深層モンテカルロ法による UNO プレイヤに関する研究

1250329　塩澤 康志　【 高度プログラミング研究室 】

1　はじめに

近年，2048 や将棋などのゲームにおいて，人間のプレイヤを凌駕する AI が次々と登場している．一方で，不完全情報ゲームであるカードゲームにおいては，依然として多くの課題が残されている．UNO においては，状態空間は非常に大きく，報酬の発生頻度が低いため，強化学習において効果的な学習が難しく，まだ改善の余地がある．本研究では，カードゲーム「UNO」において，深層モンテカルロ法（Deep Monte Carlo, DMC）によるプレイヤを評価する．

2　深層モンテカルロ法

従来のモンテカルロ法は，エピソード全体の情報を利用して行動価値関数 $Q(s,a)$ を学習する強化学習手法であり，以下の手順で方策 π を最適化する．

1. π を用いてエピソードを生成する．
2. 状態 s, 行動 a について，$Q(s,a)$ を計算し，更新する．
3. エピソード中の各状態 s に対して，
   $\pi(s) \leftarrow argmax_a Q(s,a)$ によって方策を更新する．

深層モンテカルロ法を実現するため，ステップ 1 では ε-グリーディ法を用いる．また，ステップ 2 では割引累積報酬を適用し，長期的な報酬を考慮した学習を行う．従来の Q 学習では，$Q(s,a)$ は状態と行動の組み合わせごとにテーブルとして保持されるが，状態空間が大きくなると，この方法では学習が困難となる．そこで，深層モンテカルロ法では，テーブルを直接更新する代わりに，平均二乗誤差を最小化することでニューラルネットワーク (Q-Network) を更新する．

3　実験

本研究では，UNO における AI プレイヤの学習と評価を行うため，公式ルール [2] に基づいた 2 対 2 のチーム戦形式を採用した．また，ルールベースプレイヤを基準として設定し，その戦略に基づいたデータを用いて DMC プレイヤの学習を行った．本研究におけるルールベースプレイヤは先行研究 [1] と同様に以下のような戦略をとる．

- カードを出すことができる場合，カードを引かずに出すことのできるカードを出す．
- 手札のカードにおける色や数字，記号のうち，最も多いものを優先して出す．
- ワイルドカードを出すことができる場合，ある確率で選択する．

初めに，ルールベースプレイヤに 100 万ゲームの対戦をさせ，そのデータを Q-Network を用いる DMC プレイヤに学習させた．その後，学習済み DMC プレイヤ 2 体とルールベースプレイヤ 2 体を対戦させた．

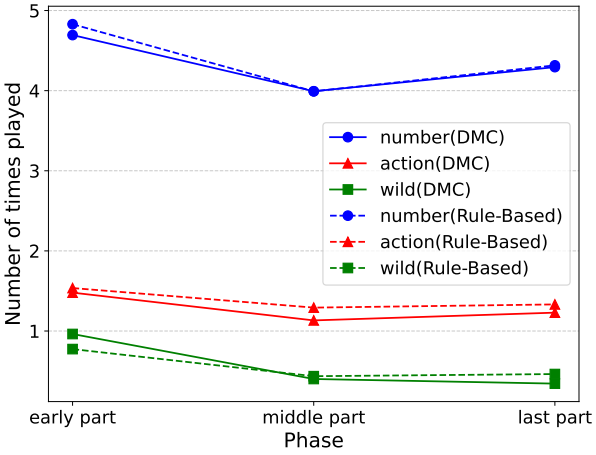

図 1　場面別の各カードが出された平均枚数

4　実験結果

ルールベースプレイヤとの 200 万ゲームにおける DMC プレイヤの平均勝率は約 45.12%であった．図 1 は，序盤・中盤・終盤におけるそれぞれのチームが出した各カードの平均枚数を示したものである．どちらのプレイヤも同じような変化をしていることが分かる．ルールベースプレイヤと比較すると，DMC プレイヤは記号カードを温存しており，ワイルドカードを序盤に出していることが分かる．また，候補のカードの中に異なるスコアのカードが複数存在するとき，ルールベースプレイヤ約 25.34%，DMC プレイヤは約 25.54%の割合でスコアが最大であるカードを出していた．

5　まとめ

本研究では，UNO における DMC プレイヤの評価を行った．結果として，DMC プレイヤはルールベースプレイヤとの対戦における勝率は上がらなかったものの，ルールベースプレイヤと同じような戦略を取ることができることが分かった．学習の第 2 段階として強化学習を行うことで改善がみられるか確認することは今後の課題である．

参考文献

[1] X. Yang, X. Liu, W. Lin, “Multi-DMC: Deep Monte-Carlo with Multi-Stage Learning in the Card Game UNO”, 2024 IEEE Conference on Games (CoG), 2024.

[2] Mattel, “ウノ | Mattel Games マテル ゲーム”, https://mattel.co.jp/toys/mattel_games/mattel_games-10936/#howToPlay, 2025 年 2 月 1 日閲覧.