jh230044

深層強化学習を用いた麻雀 AI に関する研究

鶴岡慶雅 （東京大学）

概要

確率的な結果が出るゲームでは、限られたデータからエージェントのパフォーマンスを評価することは困難である。モンテカルロサンプリングの結果は、分散が大きいため信頼性のある指標とはならない。特に麻雀のような不完全情報ゲームは、巨大な状態空間を持つため、エージェントの評価が難しい。例えば、Suphx は、オンライン麻雀で人間と 5,760 試合対戦し、そのパフォーマンスを評価するのに 4 ヶ月の時間を要した。本研究では、麻雀プレイヤーの評価方法として、平均順位の不偏推定値であり、かつ分散の小さい推定値を求める方法を提案する。提案手法は麻雀の巨大なゲーム木と盤面情報を扱うために、3 つの技術を導入する。ゲームのサブゲームへの分割、ツモ牌、配牌、およびドラ表示牌による分散への対処、ニューラルネットワークの導入である。オンライン麻雀の牌譜を使用してデータセットを作成し、ニューラルネットワークベースの価値関数を訓練した。オンライン麻雀の牌譜において提案手法を評価し、推定順位が平均順位の不偏推定量であり、推定順位の分散が平均順位の 45.5% 減少していることを確認した。その結果、プレイヤーの能力を正確に評価するために必要なゲーム数が 45.5% 削減される。

1 共同研究に関する情報

1.1 共同研究を実施した拠点名

- 東京大学 情報基盤センター
- mdx

1.2 課題分野

- データ科学・データ利活用課題分野

1.3 共同研究分野 (HPCI 資源利用課題のみ)

- 超大規模データ処理系応用分野

1.4 参加研究者の役割分担

- 鶴岡慶雅 (東京大学): 研究代表者、総括
- 松井勇佑 (東京大学): 共同研究者、研究補助
- 大神卓也 (東京大学): 麻雀プレイヤーの実力推定に関する研究、開発
- 天野克敏 (東京大学): 麻雀プレイヤーの実力推定に関する研究、開発

2 研究の目的と意義

本課題では、麻雀プレイヤーの実力をより少ない試合数で評価するための手法について研究を行った。当初の提案では、麻雀 AI の活用範囲を広げることを目的とし、既存の麻雀 AI にはない工夫を行うことをテーマとして掲げていた。その一環として、プレイヤーの実力を正確に推定するための方法を考察した。

麻雀 AI の導入目的の一つは、プレイヤーの実力を正確に推定することである。これを実現するために、強力な麻雀 AI とプレイヤーの選択の類似度を測る手法が考案されてきた。しかし、AI とプレイスタイルが異なるプレイヤー

に対しては、正確な評価が難しいという課題が存在する。

麻雀プレイヤーの公平かつ効率的な実力の評価に関する研究は、これまで十分に行われてこなかった。麻雀における実力評価のコストを低減することは、麻雀の楽しみ方に大きな影響を与えると考えられる。そのため、本研究では、麻雀プレイヤーの公平かつ効率的な実力評価に集中して取り組んだ。

不完全情報ゲームにおいては、確率論的な状態遷移が存在するため、毎回の試合結果にばらつきがある。このため、プレイヤーの実力を評価するには膨大な試合数が必要となる。例えば、麻雀では 1,000 試合（およそ 500 時間）の試合結果だけでは十分にプレイヤーの実力を評価することが難しいと言われており、人間のプレイヤーにとって大きな負担となっている。

プレイヤーの実力評価に必要な試合数の削減に関する研究は主にポーカーで行われてきた。ポーカーでは試合数の削減が可能であることが示されているが、ポーカーよりも規模が大きいゲームにおいても同様の手法が適用可能かどうかは未知数である。そこで、本研究では麻雀においてプレイヤーの評価に必要な試合数を削減する手法を検討した。麻雀はポーカーよりも不完全情報の大きさやゲーム木の長さが大きいため、研究対象として適切であると考えられる。麻雀においても試合数の削減が可能であることを示すことで、他の大規模な不完全情報ゲームへの適用可能性を示すことができる。

