教師あり学習と強化学習を用いた麻雀AIの開発

松田, 真治
九州大学大学院システム情報科学府

伊東, 栄典
九州大学情報基盤研究開発センター

出版情報：JSAI Technical Report, SIG-KBS. 124, pp.28-33, 2021-11-15. The Japanese Society for Artificial Intelligence
バージョン：

教師あり学習と強化学習を用いた麻雀 AI の開発

Development of Mahjong AI Using Supervised Learning and Reinforcement Learning

松田 真治[1]　伊東 栄典[2]

Shinji Matsuda[1] and Eisuke Ito[2]

[1]九州大学システム情報科学府

[1]Graduate School of Information Science and Electrical Engineering, Kyushu University

[2]九州大学情報基盤研究開発センター

[2] Research Institute for Information Technology, Kyushu University

Abstract: In recent years, game AI has been remarkably evolved. Some game AIs have outperformed top human players for complete information games such as Chess, Shogi, and Go. Compared with AI for perfect information games, game AI is not so strong for imperfect information games such as Poker and Mahjong. We research and development Mahjong AI using machine learning method. In this paper, we divide the internal Mahjong AI function into two parts: supervised learning and reinforcement learning. In the supervised learning part, the AI learns the choices of tiles using the top-ranked human players in Tenho's game records. In the second part, the AI is reinforced selection function during many Mahjong games against three AIs created in the first step supervised learning.

1. はじめに

近年のゲーム分野における AI の進化はめざましく、特にチェス・将棋・囲碁などの完全情報ゲームにおいて、AI は人間のトッププレイヤーを凌駕する成績を残している[1]。近年では不完全情報ゲームであるポーカーや麻雀 AI の開発も活発である[2]。

完全情報ゲームの分野においては、教師あり学習を使った AI や強化学習を使った AI、またはそれらを併用した AI が多数存在している。一方、不完全情報ゲームである麻雀の AI では、教師あり学習を用いた研究は複数存在するものの[3,4,5,6]、強化学習を用いた麻雀 AI は未だ少ない。

我々は、教師あり学習と強化学習の２段階を組み合わせた麻雀 AI 開発を試みている。１段目の教師あり学習パートでは「天鳳」のゲーム記録である牌譜を用いて、実力上位プレイヤーの選択を模倣する AI を作成する。２段目の強化学習パートでは，教師あり学習で作成した AI をベースとしながら、麻雀 AI 同士を対戦させて AI を強化させる。

本論文の構成を述べる。第２節では関連研究を述べる。特に，本論文でも使用した先行研究で提案されたデータ構造を説明する。第３節では教師あり学習による麻雀 AI 開発について説明する。第４節では強化学習による麻雀 AI 開発を述べる。第５節では順位予測モデルの開発について述べる。第６節では自作した麻雀ゲーム環境について説明する。最後に第７節でまとめと今後の課題を述べる。

2. 関連研究

文献[3]で Gao らは、牌譜データを教師データとして、CNN を用いて教師あり学習する手法を提案している。特に先行研究で麻雀の手牌などの表現に用いたデータ構造「One-hot Structure」に代わり、データ構造「Data Plane Structure」を提案している。

表 2.1.に One-hot Structure の例を示す。この構造では 34 行×5 列のテーブルを用意し、行を各麻雀牌の種類に、列をその牌の枚数(0-indexed)とみなす。例えば、0 種類目の牌が 0 枚なら 1 行目の 1 列目に 1 を、4 枚なら 1 行目の 5 列目に 1 を書き込むようにする。

表 2.2.に表 2.1.と同じデータを文献[3]で提案された Data Plane Structure を使って書き換えた例を示す。この構造では 34 行×4 列のテーブルを用意し、行を各麻雀牌の種類に、列をその牌の枚数(1-indexed)とみなす。Data Plane Structure の特徴はある牌を n 枚持っているとき、それはある牌を n-1 枚、n-2 枚、…、1 枚持っていることを同時に満

