反実仮想報酬を用いた強化学習による和了判断の学習

飯塚 詞也[1,a)]　森山 甲一[1,b)]　島 孔介[1]　松井 藤五郎[2]　武藤 敦子[1]　犬塚 信博[1]

概要：麻雀 AI における深層強化学習の研究は広く行われており，プロレベルの成績を達成したシステムも登場している．これらの研究は主に打牌や副露を対象としており，和了可能な局面で即座に和了するか見送るかを選択する和了判断の研究は著者の知る限り存在しない．和了判断における見送りは，手役や点数状況によっては和了を上回る得点が期待できる重要な意思決定である．見送りが有利な局面は希少であるが，そのような局面での判断が対局の勝敗を左右することもある．しかし，希少であるがゆえに見送りの報酬信号が得にくく，また和了を選択すると確定の正の報酬が得られるのに対し，見送りの報酬は対局の終局まで不確定であるため，通常の強化学習では見送り判断を適切に学習することが困難である．本研究では，選択しなかった行動の結果を反実仮想的にシミュレートし，両行動の得点差を報酬として用いる学習手法を提案する．通常報酬の強化学習を比較手法とした実験より，通常報酬の強化学習では見送りが有利な局面をほとんど学習できないのに対し，反実仮想報酬を用いることで見送り判断の学習が促進され，より高い得点を達成することを確認した．

Learning Winning Hand Decisions in Mahjong via Reinforcement Learning with Counterfactual Rewards

Fumiya Iizuka[1,a)]　Koichi Moriyama[1,b)]　Kosuke Shima[1]　Tohgoroh Matsui[2]　Atsuko Mutoh[1]
Nobuhiro Inuzuka[1]

Abstract: Research on deep reinforcement learning for mahjong AI has been conducted extensively, and systems achieving professional-level performance have emerged. These studies have primarily targeted tile discarding and calling decisions, and to the best of our knowledge, no study has addressed the winning hand decision — choosing whether to win immediately or skip when a winning hand is available. Skipping in the winning hand decision is an important decision, as it can lead to a higher expected score than winning immediately, depending on the yaku and the score situation. Although situations where skipping is advantageous are rare, decisions in such situations can sometimes determine the outcome of a game. However, precisely because such situations are rare, reward signals for skipping are difficult to obtain; furthermore, while winning yields a certain positive reward, the reward for skipping remains uncertain until the end of the game. As a result, it is difficult for standard reinforcement learning to learn appropriate skip decisions. In this study, we propose a learning method that counterfactually simulates the outcome of the unchosen action and uses the score difference between the two actions as a reward. Experiments using reinforcement learning with standard rewards as a baseline confirm that standard reinforcement learning can barely learn to skip in situations where skipping is advantageous, whereas using counterfactual rewards promotes the learning of skip decisions and leads to higher scores.

1. はじめに

麻雀 AI における深層強化学習の研究は広く行われており，打牌や副露など様々な意思決定を対象とした手法が提案されている．中でも，Suphx[1] はオンライン麻雀においてプロレベルの成績を達成しており，深層強化学習が麻雀

[1] 名古屋工業大学
Nagoya Institute of Technology
[2] 中部大学
Chubu University
[a)] cmt14003@ict.nitech.ac.jp
[b)] moriyama.koichi@nitech.ac.jp

© 2026 Information Processing Society of Japan

AI に有効であることが示されている．

麻雀の意思決定の一つとして，和了可能な局面で即座に和了するか，より高い得点を狙って見送るかを選択する和了判断がある．見送りが有利な局面は希少であるが，そのような局面では和了を見送ることで高い得点が期待でき，1 回の判断が対局の得点に大きく影響するため重要な意思決定である．しかし，従来の麻雀 AI の研究において，和了判断を専門に扱った研究は著者の知る限り存在しない．

和了判断の学習には複数の問題が存在する．第一に，見送りが有利な局面が希少なため，学習サンプルに強い偏りが生じる．このため見送り行動が十分に学習されにくい．第二に，和了を選択した場合は確定の正の報酬が即時に得られるのに対し，見送りの報酬はその後の結果に依存するため不確定である．これらの要因により，通常の強化学習では和了方向に強くバイアスがかかる．さらに，見送りを選択する機会自体が減少するため，見送り行動に対する報酬信号がますます得られにくくなるという悪循環が生じる．

