強力なプレイヤとの対戦を通じた Deep Monte Carlo 型強化学習の大貧民 AI の構築

浅島　柊生　【 高度プログラミング研究室 】

1　はじめに

近年，強化学習を用いたゲーム AI の研究が盛んに行われており，自己対戦によって方策を改善する手法が高い性能を示している. 特にモンテカルロ法とニューラルネットワークを組み合わせた手法は，複雑な状態空間を持つゲームにおいて有効であることが知られている. [1]

一方，自己対戦型強化学習では，学習初期における方策の弱さが学習の不安定性を引き起こし，十分な性能向上が得られない場合がある. そのため，学習初期にどのような対戦相手を用いるかは，学習効率や最終的な性能に影響を与える重要な要因であると考えられる.

本研究では，大貧民を対象として，実対戦によって生成されたエピソードの最終結果を教師信号として学習を行う Deep Monte Carlo 型強化学習を扱う. 特に，学習初期に用いる対戦相手の強さの違いが学習過程および最終性能に与える影響に着目し，強力な既存プレイヤを用いる場合と，ルールベースプレイヤを用いる場合の比較を行う.

2　Deep Monte Carlo 型強化学習

本研究で用いる Deep Monte Carlo 型強化学習は，方策に従って生成された対戦エピソードを最後まで実行し，ゲーム終了時の最終順位を報酬としてニューラルネットワークが行動の評価値を学習する手法である. 本手法はブートストラップを用いず，エピソード完結後の結果に基づいて価値を推定する点に特徴がある.

3　実験方法

本研究では，Deep Monte Carlo 型強化学習を用いた大貧民 AI を構築し，学習初期に用いる対戦相手の違いが学習過程および性能に与える影響を検討する．比較対象として，学習初期に強力な既存プレイヤである Blauweregen を対戦相手とするモデルと，デフォルトのルールベースプレイヤを対戦相手とするモデルの 2 種類を用意した．各モデルに対し，学習初期段階を含めて合計 120000 試合の対戦を通じて学習を行った.

学習過程の評価として，5000 試合ごとに 1000 試合の性能評価を実施し，評価時の対戦相手には，それぞれ学習初期に用いたプレイヤ（Blauweregen またはルールベースプレイヤ）を用いた．また，20000 試合ごとに自己対戦の割合を段階的に増加させることで，学習初期に用いる対戦相手の違いが学習効率および最終的な性能に与える影響を比較・評価した．両モデルにおいてニューラルネットワーク構造，損失関数および最適化手法は共通とし，学習に用いる対戦相手のみを変更した.

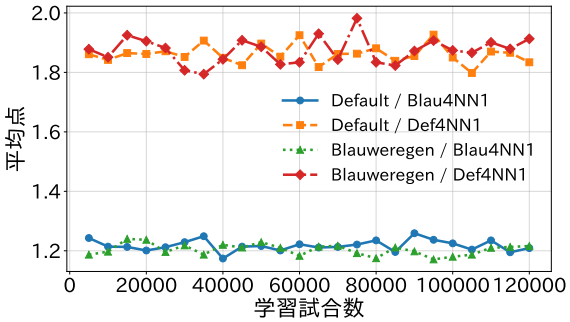

図 1: 学習後の平均点 (凡例は，学習 / 性能評価に用いた対戦相手)

4　実験結果

Blauweregen を教師として学習を行った場合と，ルールベースプレイヤを教師として学習を行った場合について，学習後のプレイヤの平均順位を図 1 に示す.

図 1 に示すように，Blauweregen を教師とした場合およびルールベースプレイヤを教師とした場合のいずれにおいても，学習初期から学習終了までの平均順位に大きな改善は確認されなかった．本研究の結果から，学習初期に強力なプレイヤを教師として用いた場合であっても，価値関数のみを用いた自己対戦型学習では，戦略的な大貧民 AI の構築には十分ではないことが示唆された.

5　まとめ

本研究では，大貧民を対象として，Deep Monte Carlo 型強化学習における学習初期の対戦相手の強さが学習に与える影響について検討した. 強力な既存プレイヤである Blauweregen を教師として学習を行う場合と，ルールベースプレイヤを教師として学習を行う場合を比較した結果，いずれの条件においても平均順位の大きな改善は確認されなかった．この結果から，学習初期に強力なプレイヤを教師として用いた場合であっても，価値関数単体に基づく自己対戦型学習では，戦略的な大貧民 AI の構築には十分ではない可能性が示唆された. 今後の課題として，方策学習や探索機構の導入，評価指標の多様化などを通じて，より戦略的な行動獲得手法について検討する必要がある.

参考文献

[1] D. Zha, K. Li, Y. Cao, H. Xiong, Y. Wang, and J. Zhang, “DouZero: Mastering DouDizhu with Self-Play Deep Reinforcement Learning.” ICML 2021: 12333–12344, 2021.