さらに、麻雀は日本において人気のあるゲームの一つである。例えば、日本における麻雀のプロリーグである M リーグは数百万人の観客を集める人気を誇っている。麻雀におけるプレイヤーの実力評価が効率化されれば、麻雀の楽しみ方の多様化を促進することができる。

3　当拠点公募型研究としての意義

麻雀 AI は学際的な研究対象であり、複数の専門分野にわたる研究室のメンバーが取り組んでいる。麻雀 AI の研究には、ゲーム理論だけでなく、強化学習や自然言語処理などの知見も必要とされる。このため、麻雀 AI の研究は多様な分野にまたがり、共同研究として実施することが求められる。

さらに、麻雀 AI の研究には膨大な計算資源が必要である。麻雀は盤面の複雑さとゲームの長さが特徴であり、これらを処理するためには大規模なニューラルネットワークを使用する必要がある。また、同時に処理する盤面の数が多いため、使用する GPU のメモリが大量に必要となる。

また、麻雀プレイヤーの実力を分析するために、インターネット麻雀「天鳳」[*1]の数年分のデータを使用した。これらのデータは非常に大きく、処理には大量のストレージが必要である。

4　前年度までに得られた研究成果の概要

今回が新規の課題であったため、省略。

5　今年度の研究成果の詳細

5.1　概要

今回行った研究の概要を図 1 に示す。プレイヤーの牌譜をもとに、各試合について運（luck）を計算する。実際の順位から運を差し引くことによって、推定順位を求める。推定順位は以下の 2 つの性質を満たす。

1. 実際の順位の不偏推定量である。

[*1] https://tenhou.net/

2. 実際の順位よりも分散が小さい。

実際の順位の不偏推定量であるため、プレイヤーの実力を偏りなく評価できる。さらに、分散が小さいため、同じ幅の信頼区間を求めるために必要な試合数を削減できる。

ポーカーにおける実力推定の先行研究を再現し、その手法を麻雀に適用するための工夫を行い、MJ-DLVAT と名付ける。具体的には、以下の 3 つの工夫を行った。

1. 深層ニューラルネットワークの導入。
2. サブゲームへの分割。
3. 複数の運要素に関する分散削減。

これにより、ゲームのサイズが大きい麻雀でも実力推定が可能になった。

Fig. 2 は提案手法の概略を示している。麻雀の試合を局と呼ばれるサブゲームに分割し、それぞれのサブゲームについてプレイヤーの実力を推定する。各サブゲームの結果を合算することで、試合全体におけるプレイヤーの実力を推定する。

また、運の要素としてツモ、裏ドラ、配牌の 3 種類を考慮する。それぞれの要素は異なる性質を持つため、異なるアプローチで分散を削減する。

5.2 サブゲームへの分割

提案手法では、ゲーム内のすべてのランダムな状態遷移について総和をとる必要がある。しかし、ランダムな状態遷移は 1 試合ごとにおよそ 600 ほどあり、計算量が膨大となるため、実行が困難であった。そこで、局と呼ばれるサブゲームに着目し、局終了時点での利得を推定するモデルを用いることで問題を分割した。この分割により、不偏推定の性質が保たれることが示された。

5.3 複数の運要素に関する分散削減

麻雀はルールが複雑で、ゲーム内で起こる確率的な状態遷移にはいくつかの種類がある。本研究では、その中でも特に重要な以下の 3 つの要素に注目した。

1. 裏ドラ
2. ツモ
3. 配牌

これらの要素はそれぞれ性質が異なるため、それぞれに対して異なるアプローチを用いて分散の削減を行った。

5.4 深層ニューラルネットワークの導入

先行研究では、分散の削減に用いる価値関数を線形関数で近似していた。一方で、Suphx などの麻雀 AI は、複雑な盤面情報を処理するために畳み込みニューラルネットワークを用いている。本研究でも、3 に示すような畳み込みニューラルネットワークを用いることで、価値関数を近似した。具体的なアルゴリズムは Algorithm 1 に示している。