そこで本研究では，選択しなかった行動の結果を反実仮想的にシミュレートし，両行動の得点差を報酬として用いる学習手法を提案する．この手法では，実際に選択した行動が和了・見送りのいずれであっても，両行動の結果を比較した報酬が得られる．そのため，学習サンプルの偏りにかかわらず，見送りが選択されない局面においても見送り結果を踏まえた学習が可能となり，希少な見送り局面に対しても明確な報酬信号を与えることができる．

本研究では，簡易麻雀環境において提案手法の有効性を検証するため，通常報酬の強化学習および教師あり学習との比較実験を行う．本稿の構成は以下の通りである．2 章で準備として麻雀と強化学習の基礎を述べ，3 章で提案手法を説明する．4 章で実験設定と結果を示し，5 章でまとめと今後の課題を述べる．

2. 準備

2.1 麻雀

麻雀は，牌を用いて通常 4 人で行うゲームである．牌には数字牌（1〜9 の数字が書かれた牌が 3 種類）や字牌などがあり，合計 136 枚から構成される．各プレイヤーは手牌 13 枚を持ち，自分の番にツモ（牌山から 1 枚引く）を行って 14 枚にした後，打牌（不要な牌を 1 枚捨てる）を行う手順を繰り返す．

手牌が 4 組の面子（3 枚で構成される組み合わせ）と 1 組の雀頭（同じ牌 2 枚の組み合わせ）などの和了形を満たし，役（和了時に必要となる特定の手牌構成や状況）が成立すると和了できる．自分のツモで和了形が完成した場合はツモ和了，相手の捨て牌で完成した場合はロン和了となり，成立した役の種類・数と手牌の構成に応じた点数を獲得して局が終了する．誰も和了せずに牌山が尽きた場合は流局となる．

本研究が対象とする和了判断とは，聴牌（あと 1 枚で和了できる状態）で上がり牌を得た際に，即座に和了するか，より高い得点を狙って見送るかを選択する意思決定である．見送りを選択した場合，手牌を進めてより高い役や点数を狙える可能性がある一方，相手に先に和了されたり流局したりするリスクも生じる．また，聴牌中に自分の和了牌が自分の捨て牌に含まれている状態を振聴（フリテン）といい，この状態ではロン和了ができない．手牌が進んで和了牌が変わり，捨て牌に含まれなくなると振聴は解消される．したがって和了判断は，確定した点数を直ちに得る行動と，不確定なより高い得点を期待して待つ行動を天秤にかける意思決定である．

2.2 強化学習

強化学習は，エージェントが環境との相互作用を通じて累積報酬を最大化する方策を獲得する枠組みである．深層強化学習は深層学習と強化学習を組み合わせた手法であり，ニューラルネットワークによって高次元の状態から直接方策や価値関数を近似することで，複雑な問題にも適用できる．

本研究では Proximal Policy Optimization（PPO）[2] を用いる．PPO は方策ベースの強化学習手法の一つであり，安定した方策更新を目的として提案された手法である．更新前後の方策の確率比 $\rho_t(\theta) = \pi_\theta(a_t \mid s_t)/\pi_{\theta_{\text{old}}}(a_t \mid s_t)$ をクリッピングにより制御することで，方策の更新が急激にならないように抑制する．

アドバンテージの推定には Generalized Advantage Estimation（GAE）[3] を用いる．TD 誤差

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \tag{1}$$

を基に，アドバンテージ $\hat{A}_t$ を

$$\hat{A}_t = \sum_{l=0}^{T-t} (\gamma\lambda)^l \delta_{t+l} \tag{2}$$

と求める．$\hat{A}_t$ は状態 s_t における行動 a_t が平均的な行動と比べてどれだけ優れていたかを表す指標であり，λ は長期報酬の考慮度合いを調整するハイパーパラメータである．

