JAIST Repository
https://dspace.jaist.ac.jp/

Title	花札のこいこいにおける方策勾配法とNeural Fitted Q Iteration の適用
Author(s)	佐藤，直之；池田，心
Citation	ゲームプログラミングワークショップ2017論文集，2017：64-71
Issue Date	2017-11-03
Type	Conference Paper
Text version	publisher
URL	http://hdl.handle.net/10119/16089
Rights	社団法人　情報処理学会，佐藤直之，池田心，ゲームプログラミングワークショップ2017論文集，2017，pp.64-71.　ここに掲載した著作物の利用に関する注意：本著作物の著作権は（社）情報処理学会に帰属します。本著作物は著作権者である情報処理学会の許可のもとに掲載するものです。ご利用に当たっては「著作権法」ならびに「情報処理学会倫理綱領」に従うことをお願いいたします。 Notice for the use of this material: The copyright of this material is retained by the Information Processing Society of Japan (IPSJ). This material is published on this web site with the agreement of the author (s) and the IPSJ. Please be complied with Copyright Law of Japan and the Code of Ethics of the IPSJ if any users wish to reproduce, make derivative work, distribute or make available to the public any part or whole thereof. All Rights Reserved, Copyright (C) Information Processing Society of Japan.
Description	

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

花札のこいこいにおける方策勾配法と Neural Fitted Q Iterationの適用

佐藤 直之[1,a)]　池田 心[1,b)]

概要：花札の「こいこい」ゲームは交互 2 人零和不完全情報ゲームの一種で，様々な媒体で多くの人に遊ばれているが研究例が少なく，人間の上級者に匹敵する人工プレイヤが開発されたという話も聞かない．そのため我々は強化学習の方策勾配法と Neural Fitted Q Iteration を用いて強い「こいこい」プレイヤの実装を試みた．それぞれ盤面の低級な特徴量 268 個を入力に用いた人工ニューラルネットワークを状態行動価値の推定に用い，簡単なルールベース人工プレイヤとの反復対戦を通じて適切なパラメータの学習を行った．その結果それぞれ対戦相手から搾取した平均スコアは-0.3 点と 0.5 点となった．

Applying Policy Gradient method and Neural Fitted Q Iteration for Hanafuda Koi-Koi game player

Sato Naoyuki[1,a)]　Ikeda Kokolo[1,b)]

Abstract: Koi-koi game, which is played using Hanafuda playing cards, is a Japanese traditional card game classified as two players turn based imperfect information zero sum game. There are few research article focusing on this game even though this game is popular in Japan. Therefore, we tried to make strong Koi-koi game player by applying two types of reinforcement learning methods. We applied policy gradient method and neural fitted Q iteration. Each player played games against an artificial player which we constructed making its decision in a simple rule based manner. Over 1,000 times game, policy gradient player gained -0.3 score per game and neural fitted Q iteration player gained 0.5 scores in average.

1. はじめに

花札は日本で古来から親しまれてきたカードゲームの 1 つである．簡単なルールと手ごろなゲームサイズを持ち，スマートフォンのアプリとして手軽に遊ばれたり，ビデオゲームの商業タイトルの中でのミニゲームとして登場したりする．しかし一方で花札を対象とした研究は例が非常に少なく，人間の上級者より十分に強い人工プレイヤが作られた例も我々の知る限りでは無い．

そこで我々は強化学習により強い花札の人工プレイヤ作成を目指す．花札を使った遊び方のうち我々は特にルールが簡明な「こいこい」ゲームに着目する．これは交互 2 人ゼロ和不完全情報ゲームで，同様の不完全情報ゲームでは麻雀に形式が似ている．既存の麻雀プレイヤ研究では上級者棋譜の教師あり学習がまず基礎の部分に適用された [1] が，花札ではそうした上級者の棋譜が大量に用意しづらい．よって我々は強化学習を用い，適度な強さの単純なルールベースプレイヤを相手に少しずつ訓練する事で強いプレイヤの獲得を目指す．