5.5 データセットの作成

データセットとして、オンライン麻雀サイト「天鳳」の最上位のフィールドである鳳凰卓の試合記録を使用した。訓練データおよび検証データとしては、2017 年、2018 年、および 2020 年の記録から約 56 万試合のデータを使用した。また、モデルの評価のために、2019 年と 2021 年の記録から、少なくとも 500 試合をプレイした 102 人のプレイヤーのデータを用いてプレイヤーごとのデータセットも作成した。このデータセットはプレイヤーごとにグループ化され、各プレイヤーのサブゲームごとの利得の変動がラベルとして付けられている。

5.6 実験結果

検証用データにおける実際の順位の分散は 0.119 であった。訓練後のモデルでは、推定順

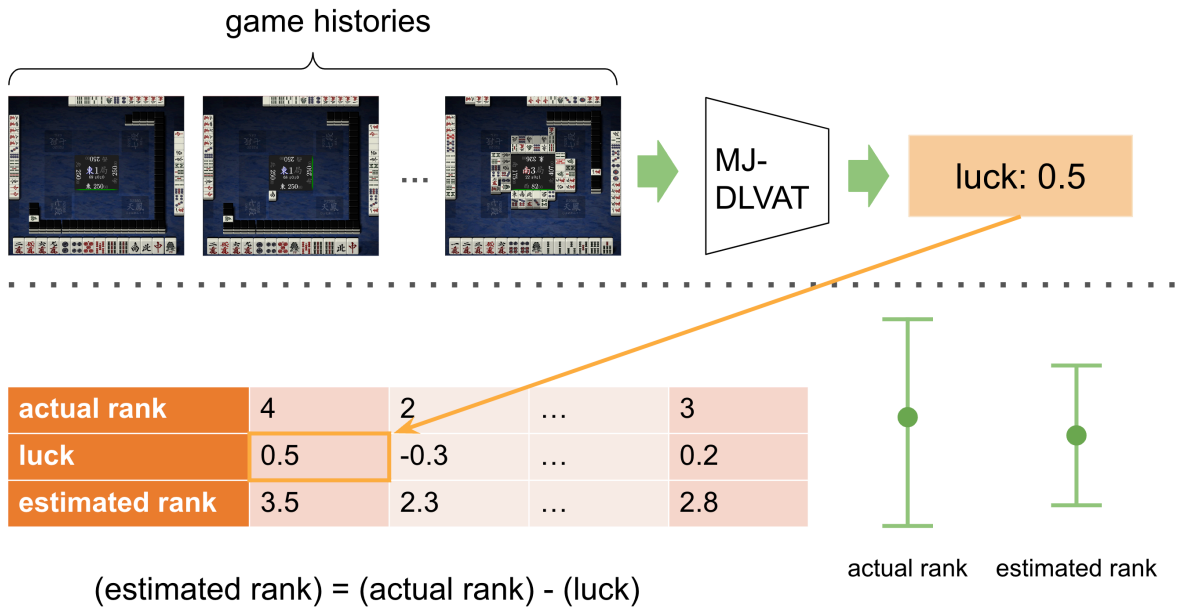


図 1: 提案するプレイヤー評価の枠組み。MJ-DLVAT はゲーム履歴を分析し、各ゲームにおける運の値を算出する。表には、実際の順位（actual rank）、運（luck）、および運を取り除いた後の推定順位（estimated rank）が示されている。推定順位は実際の順位に比べて幅の狭い信頼区間を持ち、少ない試合数での実力評価が可能となる。