PPO の目的関数は，クリッピング付き方策利得 L^{CLIP}，価値関数損失 L^{VF}，エントロピー正則化項 S の 3 項で構成される．

$$L(\theta) = L^{\text{CLIP}}(\theta) - c_1 L^{\text{VF}}(\theta) + c_2 S(\theta) \tag{3}$$

クリッピング付き方策利得は

$$L^{\text{CLIP}}(\theta) = \hat{E}_t\Big[\min\Big(\rho_t(\theta)\hat{A}_t,\ \text{clip}(\rho_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\Big)\Big] \tag{4}$$

であり，方策比が $1 \pm \epsilon$ の範囲を超えた場合に更新量を抑制する．c_1, c_2 はそれぞれの重みであり，エントロピー正

© 2026 Information Processing Society of Japan

則化は探索を促進する役割を持つ．また，勾配爆発を防ぐために勾配クリッピングを用いる．

3. 反実仮想報酬法

麻雀環境において，通常の強化学習では対局終局時に得られた点数をそのまま報酬として用いる．和了判断においては，和了を選択すれば確定した正の報酬が即座に得られるのに対し，見送りを選択した場合の報酬は対局終局まで不確定である．また，和了が有利な局面は見送りが有利な局面より圧倒的に多いため，エージェントは和了を頻繁に選択し続けることで常に正の報酬を得やすい．その結果，和了への傾向がさらに強まり，見送り行動を学習する機会が極めて少なくなる．

本研究では，実際に選択した行動の点数から選択しなかった行動の点数を引いた値を報酬として用いる反実仮想報酬法を提案する．和了可能な局面 s において，和了した場合に得られる点数を $r_{\mathrm{hu}}(s)$，見送った場合に最終的に得られる点数を $r_{\mathrm{skip}}(s)$ とする．$r_{\mathrm{hu}}(s)$ は手牌・ドラなどから既に確定している．

$r_{\mathrm{skip}}(s)$ は選択した行動に応じて次のように求める．

- 和了を選択した場合：局面 s から見送りを行い，以降を現在の方策に従って対局を終局まで進めたシミュレーションによって $r_{\mathrm{skip}}(s)$ を求める．
- 見送りを選択した場合：実際の対局の最終点数がそのまま $r_{\mathrm{skip}}(s)$ となる．$r_{\mathrm{hu}}(s)$ は既に確定しているため，追加のシミュレーションは不要である．

シミュレーションが必要なのは和了を選択した場合の $r_{\mathrm{skip}}(s)$ のみである．反実仮想報酬を

$$r_{\mathrm{diff}}(s) = \begin{cases} r_{\mathrm{hu}}(s) - r_{\mathrm{skip}}(s) & \text{（和了を選択した場合）} \\ r_{\mathrm{skip}}(s) - r_{\mathrm{hu}}(s) & \text{（見送りを選択した場合）} \end{cases} \tag{5}$$

と定義する．これは実際に取った行動の点数から取らなかった行動の点数を引いた値であり，正であれば実際の選択が優れていたことを示す．和了・見送りのいずれを選択した場合でも，各和了判断に対して両行動の結果を比較した報酬を与えることができる．

4. 実験

4.1 簡易麻雀ルール

通常の麻雀は副露・立直・多様な役など考慮すべき要素が多く，打牌や副露などを含む全体の学習に多大な時間を要する．そこで本研究では，和了判断の学習に焦点を絞り効率よく実験を行うため，ルールを簡略化した 2 人対戦の簡易麻雀を実験環境として用いる．

牌は 1 種類の数牌（1〜9）のみで，各 4 枚の計 36 枚を使用する．副露（チー・ポン・カン）および立直は行わない．各対局の開始時，各プレイヤーに 7 枚の手牌が配られる．以降は，ツモ→和了判定→打牌→相手のロン判定を 2 人が交互に繰り返す．和了形は「2 組の面子＋ 1 組の雀頭」の計 8 枚から成る．誰も和了しないまま牌山が尽きた場合（最大 10 巡）は流局となり，得点は発生しない．

