大貧民における学習を用いた合理的なプレイアウトによる指し手の探索手法

徳永 和真[2,a)]　長谷部 浩二[2,b)]

概要：本研究は，モンテカルロ法によるゲームの指し手の探索におけるプレイアウトの精度を，学習によって得られたモデルを用いて向上させる方法を提案する．より具体的には，プレイアウトにおける指し手の選択と相手の手札の推測のそれぞれを，CNN（Convolutional Neural Network）と LSTM（Long Short-Term Memory）によって得られたモデルをもとに行う．特にここでは，多人数不完全情報ゲームの一種である大貧民と呼ばれるトランプゲームを対象とする．

Search Method for Playing Daihinmin with Rational Playout Based on Machine Learning

Kazuma Tokunaga[2,a)]　Koji Hasebe[2,b)]

Abstract: We propose a method to improve the accuracy of playout in the Monte Carlo method with models obtained by machine learning. More specifically, the choice of move and the estimation of the opponent's hand in a playout are performed based on the models obtained by CNN (Convolutional Neural Network) and LSTM (Long Short-Term Memory). Here we focus on a card game called Daihinmin, which is a kind of multi-player incomplete information game.

1. 研究の背景と目的

近年，ゲームをプレイするコンピュータプログラムの研究開発が盛んに行われている．一般に，ゲームは完全情報ゲームと不完全情報ゲームの 2 つに大別される．前者は，各プレイヤーが盤面の情報をすべて知ることができるゲームであり，例として囲碁や将棋などが挙げられる．一方後者は，盤面の一部の情報がプレイヤーに対して隠されているようなゲームであり，例としてポーカーや麻雀などが挙げられる．

完全情報ゲームをプレイするプログラムについては近年急速にレベルが向上している．例えば AlphaZero[4] と呼ばれるプログラムは，同一のアルゴリズムを用いて囲碁や将棋，チェスのすべてにおいて人間のプロのプレイヤーに勝利するまでに至っている．一方，不完全情報ゲームついては，ポーカーをプレイする Pluribus[1] や麻雀をプレイする Suphx[2] などのプログラムが提案されているものの，対象とするゲームに特有の知識を多く用いて作成されており，未だ汎用的なアルゴリズムが開発されているとは言い難い．その主な理由としては，多人数でプレイする不完全情報ゲームでは 1 つの盤面に対する情報集合が大きいため，その盤面の最適な戦略を見つけることが困難であることが考えられる．

ゲームをプレイするプログラムを構築する際，プレイヤーの指し手の探索のためにモンテカルロ法と呼ばれる手法がしばしば用いられてきた．モンテカルロ法は，指し手をランダムに選択しながらシミュレーションを行い（これをプレイアウトと呼ぶ），その結果が勝利であれば選択し

[1] 情報処理学会
IPSJ, Chiyoda, Tokyo 101–0062, Japan
[2] 筑波大学情報学群情報科学類
College of Information Science, University of Tsukuba
[3] 筑波大学システム情報系
Faculty of Engineering, Information and Systems, University of Tsukuba
[a)] s1811368@s.tsukuba.ac.jp
[b)] hasebe@cs.tsukuba.ac.jp

た指し手に対して正の評価を与え，そうでなければ負の評価を与えるというものである．モンテカルロ法を不完全情報ゲームに適用する際には，可能な盤面を 1 つランダムに選択した上でプレイアウトを行うのが最も基本的な方法である．

こうしたモンテカルロ法に関する先行研究として，プレイアウトの方法を変更することにより探索の精度を向上させる試みがある．その基本的なアイデアは，プレイアウトにおける指し手と盤面の 2 つの選択を，ランダムではない方法に変更するというものである．前者については，プレイアウトの試行回数を多くすることでランダムな行動による影響を小さくする方法がある．また，有望な指し手をより多く探索するために，指し手の暫定的な評価値の見積もりを使用した UCT（Upper Confidence bounds applied to Trees）アルゴリズムなどが用いられたものもある．一方後者については，ランダムに固定した盤面をいくつも用意しそれぞれにおいて探索することで，盤面に依存した評価の影響を小さくする方法がある．また，盤面の固定にはプレイヤーのこれまでの行動履歴を使用して推測をすることで，その精度を向上させる試み [5] もある．しかしながら，学習を用いてプレイアウトの精度を向上させるアプローチについては未だ十分に検討されていない．

そこで本研究では，モンテカルロ法におけるプレイアウト中の行動の選択および盤面の推測を，学習によって作られたモデルをもとに行う方法を提案する．これにより，プレイアウトの精度を向上させて合理的な結果を導き，指し手の評価の精度を向上させることができると考えられる．特にここでは，多人数不完全情報ゲームの一種である大貧民と呼ばれるトランプゲームを対象とする．大貧民は，1 つの盤面に対する情報集合やゲームの平均的なプレイの長さがある程度大きいことから，推測による指し手の評価への影響を検証するために適したゲームと考えられる．このようなアプローチに関して，吉原らの研究 [7] や西野らの研究 [6] では推測の有効性の低いことが実験的に示されている．しかし，これらの結果はプレイアウト中の行動がランダムに決定されていることが原因であるとも考えられる．提案手法のような合理的なプレイアウトを導入することにより，推測の有効性の向上が期待できる．

2. 他のプレイヤーの手札の推測を利用した指し手の探索手法

提案手法では，モンテカルロ法のプレイアウトにおける指し手の選択と盤面の推測のそれぞれをランダムに行う代わりに，学習によって得られたモデルを用いて行う．以下で，これら 2 つのモデルの具体的な作成方法について述べる．

