情報処理学会論文誌 Vol.60 No.7 1325–1336 (July 2019)

自動対戦棋譜の教師あり学習による翻数予測に基づく麻雀プレイヤ

水上 直紀[1,a)] 鶴岡 慶雅[2,b)]

受付日 2018年8月27日, 採録日 2019年4月9日

概要：自己対戦を利用することで囲碁や将棋といった完全情報ゲームにおいて人間プレイヤを超えるコンピュータプレイヤが示されている．一方で不完全情報ゲームの分野である麻雀ではこのような研究は行われていない．そこで本論文では自動対戦棋譜の教師あり学習による麻雀プログラムを構築する方法について述べる．まず，人間の牌譜から教師あり学習によりコンピュータプレイヤを構築し，このプレイヤ同士を対局させることにより牌譜を生成する．次に，この牌譜を用いて手牌から和了の翻数を予測するモデルを機械学習により構築する．最終的に，この翻数予測モデルの出力と期待最終順位を用いて点数状況を考慮する麻雀プログラムを構築した．評価実験により，得られた翻数予測モデルは 4 翻以上の高い翻数の成功率を約 1 ポイント向上させることを確認した．

キーワード：強化学習，不完全情報ゲーム

Computer Mahjong Players Based on Winning Score Prediction by Supervised Learning from Self-play

Naoki Mizukami[1,a)] Yoshimasa Tsuruoka[2,b)]

Received: August 27, 2018, Accepted: April 9, 2019

Abstract: Recent reinforcement learning algorithms demonstrate that they can successfully achieve superhuman performance in the perfect information game such as Go or Shogi. However, in the domain of imperfect information game such as Mahjong, there is not much research using reinforcement learning. Therefore, this paper describes a method for building a mahjong program by supervised learning from self-play. First, a computer player is built from supervised learning from human's game records. The computer player's game records is generated by self-play. We train models that predict winning scores of a player's hand using game records. Our program decides moves based on the outputs of the prediction models and the expected final ranks. The program can predict future rewards and has obtained a skill for winning with high scores.

Keywords: reinforcement learning, imperfect information game

1. はじめに

近年，囲碁や将棋といった完全情報ゲームにおいてゲーム AI は人間のトッププレイヤを上回る実力を獲得した [1]．これらのゲーム AI の実力が向上した要因の 1 つは，自己対戦による強化学習によって指し手や局面に関する精確な評価関数の学習に成功したことにある．一方，不完全情報ゲームである麻雀では人間の牌譜を用いて学習を行うことで AI は上位プレイヤ並みの実力を獲得している [2]．

本研究の対象となる麻雀は「不完全情報」や「多人数プレイ」といった，AI の開発を難しくする性質を持っているが，「繰り返しゲーム」の性質もまた麻雀を複雑にしている要因である．麻雀は 1 局ごとに役に応じた点数を獲得し，全部で 4 または 8 局行う．すべての局が終了した時点で最も多くの点を持っているプレイヤが麻雀の勝者と見なされる．そのため現状の得点状況を考慮して手を決定する必要がある．

繰り返しゲームとしての麻雀を対象にした研究として，筆者ら [2] は期待最終順位に基づいたプログラムを構築し

1 HEROZ 株式会社
HEROZ, Inc., Minato, Tokyo 108–0014, Japan
2 東京大学大学院工学系研究科
Graduate School of Engineering, The University of Tokyo, Bunkyo, Tokyo 113–8656, Japan
a) mizukami@heroz.co.jp
b) tsuruoka@logos.t.u-tokyo.ac.jp

た．期待最終順位とは現在の局面から予想される最終的な順位のことである．この研究では期待最終順位を出力するモデルを牌譜を教師情報として学習を行った．そしてその出力である期待最終順位をシミュレーションの報酬とすることで点数状況を考慮したプログラムの構築に成功した．その結果，局単位の収支は減少する一方，トップ死守率などの順位にかかわる指標が向上し，それにともない人間との対戦成績も向上した．

この研究では実力を判断する基準は人間の牌譜との一致率や対戦成績が用いられている．牌譜との一致率で比較するとそのプログラムの選択する手は従来のプログラムよりも低下している．しかし対戦成績で比較すると実力は向上している．そのため現在のプログラムは一部の局面において人間よりも良い手を選択していると考えられる．そこで本論文ではプログラムどうしの対戦の牌譜を生成し，その牌譜を基に機械学習を行うことで元のプログラムのさらなる実力の向上を試みた．

確かに筆者ら [2] の研究により点数状況を考慮した手を選択することが一部の局面において可能になった．しかしこのプログラムが明らかに点数状況を考慮していない悪手を選択する問題も存在する．それは特に 1 局の序盤において効率の良い和了を目指していないことである．麻雀において一般に安い手は和了しやすく，高い手は和了しにくい．そのため人間プレイヤは役の難易度と和了したときの点数状況のバランスを考慮して手を選択している．たとえば最終局において 4 位のとき，人間プレイヤは相手の点数を逆転する手を作ろうとする．また 1 位のとき，高得点の和了の価値は低くいため時間をかけて高得点を目指す必要はない．そのため得点の高低よりも和了そのものを目指す．筆者ら [2] の研究において構築されたプログラムは局終盤の自分の和了したときの点数がほぼ確定した状態において攻めるべきか降りるべきかということは理解したものの，序盤において何点の手を作るべきかという長期的な戦略は苦手である．この原因は 1 局の序盤で用いるモデルでは現在の手牌が将来的に何点になるかを考慮していないモデルだからである．そこで本研究では自己対戦による棋譜から教師あり学習を行い，現在の手牌から和了できる点数を予測するモデルを構築し，それを用いて効率の良い和了を行う麻雀プログラムを構築する手法を提案する．

本論文は以下の構成になっている．初めに 2 章で麻雀のルールと用語を述べる．次に 3 章で関連研究，4 章で本研究のベースラインとなる 1 人麻雀政策について述べる．提案手法として，5 章で自己対戦の棋譜生成と教師あり学習の方法，6 章で提案手法の対戦結果について述べる．最後に 7 章で本研究の結論について述べる．

2. 麻雀のルール

本章では麻雀の得点に関するルールについて簡単に述べる．麻雀は 14 枚の牌（**手牌**）を組み合わせて役（特定の構成）を作り，役に応じた点数を得るゲームである．この点数を得る行動を**和了（ホーラ）**と呼ぶ．

点数は**翻（ハン）**と**符（フ）**によって決まる．どちらも牌の組合せによって決まるが，点数は翻数に大きく左右されるため基本的には翻数を大きくすることを念頭に手を進める．各役はその難易度によって翻数が決められており，一般的には難易度の高い役ほど翻数が高い．複数の役が成立した場合は，その翻数の合計値を点数計算に用いる．本研究で用いる役を**表 1** に示す．役を作るためにプレイヤは自分の手番において引いた牌を利用する**ツモ**と相手の捨て牌を利用する**鳴き**を用いる．

表 1　役一覧
Table 1　List of winning hand.

