6P-05

単純化されたUNOのプレイヤAIの強化学習による構築

福島 健介 †　　鈴木 徹也 ‡
芝浦工業大学 大学院 理工学研究科 †　　芝浦工業大学 大学院 理工学研究科 ‡

1　はじめに

多人数不完全情報ゲーム UNO は手札の増減や特殊なカードの存在など, 同じく多人数不完全情報ゲームである麻雀やポーカーとは異なる特徴をもっている. しかしゲーム AI の研究において UNO は麻雀やポーカーほど研究されていない.

このような背景から, 本研究では UNO の強力なプレイヤ AI を構築することを目的とする.

そこで本研究はまず UNO の特徴を維持したまま単純化したゲームとして miniUNO を考案した. そして miniUNO プレイヤ AI の構築に, 他の不完全情報ゲームで成果を上げている強化学習を利用した. さらにその構築したプレイヤ AI を対戦によって評価, 考察した.

2　多人数不完全情報ゲーム UNO

UNO は多人数不完全情報ゲームのカードゲームである. 配られた 7 枚のカードを最初にすべて場に出したプレイヤが勝者となり他のプレイヤから得点を得る. UNO は 76 枚の数字カード (各 4 色あり, 0 は 1 枚, 1-9 は 2 枚) と特殊なカードとしてリバース, スキップ, ドロー 2 の 3 種が各色 2 枚, ワイルド, ワイルド・ドロー 4 の 2 種が各 4 枚の合計 108 枚を 1 セットとしてゲームに使用する. UNO には多くのローカルルールが存在し, 公式とは異なるルールでプレイされる場合が多い.

本研究では AI プレイヤ同士の対戦を前提としているために以下のルールを採用していない.

- 手札が最後の一枚になる際に, 他プレイヤより先に「ウノ」と宣言しなければならない.
- 他に出せるカードがあるのにワイルド・ドロー 4 を使用したと考えた場合, チャレンジを行うことができる.

本研究では, プレイヤ AI の効果を調査するために公式ルールの他に一般に遊ばれているルール (カスタムルール) を利用する. カスタムルールは, 公式ルールに「ドロー 2 とワイルド・ドロー 4 の効果の先送り」というルールを追加したルールである

3　関連研究

桑原ら [1] は多人数不完全情報ゲームである大貧民に対して状態価値関数の推定を行いプレイヤ AI を作成した. 状態価値関数の推定を 2 通りの方法で行い, それぞれプレイヤ AI を作成した. 1 つは強力なプレイヤ AI の対戦ログからの教師あり学習であり, もう 1 つは自己対戦による強化学習である. その結果, どちらも強力なプレイヤ AI に準ずる実力のプレイヤを作成できた.

清水ら [2] は多人数不完全情報ゲームである麻雀に対して深層強化学習を適用してプレイヤ AI を作成した. この時, ゲームの複雑さや規模から大量の計算資源を必要とするために麻雀を単純化したすずめ雀で研究を行った. その結果, 自身の手牌のみから行動する 1 人すずめ雀プレイヤを超える実力にはならなかったが, 表現力の高いモデルを使用することで 1 人すずめ雀プレイヤに準ずる実力のプレイヤを作成できた.

4　提案手法

UNO のプレイヤ AI 作成の第一歩として本研究では単純化した UNO である miniUNO に対してプレイヤ AI を作成する.

4.1　miniUNO

UNO を単純化するために以下の変更を行った.

- 色数を 4 色から 3 色にする
- ワイルド, ワイルドドロー 4 の枚数を 4 枚から 3 枚
- 初期手札枚数を 7 枚から 6 枚

変更の結果 miniUNO では 57 枚の数字カードと 24 枚の特殊カードを使用する.

Construction of AI players for a Simplified UNO Card Game Using Reinforcement Learning
† Fukushima Kensuke, Graduate School of Engineering and Science, Shibaura Institute of Technology
‡ Suzuki Tetsuya, Graduate School of Engineering and Science, Shibaura Institute of Technology

4.2 AI プレイヤ

本研究では, 3 つの AI プレイヤを提案する.