本稿で我々は 2 種類の強化学習手法をそれぞれ独立に試みた．1 つは方策勾配法である．この手法ではパラメタライズされた方策を持つエージェントの受け取る報酬を観察し，その獲得報酬の期待値が上昇するようにパラメータを調整していく．この方策中の “目的関数（状態行動の良さを評価する関数）” として我々は人工ニューラルネットワーク（以下 ANN と呼ぶ）を用いる．

もう 1 つの手法として Neural Fitted Q Iteration（以下

[1] 北陸先端科学技術大学院大学
JAIST, Nomi, Ishikawa 923–1211, Japan
[a)] satonao@jaist.ac.jp
[b)] kokolo@jaist.ac.jp

© 2017 Information Processing Society of Japan

図 1 花札のゲーム局面．各プレイヤには公開フィールドとしての「持ち札」，非公開フィールドとしての「手札」が与えられている．山札や場札から特定の手順によってカードを持ち札に移していって，持ち札に「役」をつくってアガリを目指す．

NFQ と呼ぶ）を適用した．この手法は Deep Q Network [13]（以下 DQN と呼ぶ）手法と多くの点で類似している．NFQ は Q 学習と同様にエージェントの各状態行動の結果として受け取れる累積割引報酬の大きさを環境との試行錯誤から見積もっていく．ただし初歩的なテーブル型の Q 学習と違って，Q 関数を ANN で近似し，またエージェントの行動履歴を沢山集めてから Q 関数の更新を行う．そして DQN 手法と NFQ は ANN の更新頻度に違いがある．

2. 花札の「こいこい」ゲーム

対象ゲームについて説明する．花札はトランプのように様々な種類の遊び方を持つが，中でも最も有名な遊び方の「こいこい」を我々は扱う．図 1 はその局面の例である．このゲームは 2 人のプレイヤが交互に自分の手番に手札からカードを場に出して，「持ち札（手札とは異なるフィールドである）」を増やす．持ち札のカードの中で「役」と呼ばれる一定のパターンが完成するとそのプレイヤは「あがり」によって得点をもらう事ができる．以下に詳細を示す．

(1) **初期状態**：各プレイヤは 8 枚の手札を持ち，場には場札が 8 枚表向きに公開されている．また互いに持ち札は 0 枚ずつである．残りのカードは山札として伏せられた状態で場に積まれる．カードは全 48 枚で，12 種類の花が描かれた札 4 枚ずつから構成される．

(2) **先手プレイヤ行動-手札の提出**： まず先手のプレイヤが手札から好きなカードを場に出す．場札に，それと同じ「花」が描かれたカードがある場合には自分の持ち札に加える．その加え方のルールはやや複雑で，場に同じ花のカードが 1 枚または 3 枚だけあった場合は，自分の出した札と併せてそれら全てを持ち札に加える．しかし同じ花のカードが 2 枚だけあった場合のみ，それらのうち好きな方 1 枚と自分の出した札を自分の持ち札に加える．そして場に同じ花のカードが 1 枚も無かった場合は自分は何も持ち札に加えられず，自分が出したカードも新たな場札として追加される．

(3) **先手プレイヤ行動-山札めくり**： 続けて先手プレイヤは山札の一番上にある札を表向きにめくる．その札についても先ほどと同様の処理に基づき自分の持ち札を増やす．つまりめくったカードと同じ花が描かれたカードが場に 1 枚か 3 枚あるときは，めくったカードもあわせてそれら全てを持ち札に加える．2 枚だけあるときはその片方とめくったカードを持ち札に加え，1 枚もなければ単に場札に加える．

(4) **後手プレイヤ行動**：次に後手プレイヤが同様に手札の提出と山札のカードをめくる手続きを行う．以上の手続きは両プレイヤ交互に，どちらかのプレイヤの持ち札に「役」が完成するか提出できる手札が無くなるまで繰り返される．