翻数	役名
一	門前清自摸和（メンゼンツモ），リーチ
	槍槓（チャンカン），嶺上開花（リンシャンカイホウ）
	海底（ハイテイ），断幺九（タンヤオ）
	一盃口（イーペーコー），平和（ピンフ），役牌（ヤクハイ）
	一発（イッパツ），ドラ，赤ドラ，裏ドラ
二	混全帯幺九（チャンタ），一気通貫（イッツー）
	三色同順（サンショクドウジュン），ダブルリーチ
	三槓子（サンカンツ），対々和（トイトイ）
	小三元（ショウサンゲン），混老頭（ホンロウトウ）
	三暗刻（サンアンコウ），三色同刻（サンショクドウコウ）
	七対子（チートイツ）
三	純全帯幺九（ジュンチャン），混一色（ホンイツ）
	二盃口（リャンペーコー）
六	清一色（チンイツ）
役満（十三）	天和（テンホウ），地和（チーホウ）
	大三元（ダイサンゲン），四暗刻（スーアンコウ）
	字一色（ツーイーソー），清老頭（チンロウトウ）
	国士無双（コクシムソウ），緑一色（リュウイーソー）
	四喜和（スーシーホー），四槓子（スーカンツ）
	九蓮宝燈（チュウレンポウトウ）

麻雀は 4 人のうち誰かが役を構成し和了すると 1 局が終了し，これを定められた回数の局数を行う（通常は 4 または 8 回）．最初の持ち点は 25,000 から開始し，最終局（**オーラス**）を終了したときの得点の多さに応じて順位が決まる．

和了の方法は 2 種類あり役を完成するタイミングにより呼び方が異なる．自分の手番において引いた牌で和了することを**ツモ和了**あるいは省略してツモと呼ぶ．また相手の捨てた牌で和了することを**ロン**と呼ぶ．またロンされることを**放銃**と呼ぶ．ツモの場合，点数を残りの 3 人が和了点数を分割して支払う．またロンの場合，放銃したプレイヤがすべての和了点数を支払う．

3. 関連研究

探索を用いることである程度の向聴数（シャンテン数）において効率的な手を選択する手法が提案されている．栗田ら [3] は有向非巡回グラフを用いて 1 人麻雀の探索アルゴ

リズムを提案した．このアルゴリズムは手牌と打牌回数とゲームの進行フェーズに基づいて有向非巡回グラフの接点をまとめる．このグラフを用いることと終端ノードでの報酬を設定することで麻雀の高い専門知識を使用することなく，3，4 向聴数の手牌の効率的な手を選択することが可能になった．

不完全情報ゲームのポーカーの 1 種であるテキサスホールデムでは，Heinrich ら [4] は強化学習を用いることで **CounterFactual Regret minimization**（**CFR**）[5] をもとにしたプレイヤに迫る実力を得たと報告している．この研究では自己対戦を行い，行動価値関数と過去のプレイヤの行動をそれぞれ Q 学習と教師あり学習でモデルを構築する．特徴として今までの研究では局面の抽象化を行うことが主流であったが，この研究ではカードやチップなどの局面の状態を事前知識なしでエンコードし，そのままニューラルネットの入力とする．行動価値関数と過去のプレイヤの行動の確率分布をこのニューラルネットワークの出力としている．

囲碁において AlphaGo では指し手を確率分布で持つモデルを勝率に変換する方法として強化学習が用いられた [1]．Silver らは畳み込みネットワークで構成される政策ネットワークと線形で構成されるロールアウトポリシを用いてランダムに局面を生成し，その局面の勝率を予測するモデルを構築した．

4.　1 人麻雀政策

本章では本研究で重要な役割を果たす 1 人麻雀政策 [6] について述べる．この政策は人間の牌譜から教師あり学習によって得られたモデルであり，単純に和了に向かう行動だけを選択する．以前の研究では 1 人麻雀の手 [6] と呼んでいたが，手という表現は麻雀において複数の意味を持つため本研究では 1 人麻雀政策と呼ぶ．

麻雀では和了に向かう目的だけでなく放銃を避ける目的など異なる目的を意図してプレイヤは牌を選択する．そのためすべての牌譜を学習に用いることは 1 人麻雀政策の学習には適切ではない．そこで 1 人麻雀政策の学習ではすべての牌譜の中から和了に向かう手を含んだ牌譜を抽出する．しかし和了に向かう手は人間同士でも判断基準が異なるため，選択した意図を正確に読み取るのは困難である．そのため客観的に判断可能な誰もリーチを宣言していない状態で和了したプレイヤまたは最初にリーチを宣言したプレイヤの牌譜を用いる．人間の牌譜はインターネット麻雀サイト天鳳 [7] の鳳凰卓の牌譜である [*1]．

学習では牌譜中で実際に選択された牌と現在の重みベクトルから選択される牌の評価を近づけるように重みベクトルの調整を行う．具体的な学習方法は与えられた手牌から 1 つ牌を切ったときの手牌を想定し，その手牌から抽出される特徴量と重みベクトルの内積によって評価値を計算する．これをすべての牌について行い一番評価値の高い牌と実際の牌譜で切られた牌が異なる場合に重みベクトルを更新する．これを繰り返すことで重みベクトルは牌譜と同じ手を選択するように調整される．牌譜自体は和了に向かう手であるため，そこから得られる 1 人麻雀政策は和了に向かう手である．教師あり学習の方法として平均化パーセプトロン [8] を用いた．

筆者らの研究 [6] ではこの特徴量では役の表現力が乏しく，当時のプログラムは実際の対局において役を完成できない鳴きを頻繁に行った．そのためこの特徴量は，本研究では役を作るための特徴量として適さないと判断した．そこで本研究では特徴量の改善を行った．詳細は**表 2**，**表 3** に示す．多くの特徴量はいくつかの要素の組合せで構成されている．すなわち組合せのすべての要素が満たされるときにベクトルの値が 1 となるような特徴量である．特徴量の数は合計で 6,661,309 である．

改善した特徴量を用いた実験の結果を**表 4** に示す．表中のツモ局面とは打牌を選択する局面であり，鳴く局面とは牌譜中のプレイヤが鳴いた局面であり，鳴かない局面とは牌譜中のプレイヤが鳴ける局面で鳴かなかった局面を指す．その鳴く局面において 1 人麻雀政策と牌譜の晒した牌が一致し，さらにそこから切った牌も一致した局面数が完全一致数である．晒した牌は一致したものの切った牌が異なる場合は鳴きのみ正解数としてカウントされる．鳴いたことが重要であるため鳴きに関する正解率は完全一致数と鳴きのみ正解数の合計を鳴き局面数で割った値とする．

テストデータが異なるため単純な比較はできないが，以前の結果 [6] では鳴く局面での正解率は 84.2%であり，鳴かないときの正解率は 90.7%であった．このことから特徴量を改善することで一致率が向上し，役が完成しなくなる鳴きは減少したと考えられる．本研究ではこの 1 人麻雀政策を用いて実験を行う．

1 人麻雀政策は人間の牌譜から学習しているため，ある手牌が与えられたときに牌譜中のプレイヤが最も選択するであろう牌を切る．そのため打牌の基準となる評価値は牌の選択されやすさであり，現状の手牌が何翻で和了できるかはまったく考慮されていない．そのため 1 人麻雀政策に従うと平均的な局面においては悪手であることは少ないが，オーラスといった得点状況に応じて最善手が変わるケースにおいて悪手となりうる．次の章はこの原因を解消する手法について述べる．

5.　自己対戦の棋譜を用いた教師あり学習による役作り

前章で述べたように 1 人麻雀政策は現在の手牌の将来的な報酬を理解していない．この章では自己対戦の棋譜を用

*1　2009 年 2 月 20 日から 2015 年 12 月 31 日までに行われた対局

表 2　1 人麻雀プレイヤの特徴量

Table 2　Features of one player mahjong moves.

