論 文 の 内 容 の 要 旨

論文題目　　報酬が疎なゲームにおける強化学習を用いたプレイヤの構築

氏　　名　　水上　直紀

コンピュータゲームの研究は人工知能のテストベッドとして発展してきた.
近年では将棋や囲碁, ポーカーなどは人間のトップと同等以上の性能を出している.
ゲーム AI はルールとその特性から現実世界の問題を解決する方法の一つと考えれている.
政治や経済などといった現実世界の問題ではゲームにおける報酬がめったに観測されない疎な環境であることが多い.
そこで本研究ではこの報酬が疎であるゲームを対象に強いコンピュータプレイヤを開発するための手法を提案する.
ゲームとしては Atari と麻雀を対象とした.
強化学習や教師あり学習の手法を組み合わせることで提案手法の多くについてその有効性を示した.

多くの Atari では最近の深層強化学習の手法は人間を超える性能を出している.
しかしながら本研究の対象とする報酬が疎のゲームでは人間にはるかに及ばない.
一つの解決方法としてエージェントの学習時に明示的で賢い探索戦略を組み込むことである.
本研究では学習の進行具合を明示的に考慮した Upper Confidence Bounds (UCB)を用いた探索ボーナスを用いる効果的な探索戦略を提案する.
さらに提案手法では探索ボーナスを報酬から分離することで本来の目的関数の学習を妨害を回避する仕組みも含まれている.
これにより報酬が疎なゲームにおいてベースラインとなる asynchronous advantage actor-critic よりも性能が向上した.

麻雀は Atari と異なる性質を持つゲームではあるが同じ報酬が疎なゲームである.
本研究では麻雀の繰り返しゲームの性質に着目して一ゲームをそのまま考慮する代わりに一局単位を最適化する方法を提案する.
そのため最終順位を考慮したコンピュータ麻雀プレイヤの構築法について述べる.

Atari と異なり大量に取得可能な牌譜中に現れた点数状況から最終順位を予測するモデルの学習を行う.
モンテカルロ法のシミュレーションでの報酬を予測モデルの結果を用いることで最終順位に基づく手をプログラムは選択する.
この手法は実験により局収支を最大化するプレイヤと比較して最終順位を考慮したプレイヤの性能が高いことを示した.

次に上記の麻雀プレイヤを利用し強化学習を用いた麻雀プレイヤを構築する方法について述べる.
初めに手牌から和了点数を予測するモデルを生成した牌譜から学習する.
このモデルの結果と期待最終順位を用いて効率的な和了を行う手をプログラムは選択する.
得られたプログラムは高い点数を和了する技術を獲得したものの, 自己対戦の結果は元のプログラムに勝ち越すことはできなかった.