位の分散は 0.072 となり、実際の順位の分散よりも小さくなった。

表 1: 提案手法の分散減少率

	Reduction rate of variance
Monte Carlo	—
MJ-DLVAT (all)	45.5%
MJ-DLVAT (drawn tiles)	40.5%
MJ-DLVAT (hidden-dora)	2.5%
MJ-DLVAT (dealt tiles)	2.0%

5.6.1 分散削減の評価

訓練されたモデルを用いて、ツモ、配牌、そして裏ドラについて、提案された方法で推定順位を個別に計算した。表 Table 1 は、テストデータにおける各プレイヤーの対戦数で重み付けされた平均分散削減率を示している。提案手法で順位を推定することにより、モンテカルロ法と比較して分散を合計で 45.5% 削減した。内訳を見ると、引いた牌による分散の削減が 40.5%、配牌によるものが 2.5%、裏ドラによるものが 2.0% である。これにより、すべての種類の分散に対処することで高い性能を実現している。

ツモと配牌については、各ラウンドでの各プレイヤーによる取得牌の数は、配牌が 13 枚、ツモが 10〜20 枚であり、大きな違いはない。しかし、削減された分散の割合は大きく異なる。裏ドラについては、あるラウンドでのリーチの確率が約 20%、そのラウンドでの勝利確率が約 50% であるため、裏ドラが公開される状況は多くない。そのため、分散削減率が小さい結果となったのは納得できる。

5.6.2 エラーレートとの比較

本研究とは異なる実力推定のアプローチとしてエラーレートが挙げられる。エラーレー

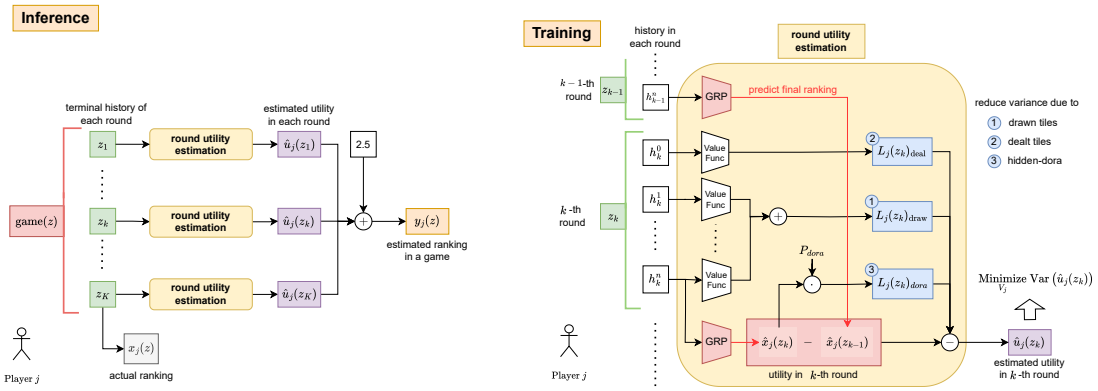


図 2: 提案手法の概略図。（左）最終順位の推定値を推測する方法。ゲームの履歴をサブゲームに分割し、各サブゲームで推定された利得を合計する。（右）それぞれの価値関数を訓練する方法。ツモ、配牌、裏ドラによって引き起こされる分散を減らすために、履歴から第 k サブゲームの効用を推定し、推定値の標本分散を最小化する。

表 2: 天鳳の卓ごとのエラーレート

卓名	エラーレート
一般卓	0.896
上級卓	0.706
特上卓	0.522
鳳凰卓	0.410

トは、AI エージェントの状態-行動価値関数 $Q(s,a)$ をプレイヤーの状態-行動ペア (s,a) の集合を用いて計算される。状態 s でプレイヤーが選択した行動 a に対する $Q(s,a)$ が小さいほど、エラーレートは大きくなる。先行研究によれば、エラーレートを用いると少ない試合数で初心者、中級者、上級者を区別できることが確認されている。(Table 2)