特徴量	次元数
通常手，七対子，国士無双の向聴数	$15 + 7 + 14 = 36$
副露数	5
向聴数，副露数（フーロ数）	$15 \times 5 = 75$
リーチが可能か	2
向聴数，副露数，min(受け入れ枚数, 20)	$15 \times 5 \times 21 = 1{,}575$
副露した種類	136
役牌の刻子の数	5
向聴数の悪化しない頭の数，役牌の対子の数	$6 \times 6 = 36$
役牌の刻子があるか，向聴数の悪化しない頭の数，役牌の対子の数， 浮いた役牌の数，min(向聴数, 4)，副露数	$2 \times 6 \times 6 \times 16 \times 5 \times 5 = 28{,}800$
色の中で最も多い色の数+染め役は不可能，副露数，混一色または清一色	$(14 + 1) \times 5 \times 2 = 150$
min(ドラ+赤ドラの数, 3)	4
役がある，ない，片上がり	3
min(向聴数, 3)，役がある，ない，片上がり，巡目	$4 \times 3 \times 18 = 216$
min(向聴数, 3)，副露数，振聴か，min(役のある待ち牌の数, 7)	$4 \times 5 \times 2 \times 8 = 320$
両面を優先したときの両面+面子の数，向聴数	$7 \times 15 = 105$
面子+ターツ+ターツ候補，向聴数	$11 \times 15 = 165$
両面を優先したときの両面+面子の数，面子+ターツ+ターツ候補，向聴数	$7 \times 11 \times 15 = 1{,}155$
min(全帯幺九の向聴数, 4)，max(全帯幺九の枚数 − 6)，全帯幺九のメンツまたはターツ候補， min(受け入れ枚数/4, 4)，min(全帯幺九の向聴数−向聴数, 3)	$5 \times 8 \times 8 \times 5 \times 4 = 6{,}400$
min(全帯幺九の向聴数, 4)，max(全帯幺九の枚数 − 6)，2,378 の暗刻があるか，副露数， 両面を優先したときの面子の数，両面を優先したときの両面の数，愚形の数， 面子の減らない 19 字牌の頭の数，面子の減らない 2,378 字牌の頭の数	$5 \times 2 \times 5 \times 5 \times 4 \times 8 \times 2 \times 2 = 32{,}000$
ドラの種類（19，28，37，46，5，役牌，オタ風），ドラの数， 見えているドラの数，現在の巡目/2，赤ドラの数	$7 \times 4 \times 4 \times 8 \times 4 = 6{,}400$
min(全帯幺九の向聴数, 4)，max(全帯幺九の枚数 − 6)，全帯幺九のメンツまたはターツ候補， min(受け入れ枚数/4, 4)，min(全帯幺九の向聴数−向聴数, 3)	$5 \times 8 \times 8 \times 5 \times 4 = 6{,}400$
両面を優先したときの両面+面子の数，愚形の数，両面対子の数，愚形対子の数，浮き牌があるか， 暗刻があるか，頭の数，min(向聴数, 3)，完全一，二向聴，そうでない，リーチが可能か，巡目/3	$8 \times 8 \times 4 \times 4 \times 2 \times 2 \times 4 \times 4 \times 3 \times 2 \times 7$ $= 2{,}752{,}512$
面子を優先したときの両面+面子の数，愚形の数，両面対子の数，愚形対子の数， 浮き牌があるか，暗刻があるか，頭の数，min(向聴数, 3)，リーチが可能か，巡目/3	$8 \times 8 \times 4 \times 4 \times 2 \times 2 \times 4 \times 4 \times 2 \times 7$ $= 917{,}504$
min(色の中で最も多い色の数の向聴数, 4)，両面を優先したときの面子の数， 両面対子+愚形対子の数，副露数	$5 \times 5 \times 8 \times 8 \times 5 = 8{,}000$
min(19 字牌抜いたときの向聴数, 4)，両面を優先したときの面子の数， 両面を優先したときの面子+愚形の数，タンヤオのドラの数，副露数，巡目/3, タンヤオの向聴数=向聴数か，min(19 字牌の受け入れ枚数, 2)，max(タンヤオ牌 − 11, 0)	$5 \times 5 \times 8 \times 4 \times 5 \times 6 \times 2 \times 3 \times 3 = 432{,}000$
両面を優先したときの両面+面子の数，愚形の数，七対子の向聴数， min(向聴数, 3)，完全一，二向聴，そうでない，リーチが可能か， 浮き牌の種類（19，28，34,567，字牌），その浮き牌の枚数	$8 \times 8 \times 8 \times 4 \times 3 \times 2 \times 4 \times 4$ $= 196{,}608$
両面を優先したときの両面+面子の数，愚形の数，二度受けの両面の数，二度受けの愚形の数， min(向聴数, 3)，完全一，二向聴，そうでない，リーチが可能か	$8 \times 8 \times 4 \times 4 \times 4 \times 3 \times 2 = 24{,}576$
min(向聴数, 4)，七対子の向聴数，向聴数の悪化しない頭の数， 両面を優先したときの両面+面子の数，リーチが可能か，完全一，二向聴，そうでない，	$5 \times 8 \times 8 \times 8 \times 2 \times 3 = 15{,}360$
min(向聴数, 3)，役牌の刻子があるか，役牌の対子があるか，両面を優先したときの面子の数， 両面+愚形，向聴数の悪化しない頭の数，min(副露数, 3)，役がある，ない，片上がり	$4 \times 2 \times 2 \times 5 \times 8 \times 4 \times 4 \times 4 \times 3$ $= 122{,}880$
両面を優先したときの両面+面子の数，愚形の数，浮き牌の最も外側の種類（19，23，3,456，字牌）， 頭と頭の組合せ，頭の数，完全一，二向聴，そうでない	$8 \times 8 \times 5 \times 16 \times 4 \times 3 = 61{,}440$
min(向聴数, 4)，七対子の向聴数，向聴数の悪化しない暗刻の数，副露数， チーがあるか，完全一，二向聴，そうでない，	$5 \times 8 \times 5 \times 5 \times 2 \times 3 = 6{,}000$

いた教師あり学習を用いてこの問題に取り組む．

図 1 は提案手法の全体像である．前章の人間の牌譜から 1 人麻雀政策を作成し，この 1 人麻雀政策を利用して AI どうしの牌譜を生成，この AI どうしの牌譜から翻数予測モデルを構築する．すなわち牌譜の生成には 1 人麻雀政策に従う AI が 4 体の 1 人麻雀でない 4 人麻雀を行う．この翻数予測モデルに以前の研究で用いた期待最終順位と和了できないペナルティやゲーム木の探索を組み合わせることで，提案手法となる序盤アルゴリズムを構築する．対戦実験では以前の研究の麻雀プログラムと新しい麻雀プログラ

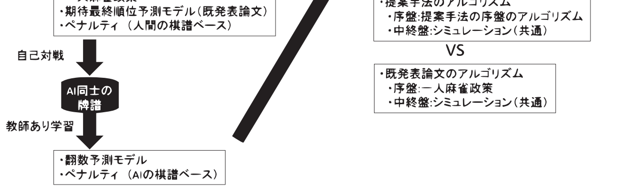


図 1　提案手法の全体像

Fig. 1　Overview of proposed method.

表 3 1 人麻雀プレイヤの特徴量
Table 3 Features of one player mahjong moves.

各字牌に対して見えている数，持っている数，ドラか，max(色の中で最も多い色の数 − 6, 0)，両面を優先したときの両面+面子の数，max(巡目, 8)，字風，場風，東南西北+三元牌	$5 \times 5 \times 2 \times 8 \times 8 \times 8 \times 4 \times 35 \times 5 = 1{,}536{,}000$
各数牌に対して数牌の種類（19，28，37，46，5），持っている枚数，ドラとの近さ（0，1，2，3，違う色）	$5 \times 5 \times 5 = 125$
連続する n 種類の数牌の持っている枚数の組合せ（$n = 2$〜6），リーチが可能	$(100 + 500 + 1{,}860 + 8{,}634 + 23{,}760) \times 2 = 69{,}708$
各色の 1 から 9 の組合せ．各数字は最高で 2	19,472
暗刻の数，対子の数，刻子にならない対子の数	$5 \times 7 \times 2 = 70$
刻子の数，対子の数，刻子にならない対子の数+対々和ができない	$5 \times 7 \times 2 + 1 = 71$
min(タンヤオの向聴数, 4)，min(タンヤオの向聴数−向聴数, 3)，max(タンヤオの枚数 − 9, 0)，副露数，max(タンヤオの頭, 3)+タンヤオができない	$5 \times 4 \times 5 \times 4 + 1 = 401$
ドラの数，タンヤオのドラ，min(タンヤオの向聴数, 4)，min(タンヤオの向聴数−向聴数, 3)，タンヤオができるか	$4 \times 4 \times 5 \times 4 \times 2 = 640$
タンヤオができるか，全帯幺九ができるか，min(19 字牌の受け入れ枚数, 3)，ありーチができるか，副露数	$2 \times 2 \times 4 \times 2 \times 5 = 160$
3 色に最も近い枚数，向聴数，副露数	$10 \times 14 \times 5 = 700$
各 3 色の可能性について，123,789 か，各数字を持っているか，min(向聴数, 3)，3 色に近づく受け入れがあるか，リーチができるか	$2 \times 512 \times 4 \times 2 \times 2 = 16{,}384$
各一通の可能性について，各数字を持っているか，min(向聴数, 4)，副露数，両面+面子+愚形 ≧ 5，両面+面子 ≧ 4，完全一，二向聴，そうでない	$512 \times 5 \times 5 \times 2 \times 2 \times 3 = 153{,}600$
一通に最も近い枚数，面子，両面，愚形−頭の数，頭があるか，副露数，一通に近づく受け入れがあるか	$10 \times 5 \times 8 \times 8 \times 2 \times 5 \times 2 = 64{,}000$
各風牌の枚数，最高 3 枚	$4 \times 4 \times 4 \times 4 = 256$
各三元牌の枚数，最高 3 枚	$4 \times 4 \times 4 = 64$
各和了牌について，枚数，翻数，巡目/3	$4 \times 9 \times 7 = 252$
各和了牌について，牌の種類（19，28，37，46，5，ダブ東南，役牌，オタ風），枚数，翻数，ツモとロンでの翻の差，リーチか，ドラ待ちか，筋待ちか，フリテンか	$8 \times 4 \times 8 \times 3 \times 2 \times 2 \times 2 \times 2 = 12{,}288$
min(役ありの和了牌数, 9)，min(役なしの和了牌数, 5)，副露数，七対子または国士無双か	$10 \times 6 \times 5 \times 2 = 600$
min(一向聴時の受け入れ, 31)，副露数，完全一，二向聴，そうでない	$32 \times 5 \times 3 = 480$
4 色の選んだ 3 色の受け入れ枚数（最大 20）までの組合せ	$20 + 231 + 1{,}771 = 2{,}022$

