークの連番が 3 枚以上ある場合は，同時に出すことができる．5 枚以上同時に出すと革命が起こる．
- しばり（ロック）：場にあるカードと同じマークのカードを出すと「しばり」状態となり，以後同じマークしか出せない．
- あがり方：どんなカードでもあがることができる．
- カードの交換：大富豪は 2 枚，カードをもらう．富豪は 1 枚．選び方は任意．強いカードをあげてもよい．大貧民は 2 枚，貧民は 1 枚強いカードを献上する．カードは自動的に選ばれ，選択できない．

本大会で使用したプログラムは，カードの配布や場の管理を行うサーバ・プログラムと，プレイヤーに対応するクライアント・プログラムから構成される．5 人のプレイヤーに対応する 5 つのクライアント・プログラムを，サーバ・プログラムにつないで対戦を行う．上記プログラムのソース・コードは，大会サイトからダウンロード可能である（図 1 参照）．

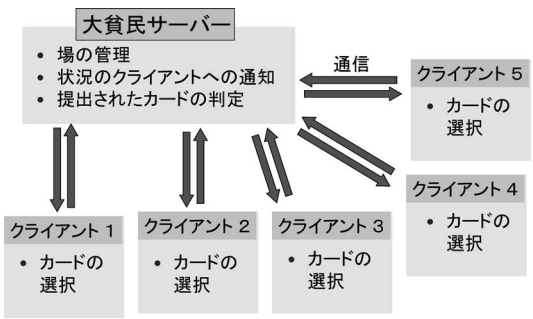


**図 1** システム構成図

## 3. 差分学習法の応用

須藤らは，不完全情報多人数ゲームである大貧民に対して，素朴な MC 法を用いたプレイヤープログラムを実装した．このプログラムは，UEC コンピュータ大貧民大会公式ウェブサイトからダウンロード可能である．そのプログラムの処理の流れは，以下のとおりである．

**1.** （合法手の列挙）自身が選択可能な合法手を列挙する．

**2.** 以下の手順を複数回行う．

- **a.** （合法手の選択）列挙した中から，1 つの合法手 $i$ を選び，その行動選択に従ってゲームの状態を遷移させる．
- **b.** （手札の配布）乱数を用いて相手にカードをランダムに割り当てる．
- **c.** （乱数によるシミュレーション）次のプレイヤーの合法手を列挙し，その内 1 つをランダムに選択する．その行動選択に従ってゲームの状態を遷移させる．自身のカードがなくなるか，相手全員がカードを提出し終え，ゲームが終了するまですべてのプレイヤーに順番に行動選択をランダムに行わせる．
- **d.** （報酬値のフィードバック）ゲームが終了したら，報酬値 $R$（順位）の値を調べる．そして，最初に選択した合法手 $i$ に報酬値を与え，合法手の評価 $X_i$ を更新する（式 (2), (3), (4) 参照）．

**3.** （報酬値の比較）複数回のプレイアウトを行ったあと，各合法手 $i$ に対する $X_i$ の大小を比較し，$X_i$ が最大の合法手を最善手と推定する．

このように，最終状態における報酬値のみを用いて，現在の盤面における合法手の評価値が決定されている．どの合法手に対してプレイアウトを行うのかについては UCB1-TUNED を用いており，合法手の評価値と選択回数から最善手である確率が最も高いと判断された合法手を優先して選択している．

### 3.1 提案アルゴリズム

須藤らが用いた MC 法は，プレイアウトの途中における盤面を評価せず，報酬値のフィードバックを行う．したがって，プレイアウトに選択した行動選択 $i$ の評価の更新値 $V$ は，報酬値 $R$ を用いて，式 (1) として表される．

$$V = R \tag{1}$$

そして，以下の式 (2)，(3)，(4) を用いて，最初に選択した合法手 $i$ に関して，総報酬値 $X_i$，選択した回数 $n_i$，平均報酬値 $X_i$ の更新を行う．

