強化学習を用いた効率的な和了を行う麻雀プレイヤ

水上 直紀[1,a)]　鶴岡 慶雅[1,b)]

概要：本論文では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．

Computer Mahjong Players with Effective Winning Strategies Based on Reinforcement Learning

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

Abstract: This paper describes a method for building a mahjong program using reinforcement learning. We train models that predict winning scores of a player's hands using game records that are generated by our program. Our program decides moves based on the outputs of the prediction models and the expected final ranks. The program has obtained a skill for winning with high scores, but the evaluation results of self-play is lower than those of the original program.

1. はじめに

強化学習とはエージェントがある環境において経験から累積される報酬を最大にする行動を自動的に学習する機械学習の一種である．この形式は汎用的であるため制御工学やマルチエージェントの分野でも用いられる．近年では，強化学習を用いることで強いゲーム AI が提案されている [1], [2], [3].

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．全ての局が終了した時点で最終的に最も多くの点を持っているプレイヤが麻雀の勝者である．麻雀は不完全情報や多人数という将棋などのゲームにはない難しい性質を持っているが，同じ手牌でも点数状況によって最善手が異なるという繰り返しゲームの性質もまた麻雀の中で難しい性質の一つである．

繰り返しゲームとして麻雀を対象にした研究として，水上らは期待最終順位に基づいたプレイヤを構築した [4]．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．まずはこの期待最終順位を出力するモデルを牌譜から教師あり学習を行った．麻雀プレイヤは中終盤においてはシミュレーションを行いその報酬によって手を決定している．そこでシミュレーションの報酬を和了した時の点数から期待最終順位に変更することで点数状況を考慮したプレイヤの構築に成功した．その結果，局単位の収支は減少したが，トップ死守率などの順位にかかわる指標が向上し，それに伴い人間との対戦成績も向上した．

上記の方法で点数状況に関しての問題が完全に解決されたわけではなく，このプレイヤが明らかに悪い手を選択する状況も存在する．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．一般に安い手は和了しやすく，高い手は和了しにくい．そのためプレイヤは和了できる確率に対して和了した時の点数状況をよくなる効率の良い和了を目指している．例えば最終局において 4 位の時には相手の点数を逆転する手を作らなければ意味がない．反対に 1 位の時は高得点の和了は必要なく，和了できれば低い点数で構わない．現在の麻雀プレイヤは終盤の自分の和了した時の点数がほぼ確定した状態に対しては攻

[1] 東京大学工学系研究科
Graduate School of Engineering, The University of Tokyo
[a)] mizukami@logos.t.u-tokyo.ac.jp
[b)] tsuruoka@logos.t.u-tokyo.ac.jp

表 1　役一覧

翻数	役名
一	門前清自摸和, リーチ, 槍槓, 嶺上開花, 海底, 役牌, 断幺九, 一盃口, 平和, 一発, ドラ, 赤ドラ, 裏ドラ
二	混全帯幺九, 一気通貫, 三色同順, ダブルリーチ, 三色同刻, 三槓子, 対々和, 三暗刻, 小三元, 混老頭, 七対子
三	純全帯幺九, 混一色, 二盃口
六	清一色
役満 (十三)	天和, 地和, 大三元, 四暗刻, 字一色, 緑一色, 清老頭, 国士無双, 四喜和, 四槓子, 九蓮宝燈

めるべきか降りるべきかということは理解したものの，反対に序盤で何点の手を作ればいいかに関しては全く理解していない．これは麻雀プレイヤの序盤で用いるモデルはこの手が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，強化学習を用いて，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プレイヤ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，3 章で関連研究を述べる．次に 4 章で本研究のベースラインとなる一人麻雀の手について述べる．提案手法として，5 章で強化学習の方法，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. 麻雀のルールと用語

この章では麻雀の得点に関するルールについて解説する．麻雀は自分の牌を組み合わせて役（特定の構成）を作り，**和了（ホーラ）**し，役に応じた点数を得るゲームである．点数は**翻**と**符**によって決まる．どちらも自分の牌の組み合わせによって決まるが，点数は翻数に大きく作用されるため，プレイヤは符についてはあまり考慮せず，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，難易度の高い役ほど，基本的には翻数が高い．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．各色の 5 の一枚が赤ドラとなっている．