たすとみなす点である。例えば、0 種類目の牌が 0 枚なら何も書き込まず、4 枚なら 1 行目の 1 列目~4 列目に 1 を書き込むようにする。

文献[3]は Data Plane Structure の利点を述べている。利点として、情報が横に広がりを持つことで CNN との相性が良くなる点、コードの拡張性が高い点、データ空間を 20%削減することができる点などをあげている。

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

3. 教師あり学習

3.1. 牌譜

麻雀での打牌選択の教師データとして、麻雀ゲームの記録である牌譜を用いる。本研究ではオンライン麻雀サイト「天鳳」[7]の牌譜を利用した。天鳳は世界最大手のオンライン麻雀対戦サービスである。天鳳では誰でも麻雀対戦できるだけでなく、対戦のログが牌譜に記録され、無償公開されている。また天鳳は実力によって一般卓・上級卓・特上卓・鳳凰卓の 4 つの卓に対戦のフィールドが分けられているため、レベルの高い卓の牌譜を使うと、実力上位のプレイヤーの選択を学習できる。そのため先行研究でも天鳳の牌譜が利用されている[3,4,5,6]。

本研究では天鳳五段以上の実力者のみがプレイできる鳳凰卓の牌譜を利用した。期間は 2016 年度から 2018 年度のものを利用した。麻雀はルールの違いによって選択にも違いが出てくるため、本研究では最も一般的なルールである「四人打ち・東南戦・赤有り・喰断么九有り」の牌譜のみを利用した。

なお牌譜の解析については小林聡氏のブログ[8]を参考にした。

3.2. データ構造および特徴量

データ構造には表 2.2. に示した Data Plane Structure を採用した。表 3.1. に教師あり学習に用いた特徴量を示す。

表 3.1. 利用した特徴

特徴	Plane 数
自分の手牌	1
自分の赤牌	1
4 人分の捨て牌	4
4 人分の副露	4
ドラ	1
対戦相手のリーチ	3
最後の打牌	1

3.3. 実験

3.3.1. 実験設定

天鳳の牌譜からランダムで場面を抜き出し、その時のゲームの状況を入力、実際に行われた選択を正解データとして教師あり学習をおこない、天鳳の実力上位のプレイヤーの選択を模倣するようなモデルを開発する。表 3.2. に示す 5 つの選択において教師あり学習をおこなう。

表 3.2. モデル

モデル	クラス数
打牌選択	34
リーチ選択	2
ポン選択	2
チー選択	4
カン選択	2

リーチ選択、ポン選択、カン選択はするかしないかの 2 択なので 2 クラス分類、打牌選択は 34 種類の牌のどれを打牌するかの 34 択なので 34 クラス分類、チー選択はするかしないかの 2 択に加えてチーの仕方が 3 択あるので計 4 クラス分類となる。

3.3.2. 教師あり学習モデルの構成

教師あり学習の設定を以下に列挙する。

- モデル：TensorFlow で構成
- オプティマイザー：Adam を使用
- エポック数：200
- バッチサイズ：256

使用したデータを以下に列挙する。

- 教師データ：80 万場面
- 検証データ：20 万場面

表 3.3.にモデル構成を示す。

表 3.3. 教師あり学習モデル構成

Layer	Output Shape
入力層	(N, 34, 4, 17)
畳み込み層	(N, 30, 3, 100)
ドロップアウト層	(N, 30, 3, 100)
畳み込み層	(N, 26, 2, 100)
ドロップアウト層	(N, 26, 2, 100)
畳み込み層	(N, 22, 1, 100)
ドロップアウト層	(N, 22, 1, 100)
平坦化	(N, 2000)
全結合層	(N, 300)
ドロップアウト層	(N, 300)
全結合層	(N, クラス数)

3.3.3. 結果

表 3.4.に教師あり学習で作成したモデルの精度(Accuracy)を示す。打牌選択以外では，実力上位のプレイヤーの選択とモデルの選択が概ね一致している。しかし打牌選択では一致率が低い。リーチやポンやチーやカンでは選択肢が少ないのに対し、打牌選択では選択肢が多いためであろう。学習量を増やす必要がある。