(5) **「あがり」と「こいこい」の選択**：片方のプレイヤが持ち札に「役」を完成させた時，そのプレイヤはただちに「あがる」事によって役に応じた得点をもらうか「こいこい」を宣言する事によってゲームを継続する事を選ぶ事ができる．「こいこい」は通常，自分が更に高い点数の役を狙えそうでなおかつ相手が役の完成から遠そうな場合に選ばれるオプションである．

(6) **ゲームの終わり**：片方のプレイヤがあがった場合に，そのプレイヤに得点が与えられてゲームが終わる．あるいはどちらのプレイヤもあがらないまま 8 枚の手札を使いきった場合にゲームが終わり，この場合はどち

© 2017 Information Processing Society of Japan

らのプレイヤにも得点が与えられない．このゲームは確率的な要素も大きく，通常は何回もゲームを繰り返してその合計スコアを競い合う．

以上がこいこいゲームのルールであるが，役の種類や特定条件下での手札の配り直し等のローカルルールが加えられる事もよくある．特に近代的なオンラインゲームとして提供される場合には，既存のものとの差別化のためか，かなり大がかりな独自の特殊ルールが導入されている事もある．

そのように様々なルールがある中で本稿で採用するルールはかなりシンプルで，認められる役は五光（10 点）四光（8 点）雨四光（7 点）三光（5 点）赤短・青短・猪鹿蝶（各 5 点）タネ・短冊・カス（各 1 点から 1 枚増加で 1 点追加）のみである．しばしば用いられる「手四」と「くっつき」は無し[*1]とする．また一定条件下での得点の倍増に関するルールも取り入れない．

3. 既存研究

不完全情報のカードゲームに関する研究はポーカーを対象にしたものが盛んである．Martin らは Counter Factual Regret と呼ばれる指標の最小化を強化学習で行う事で Heads-up limit Texas holdem ルールで ϵ ナッシュ均衡の導出に成功した [2]．また Tammelin はその発展形としての指標である CFR+を提案し学習収束速度の向上を示した [3]．さらにゲーム規模を 2 人用に限定していない多人数ポーカーにて Counter Factual Regret の適用を試みた研究 [4] やナッシュ均衡から敵モデルを構築して搾取する研究がある [5]．

また花札のこいこいと同様に，役作りとあがり，あがりを延期し更なる高得点を狙うオプションを備えた不完全情報ゲームには麻雀があるが，麻雀では上級者棋譜が使えるため教師あり学習がしばしば用いられる．水上らは教師あり学習からの麻雀プレイヤの作成 [1] とその発展形としてモンテカルロシミュレーションを加えた手法を提案した [6]．さらに麻雀では人工プレイヤの技術発展を目的とした競技用のサーバーが用意されている [7]．

花札のこいこいを扱った研究として我々は過去に，方策勾配法を用いたエージェントの性能を試験 [11] した．本稿の方策勾配法との違いは目的関数のモデルであり，前回はゲームの固有知識を利用した高級な少数の特徴量の線形和モデルを用いて方策勾配法を試験した．一方で今回は，ゲームの固有知識にあまり頼らない低級で多数の特徴量を用いた ANN をモデルとし「性能の更なる向上」と「ゲーム固有知識に頼らない学習の成功」を目指している．更に，本稿では前回扱わなかった NFQ 手法を試みている点も新規である．

[*1] 初期状態で手四の形ができていたらゲームはやり直しとし，くっつきはそのままゲームを続ける．

4. 適用手法

我々はアプローチとして強化学習の方策勾配法 [8] と NFQ [12] を選んだ．ポーカーは各状態での行動がコール・レイズ・フォルドの 3 つだけだが花札は全カードについて 48 種の行動が想定されるため行動政策の定式化が難しい．そのためポーカーの既存研究のように CFR を適用するアプローチには計算コストの大きさによる困難が予想される．また麻雀のように利用可能な上級者棋譜が見当たらないため教師あり学習も困難である．そこで強化学習による動的な強さ向上を目指した．