麻雀は四人のうち誰かが役を構成すると 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**）を終了した時の得点の多さに応じて順位が決まる．

和了に関する用語としては，ツモして和了することをツモ和了，あるいは略してツモと呼び，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全ての和了点数を支払う．

3. 関連研究

コンピュータ麻雀プレイヤに関して，水上ら [4] は期待最終順位に基づいたプレイヤを構築した．期待最終順位とは現在の点数状況から最終的な各順位をとる確率から得られる値である．この確率は牌譜から学習を行った最終結果を予測するモデルが出力する．これをシミュレーションの報酬とすることで，点数状況に応じた押引きが可能となった．天鳳 [5] における対人戦の結果は局収支を最大化するプレイヤに比べ局収支は悪化したものの平均順位は改善した．

小松ら [6] や海津ら [7] はモンテカルロ法を用いて　鳴きは行えない一人麻雀の手作りを行うプレイヤを構築した．この手法はランダムに手を選択するのではなく，和了できる組み合わせになるまでランダムに牌を足し続け，各牌の和了点の寄与を調べることで，和了に必要な牌を求めた．海津らは早和了に必要なパラメータを導入し，早和了と高得点のバランスをとった．しかしながらどの局面において早和了または高得点を目指すべきかということについては触れられていない．

不完全情報ゲームのポーカーの一種であるテキサスホールデムでは，Heinrich ら [3] は強化学習を用いることで **CounterFactual Regret minimization, (CFR)**[8] で構築されたプレイヤに迫る実力を得たと報告している．手法としては自己対戦を行い，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もつモデルをより賢くかつ勝率に変換する方法として強化学習が用いられた [1]．手法は畳み込みネットワークで構成される方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ー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．まず学習に用いる局面のステップ数をランダムに決める．ステップ数のひとつ前まではロールアウトポリシーによって手を決める．次に完全にランダムな手を決め，その手を着手した局面を学習に用いる．最終結果はその局面から方策ネットワークを用いてゲームを進めることで得る．AlphaGo は 3000 万の局面を生成し学習を行うことで局面の勝率を予測する評価値ネットワークを構築し，世界チャンピオンに勝利するまでになった．

一人ゲームとしての研究では **Deep Q Nework (DQN)**[2] と呼ばれる手法が Atari というビデオゲーム

で人間よりはるかに高いまたは人間に近いスコアをとるようになった．DQN は強化学習の一種である Q 学習における Q 関数を畳み込みネットワークで表現した．その入力には画面のピクセル画像を使用するという事前知識をほとんど用いない設定である．それにもかかわらず多様なゲームにおいても高いスコアを上げることに成功している．特にピンボールのようにボールとバーの位置関係がわかればスコアが得られる単純なゲームでは人間の 25 倍以上のスコアを達成した．

麻雀ではポーカーと異なり手を効率的に抽象化する方法は提案されておらず，抽象化したゲームとして解くことは現在のところ不可能である．また，相手を考慮しない一人麻雀としても，情報集合数は膨大になり，ゲーム木の全探索は不可能といえる．そこで本研究では強化学習を用いて，和了翻数予測モデルを構築し．そしてその予測モデルを用いて効率の良い和了を考慮したプレイヤを構築する．

4. 一人麻雀の手

この章では本研究で重要な役割を果たす一人麻雀の手について述べる．一人麻雀の手とは以前の水上らの研究 [9] で提案された教師あり学習によって得られたモデルの出力結果である．与えられた手牌から一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ものを出力結果とする．教師データとなる牌譜は最初に和了したプレイヤまたはリーチを宣言したプレイヤの牌譜を用いるため，出力結果である一人麻雀の手は和了に向かうための手である．一人麻雀の手という名前ではあるが，ツモ番だけでなく鳴ける局面でも使用することができる．教師あり学習は平均化パーセプトロン [10] を用いた．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．

以前の研究と異なる点は特徴量の改善である．以前の特徴量を使用しない理由はこの特徴量では役が完成しなくなる鳴きが多くなるからである．そのため役を作るための強化学習に利用するのは不適切と考え本研究では特徴量の改良を行った．詳細は**表 2** と**表 3** に示す．多くの特徴量はいくつかの要素の組み合わせで構成されている．組み合わせの全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．特徴量は合計で 6,661,309 である．

結果を**表 4** に示す．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 [9] 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．本研究ではこの一人麻雀の手を用いて実験を行う．