得点は基本点 1000 点に対し，ドラ 1 枚につき 2 倍，手牌に 1・9 を含まない場合はさらに 2 倍とし，上限を 8000 点とする．

4.2 比較手法

提案手法の有効性を検証するため，以下の 3 手法を比較する．

PPO は，対局終局時の最終得点をそのまま報酬として用いるベースライン手法である．和了判断時に方策から行動をサンプリングし，対局終局時に報酬を受け取る．

反実仮想報酬法（提案手法）は，3 章で述べた $r_{\mathrm{diff}}(s)$ を報酬として PPO で学習する．報酬の定義以外の更新手順は PPO と同一である．

教師あり学習は，自分は和了機会を常に見送り，相手は常に和了するというデータ収集方針のもとで対局を行う．収集したデータから，各和了判断における見送り後の結末（和了・失点・流局）を 3 クラス分類で，和了時の得点 $\hat{r}_{\mathrm{win}}$ と失点時の失点 $\hat{r}_{\mathrm{loss}}$ をそれぞれ回帰で予測するモデルを学習する．見送り後の期待得点を

$$\hat{r}_{\mathrm{skip}} = \hat{P}_{\mathrm{win}}\,\hat{r}_{\mathrm{win}} - \hat{P}_{\mathrm{loss}}\,\hat{r}_{\mathrm{loss}} \tag{6}$$

と定義する．ここで $\hat{P}_{\mathrm{win}}, \hat{P}_{\mathrm{loss}}$ は分類ヘッドが出力する和了・失点の予測確率である．推論時は $\hat{r}_{\mathrm{skip}}$ が即時の和了点を上回る場合に見送りを選択する．

4.3 モデル構造

全手法で共通の特徴抽出ネットワークとして，ResNet[4] を用いる．ResNet は，層が深くなるほど生じる勾配消失問題を緩和するため，各ブロックの入力をそのまま出力に加算するスキップ接続を持つ．入力は手牌・捨て牌・ドラ・残り牌山などを 64 チャネルの長さ 9 の 1 次元特徴量として表現し，最初の畳み込み層で 128 チャネルに変換した後，残差ブロックを 30 層積み重ねる．各残差ブロックは 2 層の畳み込みと，チャネルごとの重要度を自動的に学習して重み付けするチャネルアテンション機構から構成される．

PPO と提案手法のモデル構造を図 1 に，教師あり学習のモデル構造を図 2 に示す．

PPO と提案手法は同一のモデル構造を用いる．特徴抽出後の出力を畳み込み層で圧縮・平坦化し，和了・見送りの確率を出力する方策ヘッドと状態価値を推定する価値ヘッドの 2 つに分岐する Actor-Critic 構造である．

教師あり学習では，特徴抽出後の出力を圧縮・平坦化した後，3 つのヘッドに分岐する．分類ヘッドは見送り後の結末（和了・失点・流局）の確率を出力し，2 つの回帰ヘッ

© 2026 Information Processing Society of Japan

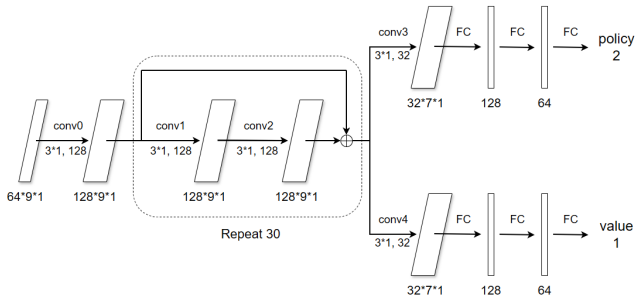


図 1 PPO・提案手法のモデル構造

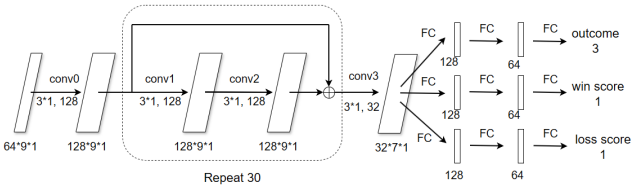


図 2 教師あり学習のモデル構造

ドはそれぞれ和了時の期待得点 $\hat{r}_{\text{win}}$ と失点時の期待失点 $\hat{r}_{\text{loss}}$ を出力する．

4.4 学習設定

学習時の対局設定として，自分と相手の両エージェントの打牌には事前学習済みの打牌モデルを共通して使用する．自分の和了判断は学習中の和了モデルで行い，相手の和了判断は見送りが有利な局面は希少であるため，常に和了するものとした．主な学習パラメータを表 1 に示す．