また方策勾配法と NFQ という 2 つの手法を選びそれぞれ独立に実験したのは，両者が異なる長所を持つためである．一般に方策勾配法のほうが学習の難度が下がると言われている（状態価値や状態行動価値を推定しなくても，報酬を最大化する行動のみ求められれば良いため）[10]．その一方，NFQ では各状態行動の期待（割引）累積報酬値が Q 関数の出力値として具体的に得られるため，今後その数値を他の目的に応用できる可能性がある．例えば，人間プレイヤを楽しませるための人工プレイヤ作成のために合計の獲得点数を相手と同じ程度に合わせようとする行動選択を行う用途が想定できる．また人間プレイヤの教育用として，最も適切な着手を提示だけでなく「各着手それぞれの具体的な良さ（獲得スコア期待値）」を示す事は，

- 自らの着手選択が最善でないときにも，他着手の中での相対的な良さのフィードバックを受け取れる
- 自分が普段選ばないような種類の着手についても良さの度合いを知ることができ，ゲーム全体への理解が深くなる

といった応用可能性から有益であると期待できる．

4.1 方策勾配法

方策勾配法は「パラメタライズされた方策で行動するエージェント」の得る期待報酬を，報酬に対する勾配方向に各パラメータを動かす事で増大させようとする．一口に方策勾配法といっても様々な流儀のものがあるが，我々が用いる方法は五十嵐らが提案した手法を参考にしている [9]．

まずエピソード開始から t 回目の行動を行う状態 s_t で行動 a_t をエージェントが選択する確率（つまり方策）を

$$\pi_a(a_t|s_t;\vec{\omega}) = \frac{exp(E_a(a_t,s_t;\vec{\omega})/T_a)}{Z_s} \qquad (1)$$

と定めているものとする．ここで $E_a(a_t,s_t;\vec{\omega})$ は，s_t での a_t の選ばれやすさを表す「目的関数」と呼ばれる指標である．T_a は温度パラメータで，方策で選ばれる行動のバラつきに影響する．Z_s は s_t でのエージェントの全可能行動についての選択確率値を 1 以下，総和 1 にし正規化するための項であり，式 (1) の右辺の分子を全可能行動に対し足し合わせる事で求められる．

© 2017 Information Processing Society of Japan

そしてある 1 回のエピソードに対応した報酬を r とし，i 回目のエピソードの報酬を r_i，報酬とそのエピソード生起確率の積 R の期待値を $E[R;\vec{\omega}]$ と表すとき，この期待値を重みベクトルの調整によって極大化しようとするためのパラメータ更新式を考える．方策が式 (1) のとき $\nabla_{\vec{\omega}}E(R;\vec{\omega})$ は一般に，

$$\nabla_{\vec{\omega}}E(R;\vec{\omega}) = \frac{1}{T}\sum_i r_i p(x_i;\vec{\omega}) \sum_{t=1}^{L_i} \Big\{ \nabla_{\vec{\omega}} E(a_{i,t}, s_{i,t};\vec{\omega}) - \sum_{a' \in A_{s_{i,t}}} \pi(a'|s_{i,t};\vec{\omega}) \nabla_{\vec{\omega}} E(a', s_{i,t};\vec{\omega}) \Big\}$$

という形で計算される．ここで $p(x_i;\vec{\omega})$ はそのエピソードの生成確率で，$\prod_{t=1}^{L_i}\pi(a_{i,t}|s_{i,t};\vec{\omega})$ に等しい（ただし L_i はエピソードの行動ステップ数）．これを用いて

$$\vec{\omega} \leftarrow \vec{\omega} + \eta \nabla_{\vec{\omega}} E(R;\vec{\omega}) \tag{2}$$

と示されるパラメータ更新を行えばよい．ただし η は小さな正の定数である．

このように式 (2) のような更新式を，エピソードを繰り返しながら重みパラメータに適用し続けて，なんらかの終了条件を満たしたときに繰り返しを打ち切るのが本稿で利用する方策勾配法の概要である．

4.2 Neural Fitted Q Iteration

NFQ [12] は Martin が提案した手法であり，ANN を利用した Q 関数学習手法という点では DQN [13] と似ている．NFQ は Q 学習と非常に多くの点が共通しているが以下 2 点が特徴である．