表 3.4. 実験結果

モデル	Accuracy (%)
打牌選択	63.5
リーチ選択	74.4
ポン選択	87.6
チー選択	83.5
カン選択	84.5

4. 強化学習

4.1. 強化学習の改善案

本研究の強化学習による AI 構築は第 6 節で述べる自作の麻雀ゲーム環境を用いている。現段階のゲーム環境は、半荘 1 回が終わるのに約 50 秒かかるため強化学習を効率よく進めることができず、強化学習の結果が出ていない。ここでは強化学習の改善案を述べる。

作成するモデルは表 3.2.と同様にする予定である。

4.2. モデル構成

TensorFlow と tf-agents を利用して DQN(Deep Q Network)を構成する。エピソード数は 150 万に設定する。

教師あり学習の際はドロップ層をモデルに組み込んでいたが、強化学習の場合は学習が進まなかったため、外した。

また強化学習を効率良くすすめるために教師あり学習で作成したモデルを使用して強化学習モデルを初期化する。

表 4.1. 改善予定の強化学習モデル構成

Layer	Output Shape
入力層	(N, 34, 4, 17)
畳み込み層	(N, 30, 3, 100)
畳み込み層	(N, 26, 2, 100)
畳み込み層	(N, 22, 1, 100)
平坦化	(N, 2000)
全結合層	(N, 300)
全結合層	(N, クラス数)

4.3. 行動決定

強化学習での行動を決める方法として、ε-greedy 法を採用する。epsilon は以下の式で決定する。

$$epsilon = \max\left(1 - \frac{episode}{90000}, 0.01\right)$$

これは 0 エピソード目に 1 で始まった epsilon が、全エピソードの 6 割である 90000 エピソード目までかけて緩やかに減少しながら 0.01 に近づき、それ以降は常に 0.01 を維持することを表す式である。

強化学習のエージェントは、選択のある局面になる度に乱数を生成し、乱数が epsilon より大きければ強化学習したモデルを使って選択をし、乱数が epsilon より小さければランダムに選択をする。つまり強化学習を始めた頃はランダムに選択をする割合が高く、終盤にかけて学習したモデルを使って選択をする割合が高くなるということである。

4.4. 報酬

第 5 節で紹介する順位予測モデルを使用する。

5. 順位予測モデル

強化学習では、選んだ行動に対して適切な報酬を設定しなければならない。ここで報酬にどのような値を利用すべきか考える。

まず、麻雀とは点棒の授受をおこなうゲームであるから、点棒の増減をそのまま報酬として設定するというアイデアがある。つまり 8000 点を失えば報酬を-8000 に設定し、8000 点を得れば報酬を+8000 に設定するというアイデアである。これは非常に直感的ではあるが、うまくいかない可能性が高い。

なぜならば麻雀において重要なのは各局の点棒授受ではなく、点の積み重ねにより決まる最終順位だからである。例えば 1000 点を受け取り、プレイヤーの最終順位が 2 位から 1 位へと上昇するような選択と、48000 点を受け取ってプレイヤーの最終順位が 2 位のままであるような選択とでは受け取る点数が低くとも、前者の方が良い。

麻雀の最終順位を最適化する強化学習を行うには、各選択がどれほど最終順位に影響を与えるかを数値化する必要がある。そこで本研究では報酬として、各選択による予測最終順位の変動を利用した。

予測最終順位の変動を報酬として利用するためには、現在の点数状況から最終順位を予測するモデルを構成しなければならない。そこで本研究では天鳳の牌譜データを利用し、入力として点数状況や局状況、正解データとして実際の最終順位を利用した教師あり学習をした。