全手法で Adam を用いて最適化し，学習率はコサインアニーリングにより初期学習率 1.5×10^{-5} から学習回数が進むにつれ減衰させる．これにより，学習初期には比較的大きな更新を行い，学習後半では学習率を低下させることで方策の収束を安定させる．

PPO および提案手法では，1 イテレーションあたり 2048 件の和了決定を収集し，ミニバッチサイズ 256 で 2 エポック更新する．クリッピング率 $\epsilon = 0.2$，価値損失の重み $c_1 = 0.5$，勾配クリッピング閾値を 0.5，割引率 $\gamma = 1.0$，GAE の $\lambda = 1.0$ とした．目標エントロピーをイテレーションの進行に伴い 0.9 から 0.2 へ線形減少させ，エントロピー正則化係数 c_2 はこの目標に方策のエントロピーが追従するよう適応的に更新する．これにより，学習初期の探索を促しつつ収束を誘導する．

教師あり学習では，自分の和了判断を常に見送りとして対局データを収集する．1 イテレーションあたり 2048 件のデータをミニバッチサイズ 256 で 2 エポック更新し，交差エントロピー損失と MSE 損失の和を最小化する．

各手法について，10 回の独立した学習を行う．

4.5 性能評価

反実仮想報酬による見送り判断の学習改善の確認と，教師あり学習との有効性比較のため，学習済みモデルを用いて評価実験を行った．

表 1 学習パラメータ

パラメータ	PPO・提案手法	教師あり学習
初期学習率	1.5×10^{-5}	
バッチサイズ	2048	
ミニバッチサイズ	256	
エポック数	2	
イテレーション数	3000	
ϵ	0.2	—
c_1	0.5	—
勾配クリッピング閾値	0.5	1.0
γ	1.0	—
λ	1.0	—

表 2 各モデルの平均点

#	PPO	提案手法	教師あり	全和了
1	0.07	0.59	0.78	0.06
2	0.16	0.73	0.50	
3	0.31	0.49	0.66	
4	0.07	0.58	0.63	
5	0.35	0.70	0.32	
6	0.26	0.73	0.55	
7	0.46	−0.28	0.67	
8	0.39	0.52	0.43	
9	0.43	0.67	0.60	
10	0.40	0.50	0.52	
平均	0.29	0.52	0.57	0.06

打牌には学習時と同じ事前学習済みの共通モデルを使用する．打牌モデルは同じ局面では常に同じ打牌を選択する決定的なモデルである．和了判断は，PPO と提案手法では出力確率の高い方の行動を選択し，教師あり学習では $\hat{r}_{\text{skip}}$（式 6）が和了点を上回る場合に見送り，それ以外は和了を選択する．手法間の比較が公平になるよう対局の乱数を統一し，1000 万局を評価した．なお，常に和了を選択する全和了方策を比較対象として加え，和了判断を学習しない場合の基準性能を確認する．全和了方策は方策に乱数を含まず対局の乱数も固定されているため，評価を 1 回のみ行った．評価指標は 1 局あたりの平均点，見送り率，および見送り成功率とした．

4.6 結果と考察

表 2 に各手法の 1 局あたりの平均点を示す．提案手法（0.52 点）と教師あり学習（0.57 点）は PPO（0.29 点）を大きく上回り，提案手法と教師あり学習の差は小さかった．PPO の平均点は全手法中最も低く，見送り判断をほとんど学習できていないことが示唆される．なお，提案手法のモデル 7 のみ −0.28 点と負値を示した．