一人麻雀の手は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．しかしながら基準となる評価値は牌の選択されやすさであり，この牌を切った時に何翻で和了できるかということとは全く関係がない．そのため一人麻雀の手は平均的な局面においては悪手を指すことは少ないが，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を指す．次の章はこの原因を解消する手法について述べる．

表 4 一致率

局面の種類	牌譜の数	完全一致数	鳴きのみ正解数	正解率
ツモ番	1,140,576	859,088	N\A	75.3
鳴く	68,397	46.945	11.731	85.8
鳴かない	252,666	240,450	N\A	95.1

5. 強化学習による役作り

前章で述べたように一人麻雀の手はこの手牌が何翻で和了できるかを理解していない．この章では強化学習を用いてこの問題を解決する．

5.1 強化学習の方法

本研究では強化学習を用いて一人麻雀の手の出力を各翻数の和了する確率に変換することを試みる．強化学習をするにあたってほかのゲームの特徴と麻雀の特徴を考慮する．バックギャモンで用いられた TD 法 [11], [12] は各コマを適当に動かしてもゴールできる，すなわち報酬が得られるためうまくいったと考えられる．しかしながら麻雀ではランダムな手を指し続けても和了することは難しいと三木ら [13] は報告している．すなわち一人麻雀の手の特徴量だけを用いて，その重みを 0 などに初期化して強化学習を行う方法は麻雀ではうまくいかない．4 章で述べたように一人麻雀の手はある程度は強いため，これをベースにすることで学習が効率的に行えると考えた．

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．1 手指すごとにパラメータを変更するのではなく，牌譜を生成しそこから教師あり学習を行う．AlphaGo は局面の勝率を出力する評価値ネットワークを構成するため，元から十分に強い方策ネットワークをベースに自己対戦を行った．

牌譜の生成法について述べる．基本的には通常の麻雀と同じで自分のツモと相手の捨て牌を利用して和了を行う．考慮すべきはプレイヤの手番での行動，すなわち自分の手番と相手の手番の行動についてである．

まず自分の手番の挙動について述べる．上記で述べたようにランダムな手を指し続けても和了することは難しく，学習が上手くいかないと考えられる．そのため自分の手番では基本的に一人麻雀の手を選択する．強化学習では局面を正確に判断するため，様々な局面に訪問することが求められる．麻雀では配牌とツモがあるため，将棋などのゲームに比べれば，何の工夫をしなくても様々な局面に訪問す

表 2　一人麻雀プレイヤの特徴量

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	$15\times5=75$
リーチが可能か	2
向聴数，副露数，min(受け入れ枚数,20)	$15\times5\times21=1,575$
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	$6\times6=36$
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	$2\times6\times6\times16\times5\times5=28,800$
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	$(14+1)\times5\times2=150$
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	$4\times3\times18=216$
min(向聴数, 3)，副露数，振聴か，min(役のある待ち牌の数,7)	$4\times5\times2\times8=320$
両面を優先した時の両面+面子の数，向聴数	$7\times15=105$
面子+ターツ+ターツ候補，向聴数	$11\times15=165$
両面を優先した時の両面+面子の数，面子+ターツ+ターツ候補，向聴数	$7\times11\times15=1155$
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)	$5\times8\times8\times5\times4=6,400$
min(全帯幺九の向聴数, 4)，max(全帯幺九の枚数-6)，2378 の暗刻があるか，副露数， 両面を優先した時の面子の数，両面を優先した時の両面の数，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 2378 字牌の頭の数	$5\times2\times5\times5\times4\times8\times2\times2=32,000$
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	$7\times4\times4\times8\times4=6,400$
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)	$5\times8\times8\times5\times4=6,400$
両面を優先した時の両面+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	$8\times8\times4\times4\times2\times2\times4\times4\times3\times2\times7$ $=2,752,512$
面子を優先した時の両面+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	$8\times8\times4\times4\times2\times2\times4\times4\times2\times7$ $=917,504$
min(色の中で最も多い色の数の向聴数,4)，両面を優先した時の面子の数， 両面対子+愚形対子の数，副露数	$5\times5\times8\times8\times5=8,000$
min(19 字牌抜いた時の向聴数,4)，両面を優先した時の面子の数， 両面を優先した時の面子+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)	$5\times5\times8\times4\times5\times6\times2\times3\times3=432,000$
両面を優先した時の両面+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 (19,28,34567, 字牌)，その浮き牌の枚数	$8\times8\times8\times4\times3\times2\times4\times4$ $=196,608$
両面を優先した時の両面+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	$8\times8\times4\times4\times4\times3\times2=24,576$
min(向聴数, 4)，七対子の向聴数，向聴数の悪化しない頭の数， 両面を優先した時の両面+面子の数，リーチが可能か，完全一, 二向聴, そうでない，	$5\times8\times8\times8\times2\times3=15,360$
min(向聴数, 3)，役牌の刻子があるか，役牌の対子があるか，両面を優先した時の面子の数， 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	$4\times2\times2\times5\times8\times4\times4\times4\times3$ $=122,880$
両面を優先した時の両面+面子の数，愚形の数，浮き牌の最も外側の種類 (19,23,3456, 字牌)， 頭と頭の組み合わせ，頭の数，完全一, 二向聴, そうでない	$8\times8\times5\times16\times4\times3=61,440$
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，	$5\times8\times5\times5\times2\times3=6,000$