表 4 人間の牌譜との一致率
Table 4 Agreement rate of game records of expert human players.

局面の種類	牌譜の数	完全一致数	鳴きのみ正解数	正解率
ツモ局面	1,140,576	859,088	N/A	75.3
鳴く局面	68,397	46,945	11,731	85.8
鳴かない局面	252,666	240,450	N/A	95.1

ムとの対局を行う．新しい麻雀プログラムと以前の研究の麻雀プログラムの差は序盤アルゴリズムが提案手法の序盤アルゴリズムに置き変わったことである．すなわち中終盤は共通のアルゴリズムを用いた序盤だけ戦略の異なる麻雀プログラムどうしの対局である．

5.1 自己対戦の棋譜を用いた教師あり学習の方法

本研究の目的は現在の手牌から局終了時に得られる報酬を予測することである．そのための予測モデルを自己対戦の棋譜を用いた教師あり学習を用いて構築する．このモデルを翻数予測モデルと呼ぶ．翻数予測モデルの出力は特定の翻数を和了する確率とする．すなわち翻数予測モデルの出力された値は 1 人麻雀政策の評価値ではなく実際の麻雀における報酬を表現する値である．この値を用いることで本研究では現在の 1 人麻雀政策の将来的な報酬を理解していないという問題に取り組む．

具体的な手法は麻雀の特徴を考慮する．麻雀ではランダムな手を選択し続けても和了し報酬を得ることは困難である [9]．すなわち特徴量の重みを 0 もしくはランダムに初期化して強化学習を行う方法は報酬による学習が行われないため，うまくいかないと考えられる．しかしながら 4 章で述べたように 1 人麻雀政策はある程度は強いため，本研究ではこれを活用した手法を提案する．

本研究では AlphaGo [1] の評価値ネットワークの学習に用いられた手法を参考にする．AlphaGo の評価値ネットワークの出力は現在の局面の勝率である．これを実現するためには局面と最終結果のペアが必要になる．そこで AlphaGo は人間の棋譜から学習した政策ネットワークをベースに 2 段階の強化学習を行い評価値ネットワークに必要な学習局面を生成した．1 段階目では現在の政策ネットワークと過去の政策ネットワークを対戦させることで政策ネットワークを改良した．2 段階目は改良した政策ネットワークを用いて評価値ネットワークに必要な棋譜を生成する．牌譜の生成アルゴリズムは AlphaGo に対局中に 1 手だけ完全なランダムな手を打たせ，その後を終局まで改良した政策ネットワークを用いた手を打たせることで実現した．これらの対局を大量に行うことで評価値ネットワークに必要なランダムな手を打った局面とその最終結果を生成した．学習の手法を参考するという点において AlphaGo における 1 段階目の生成物は政策ネットワークであり，前

章の結果をふまえると特定の局面から終局まで行う役割を1人麻雀政策は果たせると考える．本研究は2段階目の評価値ネットワークの構築法を参考にする

AlphaGoの評価値ネットワークの構成に成功した理由は，多様な局面を生成した点と考えられる．同じような局面ばかり生成して評価値ネットワークの学習を行った場合，未知の局面に対する適切な評価を評価値ネットワークが出力することは困難である．一方で完全にランダムな手を何度も用いると現実に起こりにくい局面を生成してしまい学習の効率が悪い．適切にランダムな手を混ぜることで学習に適切な未知の局面を生成することが可能になったと考えられる．

麻雀においても多様な局面をサンプリングする方法が必要になる．しかし囲碁と麻雀ではゲーム性が大きく異なるため，単純に置き換えることはできない．以下，本研究の局面の生成法について述べる．

図2は牌譜生成のフローチャートである．初期化の対象はプレイヤの配牌，山，ドラ，自風，場風，教師データとして使用する巡目でありそれぞれランダムに決定される．自風は各風が1/4の確率で選択される．場風は本研究では東風戦しか行わないものの，一般的なルールでも成り立たせるために東南戦の西入まで考慮する．実際には西入することは少ないため場風が西のデータは東と南に比べ少なくても問題ない．これを考慮して本研究では場風は東と南は4/9，西は1/9の確率で選択する．

初めにプレイヤの手番について述べる．AlphaGoの場合，2人のプレイヤのうちどちらか1人だけがランダムな手を1回打つ．4人で行う麻雀において2人以上が完全にランダムな手を打つ場合，得られる牌譜に悪手が増え学習局面としては適切でない．また誰も完全にランダムな手を打たない場合，局面の多様性がないため学習に適切な局面

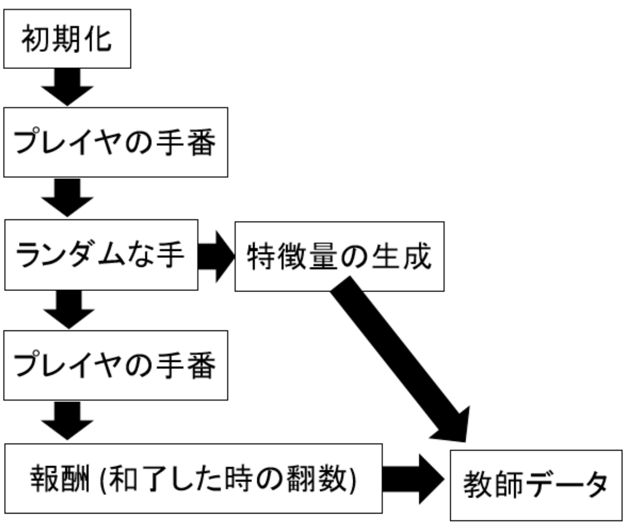


図2　牌譜生成のフローチャート
Fig. 2　Flowchart of generating game records.