MJ-DLVAT とエラーレートの手法を比較するために、散布図を用いる。Fig. 4a は、実際に得られた平均順位と MJ-DLVAT（全体）によって推定された平均順位との関係を示している。Fig. 4b は、実際に得られた平均順位と平均エラーレートとの関係を示している。これら二つの図を比較すると、MJ-DLVAT による推定順位は実際に得られた順位と高い相関があるのに対し、誤差率は得られた順位との相関が低い。

Table 2 に示されるようにエラーレートは異なる卓における選手の能力をうまく区別した。しかし、鳳凰卓内での選手の能力差を測定することはできなかった。考えられる原因として、二つの仮説がある。選手の能力差がわずかである場合には機能しない、または AI エージェントの能力よりも強い選手の能力差を区別できない。手持ちのデータと AI エージェントからは、これらの仮説に関連するさらなるテストを行うことは不可能である。

5.6.3 結果詳細

また、MJ-DLVAT の結果をそれぞれの運の要素ごとに可視化した。Fig. 5 には、MJ-DLVAT を用いて運の要素ごとにに推定された平均順位を示している。この 3 つ全てで、推定された効用と実際に得られたランキングとの間の相関は大きく、特に配られた牌とドラ表示牌において顕著である。

Algorithm 1 Learning Value Function in MJ-DLVAT

Require: subject to the following conditions

- $V_{\boldsymbol{\theta}}$: value function $\mathbb{R}^{788\times34\times1} \to \mathbb{R}^4$, satisfying the constraint that the sum of its 4-dimensional output is zero.
- N: number of iterations for each epoch.
- B: batch size.
- z: terminal history batch.
- $\boldsymbol{u} = (u_1, u_2, u_3, u_4)$: utility for each player in the batch.

for $t = 1, \cdots, N$ **do**
  sample a batch of B tuples $(z, \boldsymbol{u})$
  initialize a $\boldsymbol{Luck} = \mathbf{0}_{B\times4}$
  for all history_before_chance_player_act $h \sqsubseteq z$ in the batch **do**
    $\boldsymbol{Luck} += V_{\boldsymbol{\theta}}(h \cdot a) - \sum_{a' \in \mathcal{A}(h)} \pi^c(a'|h)\, V_{\boldsymbol{\theta}}(h \cdot a')$, vectorized over the batch
  end for
  $\hat{\boldsymbol{u}} = \boldsymbol{u}$ - $\boldsymbol{Luck}$
  $loss$ = sample_variance($\hat{\boldsymbol{u}}$), calculated over the batch
  update $\boldsymbol{\theta}$ using Adam
end for

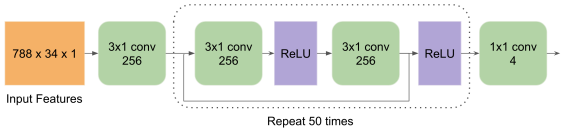
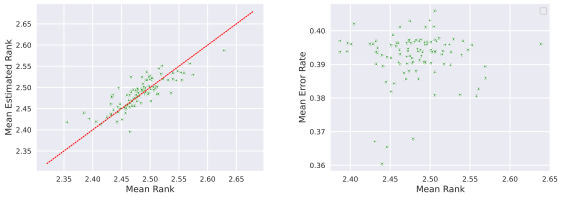


図 3: 価値関数を近似するために残差接続を持つ CNN（畳み込みニューラルネットワーク）を導入する。この構造は Suphx で提案されたアーキテクチャに類似している。

Fig. 6 には、テストデータセットで最も多くのゲーム数を持つ 6 人のプレイヤーに対して計算された 95% 信頼区間の結果が示されている。推定された順位の平均値と実際の順位の平均値が近いことが見て取れる。また、推定された順位を用いて計算された信頼区間が、実際の順位