ることが可能である．しかしながら鳴いた局面などは配牌とツモによって生成されないため，意図的に作り出す必要がある．これを実現するためには 1 局の間に一度だけランダムな手を選択し，その直後の局面を教師あり学習に使用する．

ランダムな手というのは合法手の中から一人麻雀の手の評価値関係なく一つを選択する．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（例えば 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ある．そのため鳴ける局面における合法手は鳴く鳴かないの二つではなく，鳴いて何を切るかまでを組み合わせとして合法手とする．

相手プレイヤの手番の挙動について述べる．麻雀は点数状況によって狙うべき手が変わる．同様に相手も点数状況によって狙うべき手が変わり，切られる牌も異なる．強化学習ではそれらを考慮に入れた相手が対戦相手であることが理想ではあるが，現状そのようなコンピュータプレイヤがいないことを考慮し，本研究では相手プレイヤを 2 種類用意した．まずは特定の点数を目指すプレイヤの学習が行えるのかを確認するため，なるべく弱い相手を用いる．

表 3　一人麻雀プレイヤの特徴量

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)， 両面を優先した時の両面+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み合わせ（n=2〜6），リーチが可能	$(100 + 500 + 1860 + 8634 + 23760) \times 2 = 69,708$
各色の 1 から 9 の組み合わせ．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)， あリーチができるか，副露数	$2 \times 2 \times 4 \times 2 \times 5 = 160$
三色に最も近い枚数，向聴数，副露数	$10 \times 14 \times 5 = 700$
各三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)， 三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 $\geqq$5，両面+面子 $\geqq$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) までの組み合わせ	$20 + 231 + 1771 = 480$

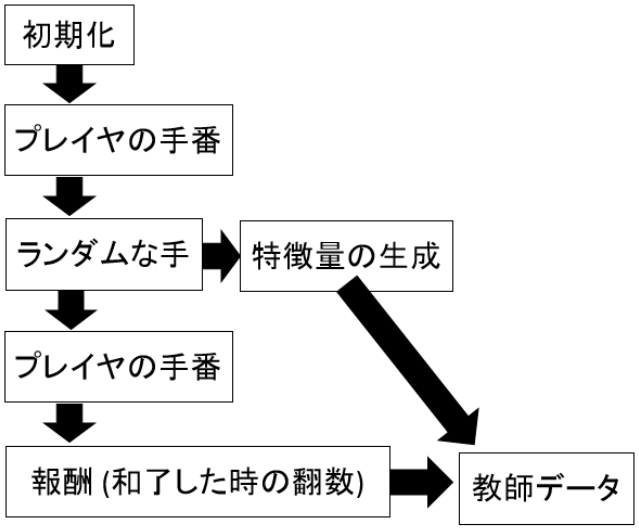

図 1　提案手法

すなわち一つ目のプレイヤはツモ切りを続けるプレイヤである．次に相手プレイヤを現実の麻雀に近づけることを考える．現実の麻雀では他のプレイヤが和了するため，一人麻雀のように 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．他のプレイヤが和了を実現するため二つ目のプレイヤは一人麻雀の手を選択し続けるプレイヤである．現実の麻雀ではプレイヤはとりあえず和了を目指すための手を選択するため，相手プレイヤとして一人麻雀の手を選択させ続けることは他のプレイヤが和了を実現するには妥当と考えた．