5.1. データ構造と特徴量

点数を 0 から 5、6 から 10、…、59990 から 59995、59996 から 60000 の 120 個の帯域に分割し、列数 120 のベクトルに対応させ、各プレイヤーの点数が含まれる帯域に対応する列を 1 で埋めた。点数が 0 以下の場合は 0 とし、60000 以上の場合は 60000 として扱う。

表 5.1. に順位予測に用いた特徴を示す。

表 5.1. 順位予測に用いる特徴

特徴
プレイヤー1 の点数
プレイヤー2 の点数
プレイヤー3 の点数
プレイヤー4 の点数
局数
本場数
供託数

5.2. モデル構成

表 5.2. に強化学習で用いたモデルを示す。

表 5.2. 順位予測モデル構成

Layer	Output Shape
入力層	(N, 840)
全結合層	(N, 256)
ドロップアウト層	(N, 256)
全結合層	(N, 512)
ドロップアウト層	(N, 512)
全結合層	(N, 1)

強化学習の設定を以下に列挙する。

- 学習モデル：TensorFlow で構成
- オプティマイザー：Adam を使用
- エポック数：200
- バッチサイズ：256

使用したデータを以下に列挙する。

- 教師データ：80 万局の牌譜
- 検証データ：20 万局の牌譜

5.3. 結果

表 5.3. に平均二乗誤差を示す。比較のため、ランダムに順位を予測するモデルと、ある局面での順位をそのまま最終順位と予測するモデルで、それぞれ 100000 回ずつ予測をした場合の平均二乗誤差も計算した。

表 5.3. 平均二乗誤差

モデル	平均二乗誤差
NN	0.763
ランダム	3.480
そのまま	1.119

表 5.3. から，本論文の NN モデルの誤差が最も少ないことがわかる。

5.4. 報酬の決定

以下の状況を考える。

- ある点数状況 A である。
- ある選択をおこない、点棒が授受される。
- ある点数状況 B になる。

点数状況から最終順位を予測するモデルを P とし

たとき、ある選択の報酬は以下の式で計算する。

$$報酬 = P(点数状況B) - P(点数状況A)$$

6. ゲーム環境構築

強化学習では、AI 同士を対戦させて AI を強くしていく。そのため AI 同士が対戦する環境が必要である。本研究では、天鳳と同じルールで麻雀対戦できる環境を自力構築した。ここでは構築したゲーム環境について述べる。

6.1. 向聴数作成機能

麻雀ゲームでは、聴牌しているか、和了っているかの判定を頻繁に行う。向聴数計算では、各局面でその都度、深さ優先探索で向聴数を計算できる。しかしこの方法は計算時間が掛かる。そこで、あらの氏の方法[9]を参考に、ありうる牌の組み合わせに対し事前に向聴数を算出し、それをハッシュ化したテーブルで管理することにより向聴数計算を高速化した。

6.2. 和了点数計算機能

麻雀の点数計算をおこなう機能を実装した。役の種類と飜数は天鳳仕様[10]で実装した。

6.3. ゲームクラス

ゲームのクラス。局の情報や山の情報などを管理したりゲームの進行を管理したりしている。ゲームアクションクラス、ゲームルーティンクラスから構成されている。

ゲームアクションクラスは、ゲームルーティンクラスやプレイヤーからのリクエストを受け、局の情報や山の情報を変更したり、プレイヤー同士の情報の相互変換を仲介したりするクラスである。例えば、プレイヤーからツモのリクエストがあった場合、山から牌を一枚取得してプレイヤーに渡したり、プレイヤーからカンのリクエストがあった場合、槓ドラをめくった後に嶺上牌をプレイヤーに渡したりする。プレイヤーが和了した場合のプレイヤー間の点棒の授受もゲームアクションクラスの仲介によっておこなわれる。

ゲームルーティンクラスは、ゲームの進行を管理するクラスである。基本的には、ツモ場面、発声 1（ツモ和、リーチ、暗槓）場面、打牌場面、発声 2（ロン和、明槓、ポン、チー）場面、ツモ場面、…と順番に場面を切り替えつつ、副露の割り込みや和了の発生にその都度、対応する実装にした。