数が増加しない．そのため本研究では特定のプレイヤ1人が1局において1回だけランダムな手を打ち，その局面を学習局面とする．すなわち生成されるすべての学習局面はランダムな手を打ったプレイヤからの視点での情報のみ使用し，残り3人のプレイヤの手牌の情報は学習局面に反映されない．また報酬もランダムな手を打ったプレイヤからの視点であり，ほかのプレイヤの和了は報酬を0として扱う．以下，その特定のプレイヤを自分プレイヤ，それ以外の残り3人を相手プレイヤと呼び，各プレイヤの手番での行動を説明する．

まず自分プレイヤの手番の挙動について説明する．上記で説明したようにランダムな手を選択し続けても和了することは困難であるため，そのような方法で生成された牌譜は学習にはあまり役に立たない．そこで自分の手番においてプレイヤは基本的に1人麻雀政策に従う．例外的に多様な局面を訪問するために1度だけ，自分プレイヤは1局の間に1度だけランダムな手を選択し，その直後の局面を学習局面に使用する．そしてそれ以降はその局面の精度の高い報酬を得るため1人麻雀政策に従う．

ランダムな手とは合法手の中から1人麻雀政策の評価値とは関係なく完全にランダムに選択された手である．ツモ局面であれば，各牌を切る（5と赤5は同一とする）手と加カンと暗カンが合法手にあたる．鳴ける局面におけるランダムな手は少し複雑であり鳴き方と切る牌の2つを考慮する必要がある．ポンやカンの場合は鳴くか鳴かないの2種類で済むが，チーの場合，牌の晒し方が最大3種類ある．また鳴いた後に切ることができる牌は，すでに完成しているメンツを鳴く行為（たとえば123から1または4をチーして1を切る）を禁止するルールが天鳳では採用されているため，すべての牌が選択可能とは限らない．そのため鳴ける局面における合法手は鳴き方と直後に切る牌を組合せによって決定する．

次に相手プレイヤの手番の挙動について説明する．得られた牌譜を使用して学習するため相手のプレイヤの挙動は牌譜の質に影響するため重要な要素である．また学習には報酬が0ばかりでなく密であるということも重要である．そこで本研究では相手プレイヤを2種類用意した．1つ目の相手プレイヤはツモ切りを続けるプレイヤである．このプレイヤはどのような局面においてもツモ切りを行い，鳴きや和了はいっさいしない．これにより相手に邪魔をされることなく和了できるため報酬が密であることが期待される．

2つ目のプレイヤは1人麻雀政策に従うプレイヤである．ツモ切りプレイヤが3人いる状況においてプレイヤが和了することは実際の麻雀と比較して容易である．現実の麻雀では他のプレイヤが和了するため，1人麻雀のように18回のツモで和了すればよいのではなく，局が終了するまでのツモはそれよりも少なくなる．この状況を実現するため，相手プレイヤの挙動を1人麻雀政策にすることで牌譜が生

成される環境を実際の麻雀の状況に近づけた．

ここでは報酬の設計について説明する．麻雀の点数は翻と符によって決まる．単純に翻と符をペアにして学習を行うとクラス分類数が増大し，学習が有効に働かない．そこで点数において符の影響は小さいため，これを無視する．すなわち報酬は 0（和了できない），1，2，3，4 翻以上とした．跳満以上は，狙ったとしても簡単に和了できないため本研究では 4 翻以上は同じカテゴリとして分類する．

麻雀では手牌が和了に近づくにつれて必要な牌の種類が少ないためツモによって手牌が更新されることが少なくなる．そのため終局までの手牌をすべて学習に用いると，同じような局面をモデルが学習してしまう．そのため学習局面は 1 局に対して 1 局面までとして 1 億局面を用意した．リーチを打った後の局面は合法手が 1 つしかないため教師局面には使用しない．また天和や地和で和了したときの局面は自分が 1 手も指していないため使用しない．

上記の方法によって現在の手牌から将来の報酬を予測に必要な牌譜生成を行う．しかしながら実際の麻雀と比較し考慮できていない部分も存在する．

- 1 人麻雀政策はリーチするかどうかは判断できないため，リーチが宣言できる局面においてリーチはすべて宣言するとした．
- 相手プレイヤが降りるといった戦略をとらないので，切られにくい牌を待つ戦略などが不当に高く評価される可能性がある．
- 1 手しかランダムな手を混ぜていないため，無理やり特定の役を狙うことは考慮に入れていない．
- 和了することが重要であるため和了が可能なときはすべて和了する．

5.2 翻数予測モデルの学習

ここでは生成した教師データを用いて翻数予測モデルを学習する方法について述べる．予測モデルの出力形式として回帰と分類が考えられる．回帰モデルを用いると出力結果は少数を含むことになるが，その値は実際の麻雀のゲームに存在せず扱いにくい．そこで翻数予測モデルは分類モデルを利用する．

分類モデルを利用するものの翻数の数字自体は大小関係が成り立ち無関係ではない．つまり 1 翻を和了するモデルの学習に 2 翻で和了した学習局面を正例とするか負例とするかは自明でない．そこで本研究では 2 種類の方法を用意した．1 つ目はちょうど特定の翻数を和了できるかどうか学習するモデルである．つまり 1 翻を和了するモデルの学習に 2 翻で和了した学習局面を負例として扱う．予測する結果は牌譜生成の報酬をもとに 0（和了できない），1，2，3，4 翻以上の 5 種類とする．現在の手牌から予想される翻数を予測するということは多クラスロジスティック回帰モデルを使用することで 5 クラスの多クラス分類問題としてとらえることができる．

出力としてソフトマックス関数を使用することにより各翻の和了できる確率として出力することができる．ソフトマックス関数は次の式で表現される．

$$P_{mc}(\mathbf{x}, h) = \frac{\exp(\mathbf{w}_h^{\mathrm{T}}\mathbf{x})}{\sum_{i=0}^{4}\exp(\mathbf{w}_i^{\mathrm{T}}\mathbf{x})} \tag{1}$$

ここで $\mathbf{x}$ は現在の手牌を表す特徴ベクトル，h は翻数である．$\mathbf{w}_h$ は各翻数 h の特徴量に対しての重みベクトルである．

目的関数は次の式で表現した．

$$L(\mathbf{w}) = -\sum_{i=1}^{N}\sum_{h=0}^{4} c_{i,h}\log(P_{mc}(\mathbf{X}_i, h)) + \frac{\lambda|\mathbf{w}|^{\mathbf{2}}}{N}$$

ここで N は学習データの事例数，$\mathbf{X}_i$ は i 番目の学習事例，$c_{i,h}$ は学習事例の結果と各翻数に対応する 2 値（1 または 0）のラベルである．λ は学習データに過学習することを防ぐ正則化の係数である．本研究では λ を 0.01 とした．

2 つ目は特定の翻数以上を和了できるかどうか学習するモデルである．つまり 1 翻を和了するモデルの学習に 2 翻で和了した学習局面を正例として扱う．手牌から予想される翻数をロジスティック回帰モデルを 4 つ構築することで表現する．4 つのモデルはそれぞれ 1 翻以上，2 翻以上，3 翻以上，4 翻以上を和了できるかどうかを予測する．

$$P_{bc}(\mathbf{x}, h) = \frac{1}{1+\exp(\mathbf{w}_h^{\mathrm{T}}\mathbf{x})} \tag{2}$$

目的関数は次の式で表現した．

$$L(\mathbf{w}) = -\sum_{i=1}^{N} c_{i,h}\log(P_{bc}(\mathbf{X_i})) + (1-c_{i,h})\log(1-P_{bc}(\mathbf{X_i})) + \frac{\lambda|\mathbf{w}|^{\mathbf{2}}}{N}$$

これら 2 つの重みベクトルの学習は確率的勾配降下法の 1 種である FOBOS [10] を用いて学習を行う．学習率は Adagrad [11] を用いて決定する．$\mathbf{x}$ は 1 人麻雀政策の学習に使用した特徴量と同じ特徴量を使用する．以後，式 (1) を用いたプレイヤを **Multi Class Player**（**MCP**）と呼び，式 (2) を用いたプレイヤを **Binary Class Player**（**BCP**）と呼ぶ．

5.3 提案手法の結果

翻数予測モデルが実際に有効に活用できるかを調べるために評価を行う．この評価では学習したモデルを使用して配牌とツモから特定の翻数（1，2，3，4）で和了できたかを調べる．特定の翻数もしくはそれ以上の翻数で和了した場合を成功とする．評価基準は成功率，すなわち成功した数と試行回数の商である．評価時の相手は牌譜生成時の相手と同じである．すなわち牌譜生成時の相手がツモ切りであ

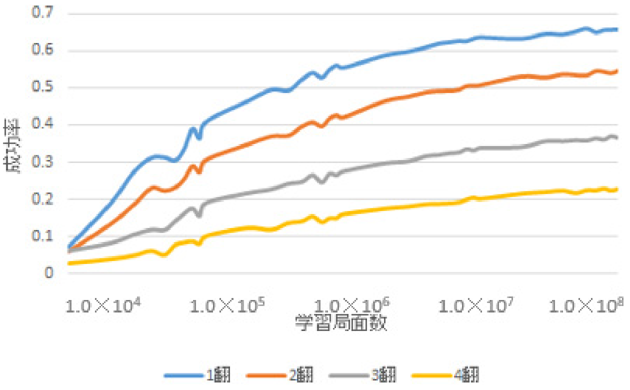


図 3　MCPvs ツモ切りにおける局面数と各翻数の成功率
Fig. 3　Success rate of MCPvs Tsumogiri.