図 1 は提案手法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設定した．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と考えた．これを考慮して場風は東と南は 4/9，西は 1/9 の確率で選択される．教師データとして使用する巡目は 1 から 1 局が終了する 1 手前の中から一様に選ばれる．当然この局が何巡で終了するかはわからない．これを調べるためにあらかじめ自分の手番では一人麻雀の手を選択した場合を行い，この局が何巡あるかを求めておく．この値を利用して教師データとして使用する巡目を決定し，同じ初期化した局面からスタートさせて教師データを生成する．

一人麻雀の手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全て宣言するとした．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ので本研究では 4 翻以上は同じとする．また符に関しては，翻の影響が点数に大きいため本研究では無視する．また和了が可能な時はすべて和了する．天和や地和で和了した時の局面は自分が一手も指していないため使用しない．教師データは学習局面は 1 局に対して 1 局面までとして 1 億局面を用意した．

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から翻数予測モデルの学習について述べる．手牌から予想される翻数を学習するモデルの構築方法は二つある．一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．すなわち，予測する結果を 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 クラスの多クラス問題として捉えることができる．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（1）で表現される．

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

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

目的関数は式 (2) で表現される．

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

ここで N はトレーニングデータのサンプル数, $\mathbf{X}_i$ は i 番目のトレーニングデータ, $c_{i.h}$ はトレーニングデータの結果と各翻数に対応する 2 値（1 または 0）のデータである．λ はトレーニングデータに過学習することを防ぐ正則化項である．本研究では正則化項 λ を 0.01 とする．

二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．手牌から予想される翻数を 2 値の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．それぞれ 1 翻以上, 2 翻以上, 3 翻以上, 4 翻以上である．

$$P_{bc}(han \leq h) = \frac{1}{1+\exp\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}\right)} \qquad (3)$$

目的関数は式（4）で表される．

$$L(\mathbf{w}) = -\sum_{i=1}^{N} \mathbf{c_i} log(P_{bc}(\mathbf{X_i})) + (1-\mathbf{c_i}) log((1-P_{bc}(\mathbf{X_i}))) + \frac{\lambda|\mathbf{w}|^2}{N} \qquad (4)$$

これら二つの重みベクトルの学習は FOBOS [14] を用いて学習を行う．学習率は Adagrad [15] を用いて決定する．$\mathbf{x}$ は一人麻雀の手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（3）を用いたプレイヤを **Binary Class Player (BCP)** と呼ぶ．

5.3 強化学習の結果

モデルが実際に有効に活用できるかを調べるためにテストを行う．テストでは学習したモデルを使用して特定の翻数 (1, 2, 3, 4) で和了することが可能かを調べる．評価基準は成功率である．特定の翻数以上を和了しても成功とする．テスト時，プレイヤは BCP は式（3）をそのまま用いるが，MCP では以下の式を用いて手を決定する．