6.4. プレイヤークラス

プレイヤーのクラス。プレイヤーの手牌や河や点数などの情報を管理している。プレイヤーアクションクラス、プレイヤージャッジクラスから構成されている。

プレイヤーアクションクラスは、ゲームの情報を AI に対して渡し、AI の出力をゲームクラスへと伝達する、いわばゲームクラスと AI クラスの仲介を担うようなクラスである。打牌の選択、副露の選択、リーチ選択などを仲介する。

プレイヤージャッジクラスはプレイヤーアクションで決定されたプレイヤーの選択が、ルール上可能かを検証するクラスである。例えば、聴牌にならない牌を切ってリーチする 選択や、フリテンロンをおこなうような選択はプレイヤージャッジクラスで弾かれるようになっている。

6.6. AI クラス

AI のクラス。プレイヤーアクションクラスから渡された情報を用いて各選択をおこなう。

6.7. 自作麻雀ゲーム環境の稼働図

図 6.1. に動作しているゲーム環境を示す。強化学習では画面表示する必要は無い。画面表示機能は、AI の動作確認と、将来人間と対戦する場合を考えて作成している。

図 6.1. 開発した麻雀ゲーム環境

7. おわりに

本研究では、不完全情報ゲームである麻雀を対象に麻雀 AI の開発を試みている。そのための手法として、牌譜を用いた教師あり学習による事前学習と、対戦による強化学習での機能強化を行う２段階での AI 実現を提案した。

教師あり学習では「天鳳」の牌譜を用いて実力上位のプレイヤーの選択を模倣した。リーチや副露の選択は上位プレイヤーの選択と一致率が高かったが、打牌選択は一致率が低く、さらなる学習が必要である。

強化学習では十分な学習経験を得られていない。

今後は計算資源の増加と、プログラムの処理効率を上げて、さらなる機能強化を目指す予定である。

参考文献

[１] Silver D., et.al., Mastering the game of Go without human knowledge, Nature, Vol.550, pp.354-359, (2017)

[２] Brown N., Bakhtin A., Lerer A., Gong Q., Combining Deep Reinforcement Learning and Search for Imperfect Information Games, https://arxiv.org/abs/2007.13544 , (2020)

[３] Gao S., Okuya F., Kawahara Y., Tsuruoka Y., Building a Computer Mahjong Player via Deep Convolutional Neural Networks, https://arxiv.org/abs/1906.02146, (2019)

[４] 青木幸聖，穴田一，鳴きを考慮した麻雀 AI，情報処理学会論文誌，Vol.61，No.4，pp.990-995, (2020)

[５] 松田真治，伊東栄典，麻雀牌譜を用いた教師あり学習による打牌推定，情報処理学会 火の国情報シンポジウム 2021, A7-3, (2021)

[６] 福山貴洋，永井秀利，中村貞吾，4 人麻雀への適用を目的とした 1 人麻雀プレイヤへの鳴きの導入，2021 年度(第 74 回）電気・情報関係学会九州支部連合大会, 04-2A-09, (2021)

[７] オンライン対戦麻雀 天鳳 / ログ，https://tenhou.net/sc/raw/, (accessed at Oct.20 2021)

[８] 天鳳の牌譜を解析する(1), https://blog.kobalab.net/entry/20170225/1488036549
天鳳の牌譜を解析する(2), https://blog.kobalab.net/entry/20170228/1488294993
天鳳の牌譜を解析する(3), https://blog.kobalab.net/entry/20170312/1489315432
天鳳の牌譜を解析する(4), https://blog.kobalab.net/entry/20170720/1500479235

[９] あらの（一人)麻雀研究所，向聴数を求めるアルゴリズム， https://mahjong.ara.black/etc/shanten/index.htm, (accessed at Oct.20 2021)

[１０] 天鳳/マニュアル，https://tenhou.net/man/#SPEC, (accessed at Oct.20 2021)