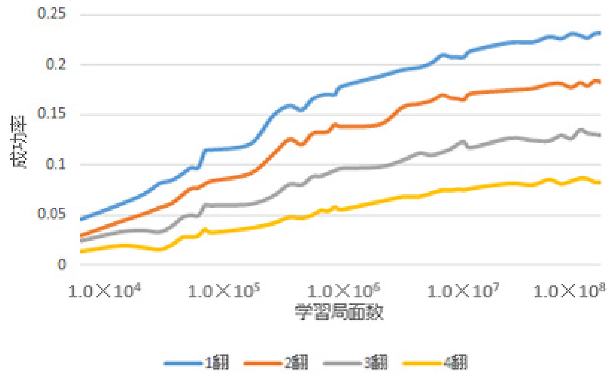


図 4　MCP における局面数と各翻数の成功率
Fig. 4　Success rate of MCP.

る場合，評価の相手もツモ切りである．同様に牌譜生成時の相手が 1 人麻雀政策である場合，評価の相手も 1 人麻雀政策である．テスト時，BCP も MCP も打牌の選択は，ある牌を切った手牌の特定の翻数で和了できる確率を求め，最も大きな確率を与えた牌を選択する．BCP は式 (2) をそのまま用い，MCP では特定の翻数以上で和了できる確率の合計を用いる．

各評価実験ごとに配牌やツモによって結果が変わらないようにするため同じ山を使用する．牌譜生成時と同じ条件にするため，和了可能な場合は特定の翻数を満たしているかどうかにかかわらず，すべて和了する．各翻数ごとに一万局を評価する．

提案手法が有効に行われているかを調べるため学習局面の数とその時点での成功率を調べる．結果を**図 3**，**図 4**，**図 5** に示す．学習局面の数は対数軸である．牌譜生成時の相手がツモ切りである場合，翻数予測モデル名に “vs ツモ切り” を付け，相手が 1 人麻雀政策の場合は何も付けない．どの学習方法においても，基本的には局面を増やすほど成功率が高くなっているため学習が上手く行われているといえる．

1 億局面を学習したときの評価の結果を**表 5** に示す．ベースラインとして予測モデルの代わりに 1 人麻雀政策を使用した結果も示す．MCPvs ツモ切りの場合，ベースラインと比較してすべての翻数において成功率が向上している．1 人麻雀政策の学習は実際の牌譜であるため相手の行動はツモ切りではない．そのためその牌譜を基に学習した 1 人麻雀政策よりも MCPvs ツモ切りは相手の行動に応じて適切に打牌を選択しているといえる．

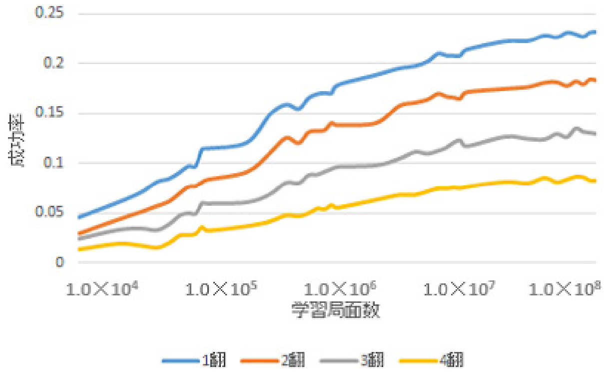


図 5　BCP における局面数と各翻数の成功率
Fig. 5　Success rate of BCP.

表 5　成功率
Table 5　Success rate of each model.

モデル	相手	1 翻	2 翻	3 翻	4 翻
1 人麻雀政策	ツモ切り	0.6411	0.5235	0.3169	0.1708
MCPvs ツモ切り	ツモ切り	0.6574	0.545	0.3663	0.2274
1 人麻雀政策	1 人麻雀政策	0.2446	0.1979	0.1278	0.0734
MCP	1 人麻雀政策	0.2318	0.1833	0.1296	0.0825
BCP	1 人麻雀政策	0.2382	0.1863	0.1267	0.0793

MCP や BCP の場合，低い翻数のときには 1 人麻雀政策に負けているものの，4 翻時には成功率が向上している．その原因は 1 人麻雀政策の学習は誰もリーチを宣言していない状態で和了したプレイヤまたは最初にリーチを宣言したプレイヤの牌譜を用いるため早くて安い点数を狙う行動が多いためであろう．現状の 1 人麻雀政策の具体的な問題点はゲーム終盤の逆転に必要な点数を作る技術がなかったことである．そのため高い点数を和了する技術が向上したことは一定の成功を収めたといえる．

5.4　最終的な順位を考慮した和了を行う麻雀プログラム

得られた翻数予測モデルを使用して最終的な順位を考慮した和了を行う麻雀プログラムを構築する．これを実現するための評価値は各翻数を和了できる確率に和了したときの期待最終順位（**Expected Final Rank, EFR**）[2] の総和に基づく．和了した際の点数の評価は重要ではあるものの，麻雀においては和了率は 2 割程度であり，8 割近くの場合和了できない．和了できなかった場合の行動をどのように評価するべきかという問題は相手の手牌や戦略に依存するため，自明な解決方法は存在しない．

この問題を解決する単純な方法は，点数移動が行われなかったと仮定して，そのときの期待最終順位を評価とする

方法が考えられる．しかし自分が和了できないときには他のプレイヤが和了しているため，期待最終順位は悪化するケースが多い．さらにこの方法では相手がツモ和了や自分が相手に放銃する可能性を無視しており，現実の麻雀を反映しているとはいえない．

そこで本研究では点数移動が行われなかったとするのではなく，現実的に起こりうるすべての点数移動を考慮し，そのときの期待最終順位を和了できないときのペナルティとして扱う．具体的には，麻雀の相手のツモ和了や放銃など，1 局で起こりうるすべての局終了時の点数状況を考慮し，そのときの期待最終順位とその状況が起きる確率との積和をペナルティとする．このペナルティの計算には和了できないときの各点数状況の確率を推定する必要がある．本研究では 2 種類の牌譜からこれらの確率を求めた．1 つ目は人間の牌譜と 2 つ目は翻数予測モデルの構築に使用した牌譜である．

このペナルティを考慮した和了を行うためのスコアの計算式は以下のようになる．

$$\begin{aligned} Score(\mathbf{x}) = & \sum_{i=1}^{4} \sum_{j \in J} P(i, j|\mathbf{x}) \times EFR(\mathbf{y}, i, j) \\ & + (1 - P(1|\mathbf{x})) \times Penalty(z) \end{aligned}$$

ここで J はロン和了とツモ和了の集合，i は和了した翻数である．$\mathbf{y}$ は i 翻で j の種類で和了したときの全プレイヤの点数とする．$P(i, j|\mathbf{x})$ は手牌 $\mathbf{x}$ が与えられたときの i 翻で j の種類で和了する確率であり，翻数予測モデルを基に算出される．各プレイヤに対するロン和了とツモ和了が起きる確率は簡単のため同じとする．すなわちロン和了が起こる確率の合計はツモ和了の 3 倍とした．関数 EFR はこの引数のときの点数状況における期待最終順位を返す．符は頻出頻度の高い 30 符とする．第 2 項は和了できない確率に前述したペナルティ項を示している．z は現在の点数状況であるため，$Penalty(z)$ は固定値でなく局ごとに異なる値を持つ．

MCP では $P(i|\mathbf{x}) = P_{mc}(i|\mathbf{x})$ とし，BCP は，$P(i|\mathbf{x})$ を以下のように置き換える．

$$P(i|\mathbf{x}) = \begin{cases} 1 - P_{bc}(i|\mathbf{x}) & \text{if } i = 0 \\ P_{bc}(i|\mathbf{x}) - P_{bc}(i+1|\mathbf{x}) & \text{otherwise} \end{cases}$$