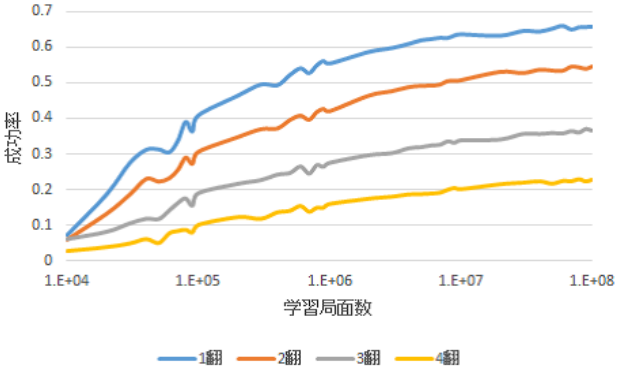

図 2　MCP+ツモ切りにおける局面数と各翻数の成功率

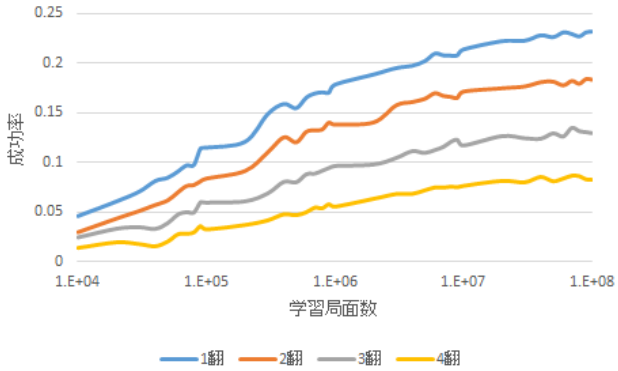

図 3　MCP における局面数と各翻数の成功率

$$Score(\mathbf{x}) = \sum_{i=h}^{4} P_{mc}(i|\mathbf{x}), \qquad (5)$$

これは特定の翻以上で和了できる確率の合計を足し合わせた値である．

各テスト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．ただし各翻ごとに使用するテスト用の山は異なる山を使用する．和了できるときは特定の翻数を満たしているかどうかに関わらず，すべて和了する．各翻数ごとに一万局をテストした．

強化学習が有効に行われているかを調べるため学習局面を少しづつ増やしながらその時の成功率を調べる．結果を**図 2**, **図 3**, **図 4** に示す．学習局面の数は対数軸である．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．またどの結果もおよそ収束していると読み取れる．

1 億局面を学習した時のテストの結果を**表 5** に示す．ベースラインとして予測モデルの代わりに一人麻雀の手を使用した結果も載せる．相手がツモ切りを行うプレイヤの時には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上がっている．相手が一人麻雀の時には，低い翻数の

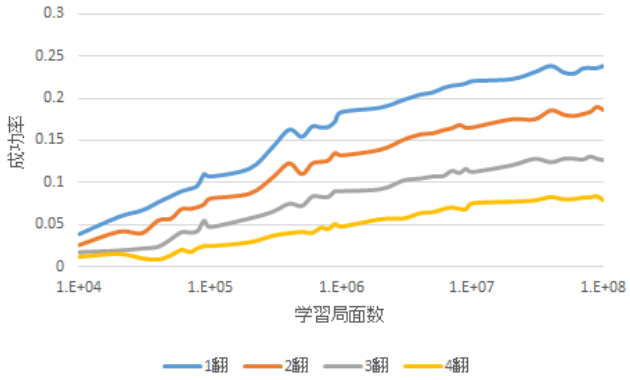

図 4　BCP における局面数と各翻数の成功率

表 5　成功率

モデル	相手	1 翻	2 翻	3 翻	4 翻
一人麻雀の手	ツモ切り	0.6411	0.5235	0.3169	0.1708
MCP	ツモ切り	0.6574	0.545	0.3663	0.2274
一人麻雀の手	一人麻雀の手	0.2446	0.1979	0.1278	0.0734
MCP	一人麻雀の手	0.2318	0.1833	0.1296	0.0825
BCP	一人麻雀の手	0.2382	0.1863	0.1267	0.0793

時には一人麻雀の手に負けているものの，4 翻時には成功率が向上している．このことから強化学習による翻数予測モデルは一定の成功を収めたといえる．

6.　対戦実験と結果

この章では提案手法によって得られた翻数予測モデルを組み込んだ麻雀プログラムと自己対戦を行う．

6.1　効率的な和了を行う麻雀プレイヤ

前章で得られた翻数予測モデルを使用して効率的な和了を行う麻雀プレイヤを構築する．効率的な和了とは最終的な順位を考慮した和了のことを指す．すなわち各翻数を和了できる確率に和了した時の期待最終順位（$Expected\ Final\ Rank$, EFR）[4] の総和に基づく．これを式にすると次の式になる．

$$Score(\mathbf{x}) = \sum_{p \in players} \sum_{i=0}^{4} \frac{P(i|\mathbf{x}) EFR(y_0 + t(i), y_p - t(i))}{4} \quad (6)$$

y_p は現在のプレイヤ p の点数，$t(i)$ は i 翻の手で和了した時の点数とする．符は頻出頻度の高い 30 符とする．和了するときはツモ和了と 3 人に対してのロン和了が均等に起こると仮定し，それぞれが起きる確率を和了できる確率の 4 で割った値とする．この計算式では和了できないときは現在の点数状況による期待最終順位を返す．

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．

$$P(i|\mathbf{x}) = \begin{cases} 1 - P_{bc}(i|\mathbf{x}) & if(i = 0) \\ P_{bc}(i|\mathbf{x}) - P_{bc}(i+1|\mathbf{x}) & otherwise \end{cases} \quad (7)$$