(a) ANN で近似した価値関数を用いている．

(b) 環境との相互作用を行うフェーズと価値関数の調整を行うフェーズが大きなかたまりで分離されている．

(a) の特徴は Q 学習の設定の範疇といえるが (b) は NFQ に特有である．

まず (a) について述べる．このエージェントの手法では通常の Q 学習と同様に，状態，行動，環境からの報酬，そして次状態を観測し，価値関数と累積割引報酬のズレが小さくなるようにパラメータを更新する．その際に NFQ は，通常のテーブル型の Q 学習と異なり，ANN の入出力により近似された価値関数を用いる．このような工夫は TD-Gammon などにも見られ，状態行動空間のサイズが膨大になる問題で有効である．

この場合には価値関数の更新は，以下の式 (3, 4) のように 2 乗誤差を小さくするように進められる．

$$L_{\vec{w}} = \sum_k \{target_k - Q_{\vec{w}}(s_k, a_k)\}^2 \tag{3}$$

$$\vec{w} \leftarrow \vec{w} - \alpha \nabla_{\vec{w}} L_{\vec{w}} \tag{4}$$

ここで $target_k$ は状態 s_k と行動 a_k に対応した価値関数の目的値であり，状態 s_k から行動 a_k を実行した「以降の累積割引報酬」である．$\vec{w}$ は ANN の重みパラメータであり，α は学習率である．

Algorithm 1 Neural Fitted Q Iteration

```
k = 0;
Q = initialized_ANN();
while k < N do
  x = 1;
  P = gen_dataset();     # {(input^l, target^l), l = 1, ..., |P|}
                         # where input^l = (s^l, a^l),
                         # target^l = r + γ max_a(Q(s^{l+1}, a))
  Q = train(Q, P);
  k = k + 1;
end while
```

こうして関数近似は膨大な広さの状態行動空間上でのエージェントの学習を可能にするが，とはいえ ANN のような非線形モデルで近似を行った場合，一般に学習が不安定になりやすいという問題点に対処する必要がある．例えば Deep Q Learning [13]（以下 DQL と呼ぶ）は NFQ と同じように「ANN を価値観数とした Q 学習」であり，Experience Replay という工夫でその対策をしている．Experience Replay は状態行動と報酬の履歴をメモリに保存し，一定量ずつランダムにサンプルしながら学習に使う方法で，パラメータの更新をなだらかにする．

NFQ では (b) の手続きによって学習を安定化させている．つまり，重みの更新を行わずに行動と報酬の履歴を蓄え続けてから，ANN の学習を行う．一度の学習に多数の訓練データを用いる点は DQL と同じだが，Q 関数の更新が飛び飛びの頻度で行われる点は「DQN でデフォルトで設計されている形式の Experience Replay」とは異なっている．この方法によって NFQ はエージェントの方策が短い周期で激しく変動するリスクを回避している．

一般的な NFQ の手順を Algorithm1 に示す．初期化した ANN を Q 関数に用い，エージェントに環境で行動させて一定数のデータを蓄えてから，そのデータを用いた教師あり学習を ANN に適用している．こうして ANN の学習が少ない頻度（DQL に比べて）のオフライン学習になるため，Rprop や Adam といった高度な最適化手法が学習に適用できる点を Martin は NFQ の利点として指摘している [12]．また NFQ で用いる ANN は，DQN の場合と違い，深層型である必要はないが我々は深層型のネットワークを利用した．

4.3 使用特徴量

我々は方策勾配法の「目的関数」と NFQ の Q 関数の両方において ANN を用いるが，両者の場合で同じ特徴量を用いているためここで述べる．我々の特徴量は「カードの位置」，「現在のターン数」，「相手の手札になさそうな月」の 3 種である．