ただし $P(5|\mathbf{x}) = 0$ とする．

ここからは以前の筆者らの研究 [2] に用いた麻雀プログラムとこの翻数予測モデルを組み合わせた提案手法の麻雀プログラムについて説明する．以前の筆者らの研究のプレイヤは序盤と中終盤において手を決定するアルゴリズムが異なる．序盤は 1 局開始時から以下の条件のいずれかを満たすまでと定義する．

- プレイヤがリーチ可能な手牌であるとき
- 相手プレイヤがリーチを宣言しているとき
- 1 人麻雀政策に従ったときの放銃率が 2%を超えるとき
- 1 人麻雀政策に従ったときの期待最終順位が 0.01 悪化するとき
- ツモが可能な牌の数が 16 枚以下

現状において問題となるのは序盤による手作りであるため，序盤のアルゴリズムのみを今回の手法に置き換える．

5.5　1 人麻雀の探索

前節で述べたように，翻数予測モデルと期待最終順位を組み合わせることで将来の報酬に基づいた手の選択が可能である．ここでは現在の手牌の報酬をより精度高く評価するため疑似的な麻雀のゲーム木を探索する．

本来の麻雀のゲーム木を探索するには自分プレイヤのほかに 3 人の相手プレイヤの手牌や行動を考慮する必要がある．このような設定ではゲーム木のノード数が膨大になり，現実的な時間内での有効な探索は行えない．そこで疑似的な麻雀のゲーム木では自分プレイヤのツモと打牌のみ（鳴きは行わない）を考慮する．ゲーム木のノードには 2 種類のノードが存在する．本研究では 1 つ目をツモ局面の**ツモノード**と 2 つ目を打牌後の**打牌ノード**と呼ぶ．2 つのノードの評価値は以下の式で表す．

$$V_{tsumo}(\mathbf{X}, depth) = \begin{cases} EFR(\mathbf{X}) & \text{if } \mathbf{X} = win \\ max[V_{move}(\mathbf{X} - Tile, depth), Tile \in \mathbf{X}] & \text{otherwise} \end{cases}$$

$$V_{move}(\mathbf{X}, depth) = \begin{cases} Score(\mathbf{X}) & \text{if } depth = Max\ depth \\ \sum_{Tile \in Tiles} P(Tile) \times V_{tsumo}(\mathbf{X} + Tile, depth + 1) & \text{otherwise} \end{cases}$$

ここで V_{tsumo}，V_{move} はツモノードと打牌ノードの評価値，$\mathbf{X}$ は手牌，$Tiles$ は 34 種類の牌，$P(Tile)$ は $Tile$ をツモる確率である．$P(Tile)$ は見えていない牌を数え上げることで求める．$depth$，$Max\ depth$ は探索を制御するパラメータであり，それぞれ現在の深さ，打ち切りまでの深さを表す．ツモノードであるルートノードの $depth = 0$ とする．

ツモノードの評価値は和了した場合そのときの期待最終順位とし，それ以外は手牌の中で最も高い打牌ノードの評価値となる．打牌ノードの評価値は終端ノードであれば前章の最終的な順位を考慮した和了を行う式により算出され，それ以外はある牌をツモったツモノードの評価値とそのツモ確率の積和とする．

上記のゲーム木探索を行うと現実的な計算時間では $Max\ depth = 2$ が限界であった．そこでより深い探索を可

能にするため枝刈りを行う．枝刈りは手牌を切る局面とツモ番の 2 つの場面においてそれぞれ異なる基準を設ける．手牌を切る局面ではすべての手牌について評価するのではなく，翻数予測モデルの上位 3 つまでを探索する．ツモ番では向聴数を減らす牌と孤立牌の関連牌のみを探索する．関連牌とは数字の前後 2 つの牌のことを指す．これ以外の牌はツモ切りが起こるとして，その場合のノードの評価値を 0 とした．そして得られた積和の値を正規化するために評価値をツモ切りが起きなかった確率で割りツモノードの評価値とした．これらの枝刈りにより現実的な計算時間に深さ 3 まで評価することが可能になった．

6. 対戦実験と結果

本章では提案手法によって得られた麻雀プログラムの強さを対戦によって評価する．以前の麻雀プログラムとの相違点は序盤のアルゴリズムである．以前の麻雀プログラムは 1 人麻雀政策に対して，本研究での序盤は翻数予測モデル（MCP や BCP）を用いる．

6.1 自己対戦における設定

自己対戦では翻数予測モデルを用いた麻雀プログラム 1 体と文献 [2] の麻雀プログラム 3 体で対局を行う．1 ゲームは東風戦で行われる．シミュレーションにかける時間は中終盤では 1 手 1.5 秒とする．また序盤の 1 手ごとの時間は以前の麻雀プログラムは 1 秒未満である．また本研究の序盤の麻雀プログラムの時間も 2 秒で計算できるであろう探索の深さとした．

6.2 自己対戦における結果

麻雀プログラムについて説明する．Penalty（自己対戦）とは自己対戦により生成された牌譜を使ってペナルティを計算した麻雀プログラムである．また Penalty（人間）とは天鳳の牌譜から特定のプレイヤが和了できなかったときの条件におけるペナルティを計算したものである．探索を行う際の深さは 3 であり，その他の麻雀プログラムは深さが 1 である．

結果を**表 6** に示す．BCP と比較して人間の牌譜を用いたペナルティや探索を加えることで実力が向上した．しかしながらいずれの麻雀プログラムも以前の麻雀プログラムと比較して大きく負け越している．自己対戦におけるペナルティが機能しない理由は現実の麻雀と比べ特定の事象の確率が乖離しているためである．すなわち局面生成時に相手がツモ切りや 1 人麻雀政策しか選択しないのは相手プレイヤの戦術としては単純すぎることが原因であろう．

表 6　順位分布

Table 6　Rank distribution.

	1 位率	2 位率	3 位率	4 位率	平均順位	試合数
MCPvs ツモ切り	0.185	0.251	0.282	0.280	2.65±0.01	30,505
MCP	0.180	0.248	0.292	0.276	2.67±0.01	62,742
BCP	0.194	0.253	0.283	0.270	2.62±0.01	44,550
BCP+Penalty (自己対戦)	0.121	0.179	0.324	0.375	2.95±0.05	1,163
BCP+Penalty (人間)	0.227	0.244	0.262	0.264	2.56±0.04	3,589
BCP+Penalty (人間)+探索（depth=3）	0.240	0.244	0.250	0.265	2.54±0.02	14,804

和了・放銃は**表 7** に示す．これらは 1 局のプレイヤの強さを測定するためによく用いられている．相手をツモ切りから 1 人麻雀政策と強くすることで得られる麻雀プログラムの和了率も向上している．しかしながら対戦相手の和了率と比較して大きな差をつけられている．このことから翻数予測モデルを組み込むことで単純な牌効率が悪化しておりその結果，平均順位も大きく悪化したのではないかと考えられる．

6.3 考察

牌効率が悪化していると分かる典型的な例は**図 6** の手牌である．この手牌の BCP による評価値を**表 8** にまとめた．この手牌から 3 翻以上の高い手を目指す場合は，人間なら索子の混一色にすることを考える．BCP においても同様に考え，萬子や筒子を切る手を高く，字牌を切る手は価値が低いと評価している．反対に 1 翻などの安い手を和了しようとすると，人間では 78p を残し白を刻子にすることを考える[*2]．その点は 7p や白を切ったときの評価が低いことから BCP も理解している．切るべき牌を考えたときに北は 5m に比べ両面になることがないためこの手牌では 1 翻を和了するためには一番価値が低いと人間は考える．しかしながら BCP は北よりも 5m を切る方が価値が

表 7　和了・放銃率

Table 7　Rate of winning and discarding a winning tiles for opponent.

	和了率	放銃率	相手の平均和了率	相手の平均放銃率
MCPvs ツモ切り	0.193	0.118	0.217	0.119
MCP	0.194	0.114	0.215	0.120
BCP	0.201	0.115	0.214	0.119
BCP+Penalty（自己対戦）	0.093	0.114	0.235	0.110
BCP+Penalty（人間）	0.202	0.123	0.226	0.123
BCP+Penalty（人間）+探索（depth=3）	0.195	0.128	0.220	0.123

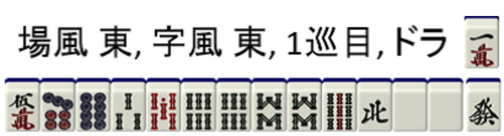


図 6　問題のある手牌

Fig. 6　A bed move example.