ここで $P(5|\mathbf{x}) = 0$ とする．

表 7　和了・放銃率

	和了率	放銃率
MCP+ツモ切り	0.193	0.118
MCP	0.194	0.114
BCP	0.201	0.115

実際に対局するときは以前の水上らの研究 [4] に用いたプレイヤにこの翻数予測モデルを組み込む．こ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．今回問題になってい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上記の手法に置き換える．

6.2　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プレイヤ 1 体と以前のプレイヤ 3 体で対局を行う．一ゲームは東風戦で行われる．中終盤のシミュレーションにかける時間は 1 手 1.5 秒とする．

6.3　自己対戦における結果

結果を**表 6** に示す．いずれのプレイヤもベースラインと比較して大きく負け越している．

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一人麻雀の手と強くすることで得られるプレイヤの和了率も向上している．しかしながら対戦相手は和了率が 0.21 ほどあり 0.01 ほど悪化している．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．

6.4　考察

牌効率が悪化しているとわかる典型的な例は**図 5** 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．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時に北は 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 を切る方が価値が高いと考えている．このケースはターツが一応そろっており，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．

表 6　順位分布

	1 位率	2 位率	3 位率	4 位率	平均順位	試合数
MCP+ツモ切り	0.185	0.251	0.282	0.280	2.65 ±0.01	30505
MCP	0.180	0.248	0.292	0.276	2.67 ±0.01	62742
BCP	0.194	0.253	0.248	0.270	2.62 ±0.01	44550

図 5　問題のある手牌

表 8　BCP による図 5 の評価値

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

7. おわりに

本研究では強化学習を用いて，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，それをもとに効率の良い和了を行うプレイヤを構築した．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一人麻雀の手を使用するより成功率が向上しており，強化学習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プレイヤは非常に弱かった．

相手を一人麻雀の手を使用する少し強いプレイヤにすることで，自己対戦の結果も少しは改善した．このことから強化学習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わかる．相手が一人麻雀の手を選択し続ける場合にはツモ切りの場合と比べ和了できないケースが増え報酬をもらえるケースが減ると学習に時間がかかり，必要な学習局面数も増えるのではないか予想する．これにより同じ特徴量を使用しているにもかかわらず，テスト時にそれよりも低い結果（特に 1 翻）しか出せなかったのではないかと考える．実際にグラフでは 1 億局面でのテストの結果が一番良かったので，学習局面を増やすことで成功率の向上の余地がある．

和了できなかった時の期待最終順位を現在の点数状況時の期待最終順位としているのも問題がある．通常和了できない場合は他の誰かがツモ和了や自分が放銃など自分の点数が減る可能性が高い．それを無視するのは，全体として楽観的になりやすく，可能性の低い高い手を狙う割に合わない戦略を取りやすくなる．この問題の改良も今後必要になる．

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] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G. et al.: Human-level control through deep reinforcement learning, *Nature*, Vol. 518, No. 7540, pp. 529–533 (2015).

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

[4] 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，*Proceedings of the 20th Game Programming Workshop*, pp. 179–186 (2015).

[5] 角田真吾：天鳳，`http://tenhou.net/` (2014).

[6] 小松智希，成澤和志，篠原　歩：役を構成するゲームに対する効率的な行動決定アルゴリズムの提案，情報処理学会研究報告. GI,[ゲーム情報学]， Vol. 2012, No. 8, pp. 1–8 (2012).

[7] 海津純平, 成澤和志, 篠原　歩：一人麻雀における打ち方を考慮した評価指標に関する研究，*Proceedings of the 20th Game Programming Workshop*, pp. 172–178 (2015).

[8] Heinrich, J. and Silver, D.: Smooth UCT search in computer poker, *Proceedings of the 24th International Joint Conference on Artifical Intelligence* (2015).

[9] 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. 1–11 (2014).

[10] Collins, M.: Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms, *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, Association for Computational Linguistics, pp. 1–8 (2002).

[11] Tesauro, G.: *Practical issues in temporal difference learning*, Springer (1992).

[12] Tesauro, G.: TD-Gammon, a self-teaching backgammon program, achieves master-level play, *Neural computation*, Vol. 6, No. 2, pp. 215–219 (1994).

[13] 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).

[14] 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).

[15] 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).