不完全情報ゲームにおける AI に関する調査

九州大学システム情報科学府 伊東研究室

松田 真治

1 はじめに

対戦ゲームは完全情報ゲームと不完全情報ゲームの大きく 2 つに分けることができる．完全情報ゲームとは囲碁や将棋のようにすべての意思決定点において，これまでにとられた行動や実現した状態に関する情報がすべて与えられているようなゲームのことであり，不完全情報ゲームとは麻雀やポーカーのように行動や状態などの情報が一部隠されているゲームのことである．多くの完全情報ゲームにおいて，強化学習と探索を組み合わせた手法を用いることで, AI は人類を超える好成績を残している．一方で不完全情報ゲームにおいて, AI が明確に人類を凌駕しているとは言い難いのが現状である．近年,Facebook 社開発のポーカーAI や Microsoft 社開発の麻雀 AI が人類相手にも好成績を残すなど，不完全情報ゲームにおけるゲーム AI の開発には注目が集まっている．

2 関連研究

文献[1]では完全情報ゲームにおいて広く使用されている従来の強化学習＋探索という手法が不完全情報ゲームと相性が悪いことを示し，不完全情報ゲームにも対応できるように改良された強化学習＋探索のフレームワークである ReBeL を紹介している。ReBeL の大きな特徴は，探索において通常の States ではなく独自の Public Belief States という概念を利用している点である．ReBeL は不完全情報ゲームであるテキサスホールデムポーカーにおいて超人的な結果を残している．

文献[2]では麻雀 AI における新たなデータ表現方法や学習アルゴリズムを提案している．特に Data Plane Structure は畳み込みニューラルネットワークと相性の良い画期的なデータ表現方法であり，水上らの先行研究(2015)と比較してモデルの精度をおよそ 8.34%向上させることに成功した．

文献[3]では麻雀 AI における新たなデータ表現方法，学習アルゴリズム，および思考ルーチンを提案している．特にグローバルな視点で収支を予測する Global Reward Prediction という予測モデル，不完全情報ゲームに適した Oracle Guiding という効率の良い学習アルゴリズム，麻雀に適した Parametric Monte-Carlo Policy Adaptation という探索のルーチンは画期的なアイデアである．本 AI は実際に天鳳に参加しており，天鳳に参加している麻雀 AI としては初めての十段到達を達成している．

2.1 Combining Deep Reinforcement Learning and Search for Imperfect-Information Games

この文献[1]では，不完全情報ゲーム向けにチューニングされた強化学習と探索のフレームワークである, ReBeL の手法の紹介と ReBeL を用いたポーカーAI の紹介をしている．

完全情報ゲームにおいて目覚ましい成績を残してきた強化学習と探索を組み合わせる手法は，不完全情報ゲームにおいてはうまく行かないことが多い．なぜなら，不完全情報ゲームにおいて，自分の最適な戦略は相手の戦略によって変わり，相手の戦略を把握しなければ適切に探索を行うことが出来ないためである．そこで本文献で紹介する ReBeL フレームワークでは，探索において通常の states ではなく，独自の public belief states という概念を導入し，不完全情報ゲームにおいても適切な探索ができるようにしている．

public belief states では通常の states に加えて推定した相手の戦略も組み込まれる．この相手の戦略を belief と呼び，相手が行動を起こすたびに，ベイズ推定を用いて belief を更新していく．相手の belief を一つに決め打ちすれば，完全情報ゲームと同じ方法で探索を行うことができるようになる．

ReBeL では強化学習のアルゴリズムとして Counterfactual Regret Minimization(反事実的後悔最小化法)が採用されている．以下に ReBeL の実戦での成績を示す．

表 1-1 ReBeL の対戦結果

対戦相手	Slum Bot	Baby Tartanian8	Top Human
成績	45±5	9±4	165±69

上の表が示すとおり ReBeL は過去のコンテストで優勝した AI やポーカーのトッププロに対しても勝ち越している．

2.2 Computer Mahjong Player via Deep Convolutional Neural Networks

この文献[2]ではオンライン麻雀対戦サービスである天鳳の牌譜を教師とした，CNN を用いた教師あり学習手法の提案をしている．特に先行研究のデータ構造を改善した Data Plane Structure と呼ばれるデータ構造は画期的である．

先行研究で入力として用いられるデータ構造は, One-hot Structure と呼ばれるものであった. One-hot Structure では各牌について 5 列存在するテーブルを用意し，各牌の枚数について，牌が 0 枚あるなら 1 列目に 1 を, 4 枚あるなら 5 列目に 1 を書き込む形式であった．本文献で紹介する Data Plane Structure では各牌について 4 列存在するテーブルを用意し，各牌の枚数について，牌が 0 枚あるなら何も書き込まず, 2 枚あるなら 1 列目と 2 列目に書き込み，4 枚あるなら 1 列目～4 列目に書き込む方式のデータ構造である．

表 2-1. One-hot Structure

	0 枚	1 枚	2 枚	3 枚	4 枚
1m	1	0	0	0	0
2m	0	0	0	0	1
・・・					
西	0	0	1	0	0
北	0	1	0	0	0

表 2-2. Data Plane Structure

	1 枚	2 枚	3 枚	4 枚
1m	0	0	0	0
2m	1	1	1	1
・・・				
西	1	1	0	0
北	1	0	0	0

Data Plane Structure の利点について，文献は様々な情報と比較しやすい，コードの拡張性が高い，データ空間を 20%削減可能などが挙げている．

次にモデルを以下の表に示すような形で構成した．

表 2-3. 特徴量

特徴	# of planes
自分の手牌	1
赤 5	1
捨て牌	4
副露牌	4
ドラ宣言牌	1
リーチ者	3
順位	4
場風	1
自風	1
過去 1 巡	13
過去 2~6 順	各 9