表 8　BCP による図 6 の評価値

Table 8　Evaluation value of Fig. 6.

牌	1 翻以上	2 翻以上	3 翻以上	4 翻以上
5m	0.353	0.338	0.221	0.111
7p	0.294	0.303	0.240	0.136
北	0.328	0.285	0.161	0.079
白	0.173	0.168	0.149	0.063

*2　p はピンズ，m はマンズである

表 9 実際の対局による図 6 の評価値
Table 9 Evaluation value of Fig. 6 in real games.

牌	1 翻以上	2 翻以上	3 翻以上	4 翻以上
5m	0.401	0.373	0.192	0.117
7p	0.325	0.323	0.298	0.203
北	0.405	0.388	0.241	0.107
白	0.203	0.201	0.164	0.080

図 7 鳴き局面における手牌
Fig. 7 A bed move example.

表 10 オーラス最下位時の図 6 の BCP と 1 人麻雀政策の評価値
Table 10 Evaluation value of BCP and one player mahjong moves.

牌	BCP の評価値	1 人麻雀政策の評価値
5m	3.889	−12,447
7p	3.797	−12,587
北	3.93	−12,326
白	3.94	−12,595

高いと考えている．このケースではターツ候補が揃っており 5m を切ってしまったために和了できないケースが少なくそこが上手く学習できないことが原因であろう．

翻数予測モデルの予測と実際の局面との比較を**表 9** にまとめた．図 6 の手牌から全員が 1 人麻雀政策に従ったときの特定の翻数以上を実際に和了できた確率を求めた．確率を近似的に求めるため各候補手に対して 1 万回試行を行った．各施行ごとに相手の手牌や山をランダムに変更した．結果である表 9 と表 8 を比較すると BCP の値は小さい．局面生成時にランダムな手を混ぜたことにより，和了しにくい手牌として学習されたことが考えられる．

実際の対局の点数状況を考慮した打牌の評価を表 9 にまとめた．本研究の 1 つの目的は逆転のために高い点数を和了することであった．そこでオーラスで満貫以上を和了できなければ最下位という状況で逆転する手を作れるかを調べた．オーラス最下位時において図 6 の手牌では 1 人麻雀政策では評価値の最も高い北を選択するが，提案手法では期待最終順位の最も低い 7p を選択する．表 9 を参照すると，7p が 4 翻以上を和了する確率が高いため，高い点数を和了するという目的としては提案手法が効果的に機能している．

もう 1 つの例は**図 7** の手牌である．この例は翻数予測モデルが役を認識できていない例である．状況はこの手牌において北が切られ，鳴くかどうかを考慮する局面である．

表 11 1 翻以上の評価値
Table 11 Evaluation value of Fig. 7.

行動	1 翻以上
パス	0.821
ポンかつ 9s	0.161

1 巡目で向聴数が 1 であるため手牌自体はよく，パスをすることで和了できる確率が高いのも納得できる．ポンをすることに関して北は役牌ではないので，鳴いても役がないことは明らかである．しかしながら翻数予測モデルでは 1 翻以上で和了できる確率が 0 になっていない．このように牌効率が悪くなった要因は役がない鳴きが頻出していることも考えられる．

7. おわりに

本研究では自動対戦棋譜の教師あり学習を用いて，現在の手牌から和了できる翻数を予測するモデルを構築し，それをもとに最終的な順位を考慮した和了を行う麻雀プログラムを構築した．牌譜生成時の相手がツモ切りの場合の学習ではテスト時の結果はどの翻数においても 1 人麻雀政策を使用するより成功率が向上しており，本研究の可能性を示すことができた．しかしながら翻数予測モデルを使用した麻雀プログラムはもとの麻雀プログラムに勝ち越すことはできなかった．

自己対戦の結果に関していえば，牌譜生成時の相手を 1 人麻雀政策に従う相手プレイヤにすることで，少しは改善した．このことから牌譜生成には相手プレイヤの実力が強化学習によって得られる実力に大きく関わっていることが分かる．そのため対戦相手を降りを行うといったより強いプレイヤにすることによって局面の質が向上し，最終的な実力も向上する可能性がある．

考察で述べたように，翻数予測モデルは鳴いたときに役があるかどうかの理解ができていない．原因は適切な局面の生成ができていないことがあげられる．局面生成時において 1 人麻雀政策に従い手を進めていた場合，役にならない手を選択されることは少ない．役にならない手牌の局面を作ることができるのはランダムに鳴いたときのみである．そのためあまり役にならない手牌の局面が生成されず，学習が有効に働かない可能性がある．解決策としては鳴ける局面を優先的にサンプリングする方法が考えられる．

もう 1 つの原因は手牌の表現力が足りないことである．和了するには向聴数を下げることが必要である．しかしながら向聴数が小さいことは役があることを保証するわけではないため，これを混同して学習が行われている可能性がある．これを解決策するためには，現在の膨大に組み合わせた特徴量を用いるのではなく，組み合わせる前の特徴量を用いて，モデルを深層学習にすることで必要な特徴量の選別を行うことが必要である．

参考文献

[1] Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al.: Mastering the game of Go with deep neural networks and tree search, *Nature*, Vol.529, No.7587, pp.484–489 (2016).

[2] 水上直紀，鶴岡慶雅：期待最終順位の推定に基づくコンピュータ麻雀プレイヤの構築，*Proc. 20th Game Programming Workshop*, pp.179–186 (2015).

[3] 栗田　萌，保木邦仁：有向非巡回グラフで表現された 1 人麻雀の探索アルゴリズム，*Proc. 22nd Game Programming Workshop*, pp.42–49 (2017).

[4] Heinrich, J., Lanctot, M. and Silver, D.: Fictitious Self-Play in Extensive-Form Games, *Proc. 32nd International Conference on Machine Learning* (*ICML-15*), pp.805–813 (2015).

[5] Heinrich, J. and Silver, D.: Smooth UCT search in computer poker, *Proc. 24th International Joint Conference on Artifical Intelligence*, pp.554–560 (2015).

[6] 水上直紀，中張遼太郎，浦　晃，三輪　誠，鶴岡慶雅，近山隆：多人数性を分割した教師付き学習による四人麻雀プログラムの実現，情報処理学会論文誌，Vol.55, No.11, pp.2410–2420 (2014).

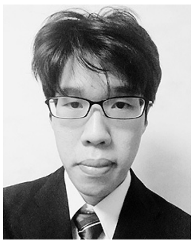
[7] 角田真吾：天鳳 (2014), 入手先 ⟨http://tenhou.net/⟩.

[8] Collins, M.: Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms, *Proc. ACL-02 Conference on Empirical Methods in Natural Language Processing – Volume 10*, Association for Computational Linguistics, pp.1–8 (2002).

[9] 三木理斗，近山　隆：多人数不完全情報ゲームにおける最適行動決定に関する研究，修士論文，東京大学 (2010).

[10] Duchi, J. and Singer, Y.: Efficient online and batch learning using forward backward splitting, *The Journal of Machine Learning Research*, Vol.10, pp.2899–2934 (2009).

[11] Duchi, J., Hazan, E. and Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization, *The Journal of Machine Learning Research*, Vol.12, pp.2121–2159 (2011).



水上 直紀　（正会員）

2015 年東京大学大学院工学系研究科修士修了．2019 年同大学大学院工学系研究科博士課程修了．博士（工学）．2018 年 HEROZ 株式会社入社．ゲーム AI に関する研究に従事．

鶴岡 慶雅　（正会員）

2002 年東京大学大学院工学系研究科電子工学専攻博士課程修了．博士（工学）．科学技術振興事業団研究員，マンチェスター大学 Research Associate，北陸先端科学技術大学院大学准教授，東京大学大学院工学系研究科准教授を経て，2018 年より東京大学大学院情報理工学系研究科教授，現在に至る．自然言語処理およびゲーム AI の研究に従事．