まず，プレイアウト中に合理的な（すなわち勝利に貢献する可能性が高いと考えられる）指し手を選択するために，各プレイヤーが場に提出するカードを選択するモデルをニューラルネットワークを用いて作成する．このモデルは，カード提出後の自分の手札や他のプレイヤーの残りのカード集合，場の状態を表す特徴量を入力として，その行動を選択した結果の状態価値を出力とする．これにより，各合法手のうちその状態価値が最も高いものを最良の指し手として選択する．入力データに，カードの集合をマークと数字からなる行列として表現していることから，局所的な特徴を見つけることができる CNN（Convolutional Neural Network）を用いてモデルを構成する．また，場の状態を表す特徴量には，場のカードのマークや数字の他にも，各プレイヤーの残りの手札の枚数やパスをしているかなどの情報を使用する．これらの特徴量を表す 1 つのベクトルと，自分と他のプレイヤーのカードの集合からなる多入力のモデルを構築する．出力の正解データには，モンテカルロ法による探索で求められた各合法手の評価値を用いてモデルの学習を行う．

次に，他のプレイヤーの手札の推測をするモデルを作成する．このモデルは，推測の対象とするプレイヤーのこれまでのカード提出履歴から残りの手役の確率分布を求める．入力データに時系列データを使用することから，神田らの研究 [5] でも有効性が示されている LSTM（Long Short-Term Memory）[3] を用いてモデルを構成する．出力の正解データは実際の残りの手役を使用する．しかし，これには行動履歴と直接関係しないような残りのカードも含まれており，学習の妨げになると考えられる．そこで，学習の効率化を図るためにルールベースを用いた正解データの重み付けを行う．すなわち，カード提出履歴から推測されるべきカードにはルールを用いてその残存確率に重みを付加する．また，このモデルのカードの集合の表現には手役を用いている．大貧民における手役は 1 枚出しや 2 枚出し，階段などがある．これは，入力データにカード提出履歴を使用していることから，一度の手番で提出ができるカードの組合せが推測されるものとして妥当であると考えられるからである．

以上の 2 つのモデルを用いて，指し手の探索アルゴリズムを作成する．まず，推測モデルを用いて他のプレイヤーの手札を仮定する．これに基づきプレイアウトを開始し，行動決定のモデルを用いて終局まで進める．しかし，モデルの出力をそのまま使用すると，プレイアウトが決定的になってしまい複数回探索をする効果がない．そこで，一定確率の下でランダムに他のプレイヤーの手札の仮定や行動を変化させることで，プレイアウトが決定的ではなく尤度の高いものを複数作成し，評価値の精度の向上を図る．その上で，探索の結果最も評価の高い指し手を選択する．

3. 結論と今後の課題

本研究では，大貧民において学習させたモデルを用いて

合理的なプレイアウトを行う指し手の探索手法を提案した．より具体的には，教師あり学習によって他のプレイヤーの手札の推測と行動の選択を決定する 2 つのモデルを作成し，それらを用いてプレイアウトの精度を高め，指し手の評価の精度を向上させることを目指した．

今後は，以上で述べた指し手の探索手法の精度を確かめるための評価実験を行うことを計画している．そのために，本研究では従来のモンテカルロ法との指し手の評価の精度を，以下の 3 つの方法で比較する．1 つ目に，同じ時間内で探索したときの評価値を比較する．提案手法では，推測や行動決定に学習させたモデルを用いているため，それぞれの探索で合理的なプレイアウトができる反面，ランダムなものに比べ 1 回の探索に時間がかかる．それらの点のトレードオフを調べるため時間制限を設けて探索をすることにより検証する．2 つ目に，最も良い評価値と次点の評価値の差を計測する．従来の手法に比べ探索の過程でランダム性が小さいことにより，各評価値に有意な差が表れるかを検証する．最後に，従来の手法と対戦を行う．大貧民をプレイするプログラムとして強化されたかを検証する．

以上の評価を行なった上での本研究のさらなる発展としては，大貧民に比べ 1 つの盤面に対する情報集合が大きい麻雀への適用が考えられる．麻雀は相手の手牌の推測を用いた打牌選択など，大貧民と似ている部分が多いと言える．こうしたゲームに対しても提案手法が有効かを検証したいと考えている．

参考文献

[1] N. Brown and T. Sandholm: Superhuman AI for multiplayer poker. *Science*, Vol.365, Issue 6456, pp.885-890, 2019.

[2] J. Li et al.: Suphx: Mastering Mahjong with Deep Reinforcement Learning. *arXiv preprint* arXiv:2003.13590, 2020.

[3] S. Hochreiter and J. Schmidhuber: Long short-term memory. *Neural computation*, Vol.9, Issue 8, pp.1735-1780, 1997.

[4] D. Silver et al.: A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, Vol.362, Issue 6419, pp.1140-1144, 2018.

[5] 神田直樹, 伊藤毅志: コンピュータ大貧民における LSTM を用いた手札推定. 情報処理学会研究報告, 2018-GI-39, Vol.2018, No.8, pp.1-8 (2018).

[6] 西野順二, 西野哲郎. 大貧民における相手の手札推定. 情報処理学会研究報告, 2011-MPS-85, Vol.2011, No.9, pp.1-6 (2011).

[7] 吉原大夢, 大久保誠也. コンピュータ際貧民における手札推定の有効性について. 情報処理学会研究報告, 2013-GI-30, Vol.2013, No.4, pp.1-6 (2013).