© 2026 Information Processing Society of Japan

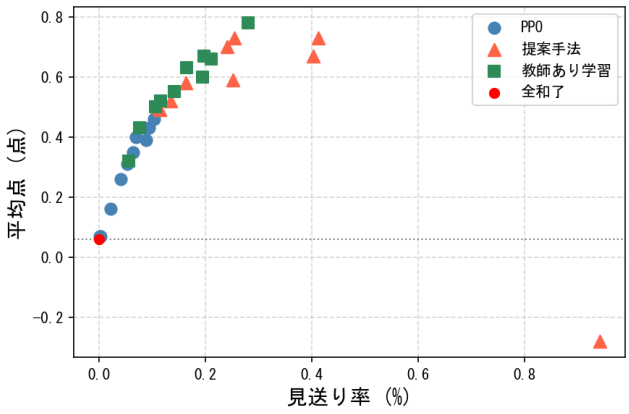


図 3　見送り率と平均点の関係

表 3　PPO の各モデルの見送り判断

#	平均点	見送り率	見送り数	平均見送り時得点	平均見送り後得点	見送り成功率
1	0.07	0.001	73	2000	3411	71.2
2	0.16	0.022	1119	1753	2716	66.1
3	0.31	0.053	2629	1807	2828	65.7
4	0.07	0.002	96	2000	3521	69.8
5	0.35	0.064	3207	1945	2923	63.9
6	0.26	0.041	2050	1918	3002	66.2
7	0.46	0.103	5165	1759	2607	64.0
8	0.39	0.088	4397	1718	2535	64.2
9	0.43	0.094	4686	1734	2590	64.7
10	0.40	0.069	3423	1982	3072	65.8

表 4　提案手法の各モデルの見送り判断

#	平均点	見送り率	見送り数	平均見送り時得点	平均見送り後得点	見送り成功率
1	0.59	0.251	12574	2524	3004	55.9
2	0.73	0.255	12765	2129	2710	59.2
3	0.49	0.114	5672	1808	2640	64.6
4	0.58	0.163	8129	1962	2672	61.3
5	0.70	0.241	12067	2004	2596	59.4
6	0.73	0.412	20625	2140	2515	54.5
7	−0.28	0.942	47348	2379	2310	44.5
8	0.52	0.135	6723	1989	2752	61.7
9	0.67	0.403	20164	2365	2715	52.3
10	0.50	0.110	5514	1788	2678	65.4

図 3 に見送り率と平均点の散布図を示す．見送り率は和了判断で見送りを選択した割合（%）である．PPO の点は見送り率 0%付近に集中しており，通常報酬の強化学習では見送り行動を学習しづらいことが確認できた．提案手法は 3 手法の中で最も高い見送り率を示し，反実仮想報酬の導入によって見送り行動が促進されたことが分かる．散布図全体では見送り率が高いほど平均点も高い傾向が見られるが，提案手法のモデル 7 のように見送り率が極端に高くなると逆に平均点が大きく下がることも分かった．見送り率 0.2〜0.4%程度で高い平均点が得られており，適切な見送り判断が得点向上に寄与することが確認できた．

表 3・表 4 に PPO および提案手法の各モデルの見送り判断の詳細を示す．平均点は表 2 と同じ 1 局あたりの点数の平均，見送り率は和了判断で見送りを選択した割合（%），平均見送り時得点は見送りを選択した局面で和了していれば得られたはずの点数の平均，平均見送り後得点は見送りを選択した後の対局で実際に得た点数の平均，見送り成功率は見送り後得点が見送り時得点を上回った見送り局面の割合（%）である．

PPO の見送り率は 0〜0.10%と低く，通常の強化学習では見送り判断をほとんど学習できていないことが改めて確認できる．PPO の見送り成功率は 63.9〜71.2%と高いが，これは PPO が低得点で明らかに見送りが有利な場面で見送りを選択しているためだと考えられる．一方，提案手法の見送り率は 0.11〜0.94%と高く，反実仮想報酬の導入によって見送り行動が促進されたことが分かる．提案手法のモデル 7 を除く全モデルで平均見送り後得点が平均見送り時得点を上回っており，提案手法が有効な見送り判断を学習できていることが分かる．