© 2017 Information Processing Society of Japan

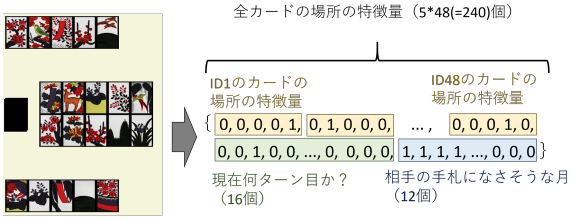

図 2　局面の特徴量ベクトル化

まず「カードの位置」の特徴量を以下のように用意した．花札に使用される 48 のカードそれぞれに ID をつけて，行動判断の主体となるエージェント（自分）から見てそれぞれのカードが以下の 5 種のうちどの場所にあるかを特徴とする．

- 自分の手札の中
- 自分の持ち札の中
- 相手の持ち札の中（※持ち札は場に公開される）
- 場札の中
- それ以外の場所，つまり自分からは見えないどこか（※具体的には相手の手札の中か山札の中）

こうして「ID1 のカードは自分の手札にあるか？」,「ID1 のカードは自分の持ち札に含まれるか？」, . . . ,「ID48 のカードは自分から見えない位置にあるか？」という $5\times48(=240)$ 種のバイナリ特徴がゲーム局面を表現するため使われる．

次に現在のターン数に関する特徴量は，「現在ゲームの 1 ターン目である」,「2 ターン目である」というバイナリ特徴である．花札のターンは 1 ターン目から最大 16 ターンまでしかないので，この特徴量は 16 個である．

最後に「相手の手札になさそうな月」とは，エージェントの相手プレイヤが見逃した場札の月を表す特徴である．例えば，場に 4 月の札と 8 月の札があるとき相手プレイヤが 4 月の札も 8 月の札も手札から出す（キャプチャ）事をせず，単に別の月の札を場に提出した（非キャプチャ）場合，相手の手札には 4 月の札も 8 月の札もないことが推測される．なぜならば，花札において非キャプチャ手はキャプチャ手よりも一般に大きく不利な行動となるためである．この特徴量は他 2 つと比べて強くゲーム固有の専門的知識に依るものだが，しかし人間がよく利用してゲーム勝敗への影響も大きい事が予想されるため，本稿ではひとまず導入の上で実験を行った．

まとめると，図 2 のように合計 268 個の特徴量が盤面の表現に使われる．方策勾配法でも NFQ 手法でも，「ある状態 s_1 におけるある行動 a_1」の良さは，s_1 に a_1 を適用した直後の盤面の特徴ベクトルをニューラルネットワークの入力にしたときの出力値として得る．

5.　実験 1 ：方策勾配法プレイヤ

まず我々は方策勾配法を用いた人工プレイヤの性能を評価するため対戦実験を試みた．

5.1　実験条件

5.1.1　提案プレイヤ

我々の提案手法による人工プレイヤは 4.1 項で示した通りの方策勾配法で動作し，式 (1) の「目的関数」には 4.3 項に示した特徴量の ANN を用いた．1 エピソードは 1 ゲームが終わるまでであり，ANN の受け取るエピソード報酬 r は提案プレイヤの終局スコアから相手役プレイヤのスコアを差し引いたものである．また式 (1) での温度 T は常に 1.0 とし，式 (2) の学習率を 0.1 とした．

5.1.2　相手プレイヤ

対戦相手も人工プレイヤで，ゲーム知識に基づく If-then ルールで処理を書き下したルールベースプレイヤである．このルールベースプレイヤは，「完全にランダムに行動を決めるプレイヤ」と 1,000 戦して平均獲得点数 2.52 点をおさめる程度の強さがあった．具体的には「“松の光” は 20 点」「“梅のカス” は 1 点」という具合に事前に各札に点数を割り当てておき，自分が札を取れる場合は点数の合計が最大になる取り方を選び，自分が手札を場札に提供せざるを得ない状況ではなるべく点数最低の札を出す．このプレイヤの挙動は本稿の特徴量を用いた ANN で表現可能である．