P_{rl} 強化学習で作成したプレイヤ

P_{ml} 対戦ログからの教師あり学習で作成したプレイヤ

P_{mrl} P_{ml} にさらに強化学習を適用したプレイヤ

提案プレイヤの効果を比較するためにヒューリスティックプレイヤ P_H を利用する.

強化学習のアルゴリズムとして先行研究 [2] で使用されている PPO[3] を使用し, 学習モデルにニューラルネットワークを使用する深層強化学習によって学習を行った. 学習モデルの構成は中間層が 2 層の全結合モデルを使用した. 中間層のノード数や学習 step 数については実験により決定した.

強化学習は順位と点数を報酬として 3 つの P_H との対戦の中で報酬を最大化する方策を学習した.

教師あり学習では 4 つの P_H の対戦ログを利用し, P_H の行動を模倣できるように学習を行った. そのため, 教師あり学習で作成した P_{ml} は学習元の P_H と同じ戦略を学習すると予想される. また, 強化学習で使用したのと同じ学習モデルを教師あり学習でも使用した.

5 実験

提案プレイヤと 3 つの P_H との対戦を 1 万回行いその平均で効果の評価を行う. また, UNO の 4 人対戦では平均得点が 0 , 平均順位が 2.5 位となる. 表中のミス率はプレイヤの全行動の中でルール上正しくない行動をした回数の割合である.

5.1 実験 1 公式ルール

初めに公式ルールに近いルールで対戦を行いプレイヤを比較する. 実験結果を表 1 に示す.

表 1 ヒューリスティックプレイヤとの対戦

	P_{rl}	P_{ml}	P_{mrl}
平均点数	3.151	0.058	0.663
平均順位	2.466	2.481	2.533
ミス率	1.828	0.000	1.062

表 1 から P_{rl} は P_H に対して優位性を持っているといえる. P_{ml} と P_{mrl} は P_H と同等程度の実力しかないが P_{rl} と比較するとミス率が低い. これは対戦ログからの学習によりルールに沿った行動ができるが, 勝利に近づく戦略の学習が足りないからだと考えられる.

5.2 実験 2 カスタムルール

次にローカルルールを取り入れたカスタムルールで対戦を行いプレイヤを比較する. 実験結果を表 2 に示す.

表 2 ヒューリスティックプレイヤとの対戦

	P_{rl}	P_{ml}	P_{mrl}
平均点数	7.475	0.187	4.835
平均順位	2.402	2.487	2.441
ミス率	2.093	0.003	1.707

表 1 から P_{rl} と P_{mrl} は P_H に対して優位性を持っているといえる. 表 1 と表 2 を比較すると強化学習を利用した P_{rl} と P_{mrl} は公式ルールよりもカスタムルールでの実験で高い順位と点数を出しているがミス率は悪くなっている. これは, ルールが複雑になったことによりヒューリスティックな戦略よりも強化学習が有利になっているが, その反面で複雑なルールを学習しきれていないためだと考えられる.

6 終わりに

本研究では, 教師あり学習, 強化学習, そしてそれら 2 つを合わせた手法の 3 通りで単純化した UNO の AI プレイヤを作成しそれらの効果を実験により調査した. その結果, 強化学習で作成したプレイヤが最も強いがルールの学習においては教師あり学習の適用が効果的であることを示した. また, 2 つの学習手法を合わせることで中間的なプレイヤを作成することができた.

今後の課題として, UNO の特性に適した学習モデルを考案することが挙げられる. 例えば UNO では 相手プレイヤの行動から手札を推定することが重要なため行動履歴も扱えるモデルに変更することが考えられる.

参考文献

[1] 桑原 和人, 保木 邦仁. 大貧民の期待順位 (状態価値) の強化学習. 研究報告ゲーム情報学（GI), 2018-GI-39(7), 1-8 (2018-02-23)

[2] 清水 大志, 田中 哲朗. 深層強化学習を用いた麻雀プレイヤの構築. ゲームプログラミングワークショップ 2020 論文集, pp.147-154, 2020-11-06

[3] Schulman, John and Wolski, Filip and Dhariwal, Prafulla and Radford, Alec and Klimov, Oleg. Proximal Policy Optimization Algorithms. arXiv, preprint arXiv:1707.06347, (2017)