以下，提案手法について述べる．見送り率の低いモデル 3・4・8・10 では平均見送り時得点が 1788〜1989 点と低く，低得点の和了を見送る局面が中心であったと考えられる．一方，見送り率が高いモデルでは平均見送り時得点が 2000 点を超えており，より高得点の和了局面でも見送りを選択していることが分かる．ただし，見送り率が高いモデルほど見送り成功率が低い傾向があり，見送りが有利でない局面でも見送りを選択する無理な見送りが増えている可能性も考えられる．モデル 7 を除く全モデルで見送り成功率は 52.3〜65.4%と 50%を超えており，見送り判断が平均的に得点向上に寄与していることが確認できる．ただし，最も高いモデルでも 65.4%にとどまり，約 4 割の見送りは失敗している．これは見送りが本質的に不確定な意思決定であることを示しており，見送り判断の学習が依然として困難であることを示唆している．モデル 7 は見送り成功率が 44.5%と 50%を下回っており，平均見送り後得点（2310 点）が平均見送り時得点（2379 点）を下回っていることから，得点上不利な局面でも見送りを選択していたことが平均点を負値に押し下げた原因と考えられる．これは，学習初期に形成された見送り方策が，学習率の減衰によって十分に修正されなかった可能性が考えられる．ただし，原因の特定には学習過程の詳細な分析が必要である．教師あり学習は常に見送りを選択してデータを収集するため，収集データが方策の変化に依存せず，10 モデル全てで安定した正の平均点を達成した．

5.　まとめと今後

本研究では，麻雀の和了判断を対象に，和了と見送りの両結果をシミュレートしてその得点差を反実仮想報酬として用いる PPO 学習手法を提案した．簡易麻雀環境を用い

© 2026 Information Processing Society of Japan

た比較実験の結果，通常の PPO では見送りが有利な局面をほとんど学習できないのに対し，提案手法は PPO を大きく上回る平均点を達成し，教師あり学習に近い性能が得られた．一方で，提案手法では一部のモデルで見送りを過剰に選択し平均点が負値に落ち込む不安定さも見られ，学習の安定化が課題として残った．

今後の課題を以下に挙げる．第一に，学習の安定性向上である．提案手法では一部のモデルで学習が不安定になったが，その原因は明らかでなく，詳しい調査が必要である．教師あり学習で得られる見送りの期待値をベースラインとして組み合わせるなど，安定化手法の検討も求められる．第二に，実験規模の拡大である．本研究では単一の対戦相手・ランダム配牌という限定的な設定で 10 回の学習を行ったため，対戦相手や配牌分布を変えた追加実験や，より多数回の試行による統計的評価が必要である．第三に，本格的な麻雀ルールへの拡張である．本研究で用いた簡易麻雀は実際の麻雀と大きく異なる．本実験では，反実仮想報酬によって通常の強化学習では学習しにくい見送り判断が促進されること，および適切な見送り判断が得点向上に寄与することを確認できた．一方，実麻雀では 4 人対戦・立直・副露・局収支・順位状況など和了判断に影響する多くの要素が存在し，見送りが有利になる条件や適切な見送り率も本研究の設定とは異なる可能性が高い．したがって，本研究の知見をそのまま実麻雀の和了判断に適用することはできず，4 人対戦・複数種類の牌・立直や副露の導入など段階的な拡張による検証が今後必要である．第四に，他の問題への応用である．一方の行動が常に確定的な正の報酬をもたらす環境では，エージェントはその行動の利用に収束しやすく，もう一方の行動が有利な局面を学習することが困難になる．これは強化学習における探索と利用のジレンマの一形態であり，麻雀の和了判断に固有の問題ではない．類似の報酬構造を持つ強化学習問題に対して，提案した反実仮想報酬の考え方が有効かどうか，今後検討したい．

参考文献

[1] J. Li et al.: Suphx: Mastering Mahjong with Deep Reinforcement Learning, arXiv:2003.13590, 2020.
[2] J. Schulman et al.: Proximal Policy Optimization Algorithms, arXiv:1707.06347, 2017.
[3] J. Schulman et al.: High-Dimensional Continuous Control Using Generalized Advantage Estimation, arXiv:1506.02438, 2016.
[4] K. He et al.: Deep Residual Learning for Image Recognition, Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.

© 2026 Information Processing Society of Japan