(a) MJ-DLVAT (all)　(b) Error rate

図 4: 推定平均順位と実際の平均順位に関する散布図。各点がプレイヤーを表している。

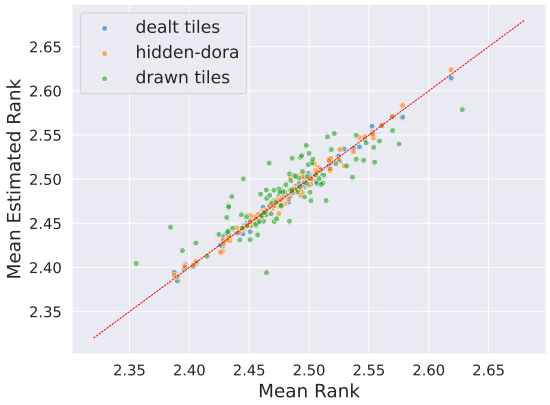


図 5: MJ-DLVAT を用いて運の要素ごとに推定した着順と実際に得られた平均順位。各点がプレイヤーを表している。

のそれよりもかなり小さいことも見て取れる。

5.7 計算機利用上の工夫

麻雀の盤面を表すデータは比較的大きい。盤面の総数も大きく、データセット全体は CPU メモリに乗り切らない。このようなデータを扱うために、IterableDataset[*2]を用いた。さらに、複数の CPU で並列にデータを読み込むことによって、訓練データの読み込みを効率化した。結果として、訓練に使用する場面数を既存の麻雀 AI に関する研究に比べて 10 倍以上増加させることができ、性能の向上に繋がった。

[*2] https://pytorch.org/docs/stable/data.html#torch.utils.data.IterableDataset

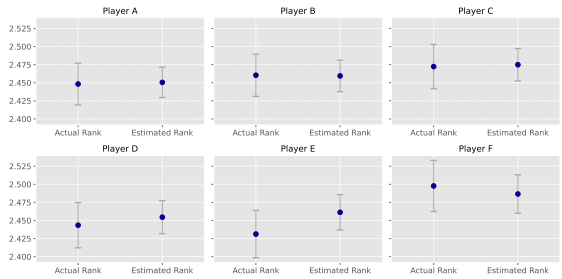


図 6: テストデータにおける 6 人のプレイヤーに関する、推定順位と実際の順位の 95% 信頼区間。

6　進捗状況の自己評価と今後の展望

前述の通り、本課題では申請時と異なるテーマに関して研究を行った。これまでの研究では、インターネット麻雀「天鳳」のデータについて、麻雀プレイヤーの実力評価に必要な試合数を 45.5% 削減することができるという結果を示すことができた。この結果をもとに、論文 [1] を投稿した。

今後の展望としては、アプリケーションの作成が挙げられる。

6.1　性能向上の可能性

提案手法では、ニューラルネットワークによって近似した価値関数を用いている。この価値関数の近似精度をさらに高めるために、使用するデータを増やすことや、モデルの改良などの工夫が必要である。これにより、試合数の削減や結果の精度向上が期待される。

6.2　実データへの適用

今回提案した手法を用いて、インターネット麻雀「天鳳」においてプレイヤーの実力の推定値を閲覧できるサービス (図 Fig. 7) を作成する。推定順位を公開することにより、ユーザーが直近の試合結果から自身の実力を推定できるサービスを提供することが可能となる。これにより、ユーザーは自身の成長や課題をより正確に把握できるようになる。

図 7: アプリケーションの参考画面

7　研究業績一覧（発表予定も含む）

学術論文　（査読あり）

なし

国際会議プロシーディングス　（査読あり）

[1] T. Ogami, K. Amano, Y. Tsuruoka, ‘MJ-DLVAT: A Deep Learning Value Assessment Technique for Mahjong’, IEEE Conference on Games, 2024

国際会議発表　（査読なし）

なし

国内会議発表　（査読なし）

[2] 大神卓也, 天野克敏, 奈良亮耶, 鶴岡慶雅 ‘分散減少法を用いた麻雀における実力推定’, ゲームプログラミングワークショップ, 2024

公開したライブラリ等

なし

その他（特許，プレス発表，著書等）

なし