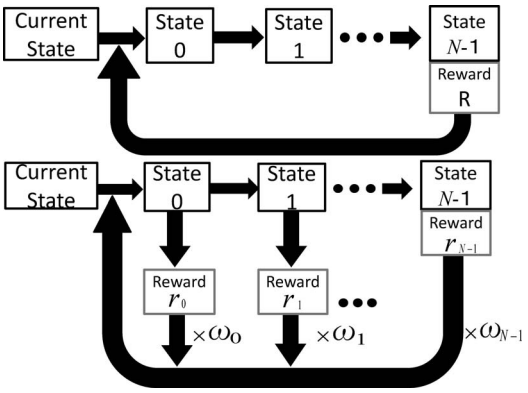


**図 2** MC 法と提案アルゴリズムの比較図

$$X_i \leftarrow X_i + V \tag{2}$$

$$n_i \leftarrow n_i + 1 \tag{3}$$

$$\overline{X_i} \leftarrow X_i / n_i \tag{4}$$

このように, MC 法では，あらゆるゲームの盤面において，ゲームの結果のみを考慮して行動選択を行う．ここで，途中経過も含めて強化学習させることにより, MC 法よりも適切な行動を選択するプログラムを実装できるのではないかと考えた．そこで, TD 学習に基づき, 式 (1) を式 (5) のように拡張し，報酬値のフィードバックについて改良した．

$$V = \sum_{t=0}^{N-1} r_t \omega_t \tag{5}$$

すなわち，プレイアウト中に生成された盤面状態 $t$ において，盤面の評価値 $r_t$ を算出する．そして，その盤面における評価の重み $\omega_t$ に従い，最初に選択した行動選択 $i$ に逐次的なフィードバックを行わせる．プレイアウトが $N$ 回の状態遷移で表されるとき，$r_{N-1} = R$ である．大貧民では，相手の手番では行動選択を行えないため，自身の手番が回ってくるまでを 1 ターンとしてフィードバックさせた．

次に，$r_t$ や $\omega_t$ をどのように求めるかが問題となる．大貧民は多人数で行われるゲームであるため，どこで報酬値の差分が発生したのかを知ることが可能である．そこで，各盤面における“ゲームに参加している人数”を評価値として用いた．これによりプレイアウトがどのような状態遷移で構成されているかを知ることができる．例えば，“5 人が競り合い惜しくも 1 位を逃がして 2 位であがったプレイアウト”では，評価値が

  

“5 → 5 → ··· → 5 → 4” という遷移となる．一方，“1 人にいち早くあがられてしまったあと，4 人で競い合い 2 位であがったプレイアウト”では，評価値が “5 → 4 → 4 → ··· → 4 → 4” という遷移となる．

MC 法では，これらのプレイアウトは同じ報酬値となる．しかし，前者のプレイアウトは何らかの要因があれば，自身が 1 位であががれる可能性があるが，後者のプレイアウトは 1 位であがれる可能性は低い．そこで前者が後者よりも優れたプレイアウトであるようフィードバックさせるために式 (6) を考案した．

$$\begin{aligned} V(0) &= r_0, \\ V(t) &= V(t-1) + \alpha(r_t - V(t-1)) \\ &= (1-\alpha)V(t-1) + \alpha r_t \quad (t>0) \end{aligned} \tag{6}$$

これにより，ターン $t$ までのプレイアウトの評価値 $V(t)$ を，ターン $t-1$ までの評価値 $V(t-1)$ とそのターンの評価値 $r_t$ を用いて求める．そのため，プレイアウトにおける状態遷移回数 $N$ によらず，1 つのパラメータ $\alpha$ で各ターンにおける評価の重み $\omega_t$ を制御できる．

以下に，$\alpha$ による $\omega_t$ の導出を示す．

$$\begin{aligned} V(N-1) &= (1-\alpha)V(N-2) + \alpha r_{N-1} \\ &= (1-\alpha)\{(1-\alpha)V(N-3) + \alpha r_{N-2}\} \\ &\quad + \alpha r_{N-1} \\ &= (1-\alpha)\{(1-\alpha)\{(1-\alpha)V(N-4) \\ &\quad + \alpha r_{N-3}\} + \alpha r_{N-2}\} + \alpha r_{N-1} \\ &\cdots \\ &= \sum_{t=0}^{N-1} r_t \omega_t \end{aligned} \tag{7}$$