5.1.3　プレイヤ共通の設定

なお，このゲームの醍醐味である「こいこい」をするかしないかの判断はどちらののプレイヤも原始モンテカルロ手法のシミュレータにより行う．すなわちこの実験に登場する人工プレイヤはどれも，役を完成させた時に「こいこい」を宣言した後のゲーム展開を 10 回シミュレーションする．そのシミュレーションの中でゲームを進めるのは先手後手ともにランダム行動プレイヤだがそのプレイヤはもはや「こいこい」はせず，あがれるチャンスには必ずあがる．このシミュレーションの平均獲得スコアが，ただちに「アガリ」を選んだ場合より高い場合は実際のゲームで「こいこい」を宣言する．

5.1.4　対戦条件

この方策勾配法プレイヤはルールベースプレイヤと 8,200,000 エピソード（GPU 無しのマシンで高速化処理なしで処理時間約 20 日間程度）にわたって対戦した．ただし方策勾配法プレイヤは常に先手番でゲームを始める．実験環境は自作プラットフォームを用いた．

5.2　結果

学習の過程における，（学習を一旦中断した状態での）1,000 戦平均スコアを図 3 に示す．最初の 1,000,000 エピソード付近まで性能は向上し続けているが，1,000,000 エピソードからおおむね −0.0 点から −0.5 点の間にスコアが分布し，ルールベースプレイヤに負け越し続けている．1,000 戦対戦平均スコアの 95 %信頼区間は ±0.30 点である事を確認しているため，学習終盤の多くの観測点において相手プレイヤより有意に弱い．同様の実験をもう 1 試行，

© 2017 Information Processing Society of Japan

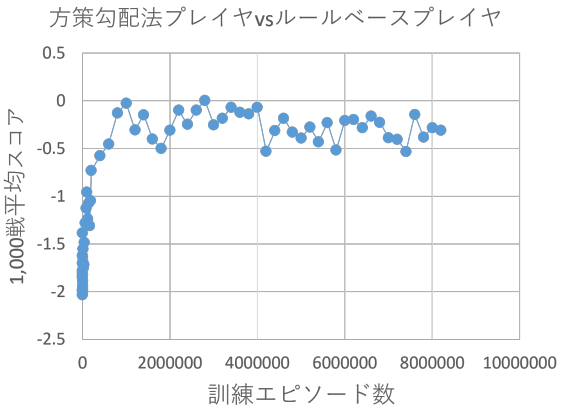

図 3 方策勾配法を用いての対ルールベースプレイヤの 3000 対戦平均獲得スコア

5,000,000 エピソードまで実行してみたが点数変化の傾向と点数そのものがほぼ同じになった．

この方策勾配法プレイヤは，原理的にはルールベースプレイヤと互角以上になる性能限界を持つ．というのも，この方策勾配法プレイヤの ANN モデルは，ルールベースプレイヤの挙動を表現可能である．そのため今回の実験でルールベースプレイヤより強くならなかった原因は，学習の条件として温度パラメータと学習率が常に一定だったり，学習設定がナイーブにすぎた点にあると我々は考えている．

6. 実験２：NFQ プレイヤ

次に我々は NFQ プレイヤの性能を評価するため対戦実験を試みた．

6.1 実験条件

6.1.1 人工プレイヤの設計

提案プレイヤは NFQ により動作する．各状態行動を項の特徴量を入力した ANN で評価し，ϵ-greedy 方策により着手を決定した．ϵ パラメータの値は，訓練データ生成時は 0.10 で，性能評価のための勝利スコア集計を行うときは 0.0 とした．

訓練データは入力が局面の特徴ベクトルで，出力がその局面から「相手プレイヤの手札」と「山札」の中身と順序をシャッフルしながらの 100 回の対戦シミュレーションから得た平均結果スコアである[*2]．データが 60,000 件たまるごとに ANN の重み更新が行った．各データは 1，3，5，7，9，11 ターン目の局面それぞれ 10,000 件ずつから成る．各更新後には性能データ集計のための 1000 回対戦をそれぞれ行った．

ANN は入力層と出力層の間に 3 層の隠れ層を持ち，それぞれに 60 個ずつの隠れニューロンを持つ．また各隠れニューロンの発火関数は ReLU 関数で，入出力層のニューロンの発火関数は線形関数である．ネットワークの学習機構は python3.5 の Keras ライブラリにより実装され，最適化の設定には「adam」を用い，学習のオプションには Early Stopping（torelance = 0）を設定した．最適化手法に関するその他の細かなパラメータは現時点での keras ライブラリのデフォルト値をあてはめた．