表 2- 3 の特徴量を用いて，捨て牌選択，ポン選択，チー選択，カン選択，リーチ選択モデルを構成し CNN を用いて教師あり学習を行った．その結果を以下に示す．

表 2-4. 実験結果

モデル	牌譜との一致率
捨て牌選択	70.44%
ポン選択	88.32%
チー選択	90.62%
リーチ選択	75.85%

水上らの先行研究と比較して，大きく変更した点は先述のデータ構造のみにも関わらず，全体として牌譜との一致率を 8.34%向上させることに成功した．この点からも Data Plane Structure が麻雀に適したデータ構造であると考えられる．

2.3 Suphx: Mastering Mahjong with Deep Reinforcement Learning

本文献[3]ではオンライン麻雀対戦サービスの天鳳に実際に参戦している麻雀 AI である Suphx で使われている技術について紹介されている．特に Global Reward Prediction, Oracle Guiding, Parametric Monte-Carlo Policy Adaptation は画期的で優れた技術である．

2.3.1 データ構造

文献[2]で紹介した Data Plane Structure を採用している．

2.3.2 Global Reward Prediction

麻雀，特に天鳳の成績において最も重要なのは最終的な着順である．そのため天鳳をプレイするときは 1 局の収支をプラスにすることだけでなく，最終的な着順を上げることを考えながらプレイを行わなければならない．つまりその局で何点の和了りをするかによっ

て最終的な順位にどのような影響があるかを考えることが重要である.そこで Suphx ではその局に何点あがるかを入力として，最終的な着順を予想する Global Reward Prediction と呼ばれる仕組みを考案している. Global Reward Prediction は GRU を組み合わせたものであり，実際の最終着順との最小二乗誤差を小さくするように学習される.

2.3.3 Oracle Guiding

不完全情報ゲームと強化学習の相性の悪さとして，見えない情報量が多くて，学習の進みが遅いというものがある．そこで Suphx では効率的な強化学習の手法である Oracle Guiding という手法を考案している. Oracle Guiding では，はじめに本来なら見ることのできない情報を全て見えるようにした状態で強化学習を行う．その後，学習を進めるにつれて徐々に見える情報を本来見えるものだけにもどしていく．このような手法を用いることで効率的に強化学習を行うことができる.

2.3.4 Parametric Monte-Carlo Policy Adaptation

麻雀では鳴きなどによってプレイの順序が変わるため探索を行うことが難しい．そこで Suphx では探索することを諦め, Parametric Monte-Carlo Policy Adaptation という手法を代替として考案した. Parametric Monte-Carlo Policy Adaptation では配牌が配られた時点で相手の配牌や山をランダムに決定し終局までシミュレーションすることを 100,000 回繰り返す．その結果を用いてその局に消極的に参加するのか，積極的に参加するのかを決定する.

2.3.5 結果

強化学習を行った結果を以下に示す．なお SL は単純な教師あり学習, RL-basic は SL + 強化学習, RL-1 は RL-basic + Global Reward Prediction, RL-2 は RL-1 + Oracle Guide を用いたモデルである.

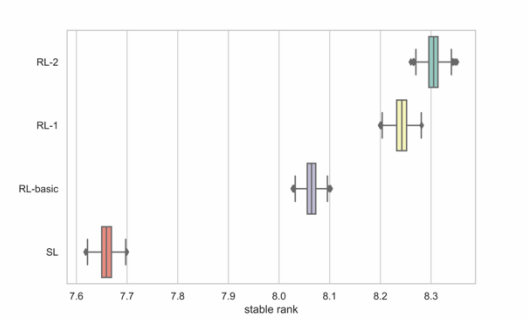

図 3.1 強化学習結果

また先述の RL-2 に Parametric Monte-Carlo Policy Adaptation を適用した場合と適用しなかったとで対戦させた場合，適用させたものの勝率が 66%となった.

次に実際の天鳳での各 AI の成績を以下に示す.

表 3-1. 天鳳での成績

	試合数	最高段位	安定段位
爆打	30,516	9	6.59
NAGA	9,649	8	6.64
人間	8,031	10	7.46
Suphx	5,760	10	8.74

表 3-1 の示すとおり, Suphx は先行の AI である爆打や NAGA を最高段位，安定段位ともに少ない試合回数にも関わらず上回っている. また最高段位はトップ層の人間と同じ 10 段であるが，安定段位はトップ層の人間をも上回る目覚ましい成績を残している.

終わりに

Facebook 社のポーカーAI や Microsoft 社の麻雀 AI である Suphx など，近年では不完全情報ゲームにおいても人間と同等，もしくは人間以上の成績を残すような AI が出てきており，この分野の研究も盛んである.

今後はこの演習で学んだ知識を生かして，麻雀 AI に関する研究を進めたいと考えている.

参考文献

[1] Noam Brown, Anton Bakhtin, Adam Lerer, Qucheng Gong. "Combining Deep Reinforcement Learning and Search for Imperfect-Information Games", https://arxiv.org/abs/2007.13544, 27 Jul 2020.

[2] Shiqi Gao, Fuminori Okuya, Yoshihiro Kawahara, Yoshimasa Tsuruoka. "Building a Computer Mahjong Player via Deep Convolutional Neural Networks", https://arxiv.org/abs/1906.02146, 5 Jun 2019.
[3] Junjie Li, Sotetsu Koyamada, Qiwei Ye, Guoqing Liu, Chao Wang, Ruihan Yang, Li Zhao, Tao Qin, Tie-Yan Liu, Hsiao-Wuen Hon."Suphx: Mastering Mahjong with Deep Reinforcement Learning", https://arxiv.org/abs/2003.13590, 30 Mar 2020.