そして，係数を比較することによりステップ $t$ における評価値の重み $\omega_t$ は，$\alpha \neq 1$ のとき式 (8), (9) で表される．

$$\omega_0 = (1-\alpha)^{N-1} \quad (t=0, \alpha \neq 1) \tag{8}$$

$$\omega_t = \alpha(1-\alpha)^{N-1-t} \quad (t>0, \alpha \neq 1) \tag{9}$$

また，$\alpha = 1$ のとき式 (10), (11) で表される．

$$\omega_t = 0 \quad (0 \leq t < N-1, \alpha = 1) \tag{10}$$

$$\omega_{N-1} = 1 \quad (t = N-1, \alpha = 1) \tag{11}$$

以上より，$\alpha = 1$ とすると MC 法と同じように終端状態のみを学習し，$\alpha$ の値を 1 より小さくするほど，遷移回数が少ない非終端状態の評価の重み $\omega_t$ を大きくさせて，プレイアウトにおける一連の状態遷移を学習する．

### 3.2 実験結果

提案アルゴリズムを実装したプレイヤープログラム “Crow” で，第 6 回 UEC コンピュータ大貧民大会に出場し優勝することができた．このとき，“Crow” 以外の決勝に残った 4 つのプログラムは，須藤らのプログラムを改良した，MC 法を用いるプレイヤープログラムであった．

“Crow” の強さを検証するため，対戦による実験を行った．まずはじめに，“Crow” のパラメータ $\alpha$ を MC 法を用いた前大会優勝プログラム “snowl” と同じ動作を行うよう $\alpha = 1.0$ と設定し，“Crow” と “snowl” の強さを比較した．そこで，1,000 ゲームの対戦を行わせる実験を 100 回行った．また，5 つのクライアントで対戦が行われることから前々大会優勝クライアントを 3 つを用いた．比較のため，昨年の優勝プログラムのプレイアウト回数に合わせて，5 つすべてのプログラムのプレイアウト回数を 2,000 回とした．その結果，“Crow” の 49 勝 50 敗 1 引き分けの結果となった．これにより，提案手法が MC 法と同じ動作を行っていることが示唆された．

次に，“Crow” をプレイアウトにおける途中経過を含めて強化学習を行うようパラメータの値を $\alpha = 0.9$ として，同様の対戦させる実験を行った．この結果，“Crow” の 54 勝 46 敗となった．この結果より，差分学習法を用いた提案アルゴリズムが有効であることが示唆された．

## 4. おわりに

不完全情報多人数ゲームである大貧民に対して，TD 学習の考えに基づき，差分をとり最善手を推定するアルゴリズムを導入した．このアルゴリズムは従来手法である MC 法より強いプレイヤープログラムとなり得ることが示された．また実験より，これらのアルゴリズムは手札枚数が少なく，プレイアウトが短かい状態遷移で表されるにもかかわらず，異なる合法手を最善手としたため，根本的な差違があることが確認されている．この差違と “大貧民というゲームにおける強さ” との関連性を示すことは今後の課題である．

### 参考文献


[1] P. Auer, N. Cesa-Bianchi and P. Fischer, Finite-time Analysis of the Multiarmed Bandit Problem, *Machine Learning*, **47** 235–256, 2002.

[2] 小沼啓，西野哲朗，コンピュータ大貧民に対するモンテカルロ法の適用，情報処理学会 第 25 回ゲーム情報学研究会資料集，2011.

[3] 美添一樹，モンテカルロ木探索—コンピュータ囲碁に革命を起こした新手法，情報処理，**49**(6), 686–693, 2008.
[4] 西野順二，西野哲朗，大貧民における相手手札推定，情報処理学会研究報告 2011–MPS–85，**9**, 2011.
[5] 西野哲朗，第 1 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，**48**(8), 884–888, 2007.
[6] J. Long, N. R. Sturtevant, M. Buro and T. Furtak, Understanding the Success of Perfect Information Monte Carlo Sampling in Game Tree Search, Proceedings of the 24th. AAAI Conf., AAAI, 134–140, 2010.

  