[*2] そのためブートストラップが一切行われず，Q 学習というよりモンテカルロ法である．しかし全ての意思決定で終端状態に遷移する特殊なケースの Q 学習とみなす事もでき，説明の簡便さを重視して本稿の説明では Q 学習および NFQ 手法の一種と位置づけ記述している．

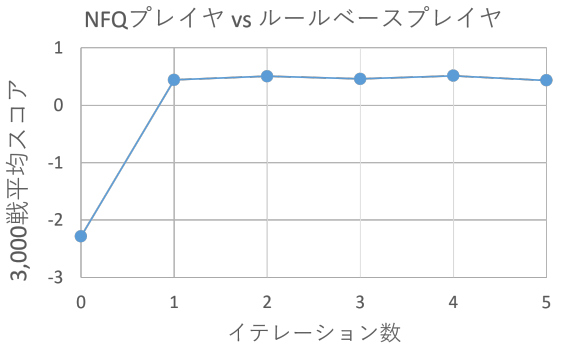

図 4 NFQ を用いての対ルールベースプレイヤの 3000 対戦平均獲得スコア

6.2 相手プレイヤと共通の設定

前項の方策勾配法プレイヤと同様の相手役ルールベースプレイヤと対戦した．さらに，13 ターン目以降に両者が共通のルールベースにより動く事と「こいこい」宣言の判断を両者がシミュレーションにより行う点も前項と同様である．

6.3 対戦の設定

この NFQ プレイヤはルールベースプレイヤと，重みが 5 回更新されるまで対戦した．重み 1 回の更新につき GPU 搭載のマシンで約 5 時間（その大部分が訓練データの生成にかかる）を要した．方策勾配法プレイヤは常に先手番でゲームを始める．実験環境は自作プラットフォームを用いた．

6.4 結果

イテレーションごとの獲得スコアの推移を図 4 に示す．初期状態では-2 点ずつ搾取されているが，1 回重み更新すると +0.5 点くらいの搾取をルールベースプレイヤから行えている．この試行の 95 %信頼区間は ±0.16 点である．また「こいこい」は先手有利なゲームであり提案プレイヤは常に先手で対戦したが，ルールベースプレイヤ同士の 3000 回対戦だと先手の平均獲得スコアは +0.051 点である．

つまり標準偏差や先手番の有利を考慮に含めても，提案プレイヤはルールベースプレイヤより有意に強くなった事が解る．1 イテレーション目以降に有意な性能向上は見ら

© 2017 Information Processing Society of Japan

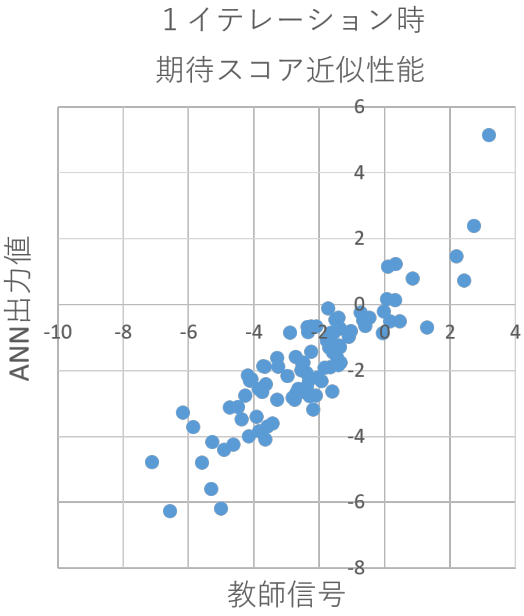

図 5 1 イテレーション後の ANN による局面結果スコア予測．訓練およびテストデータからランダムに 95 件のサンプルを選び，教師信号を横軸に，入力ベクトルからの ANN 出力を縦軸にプロットしている．

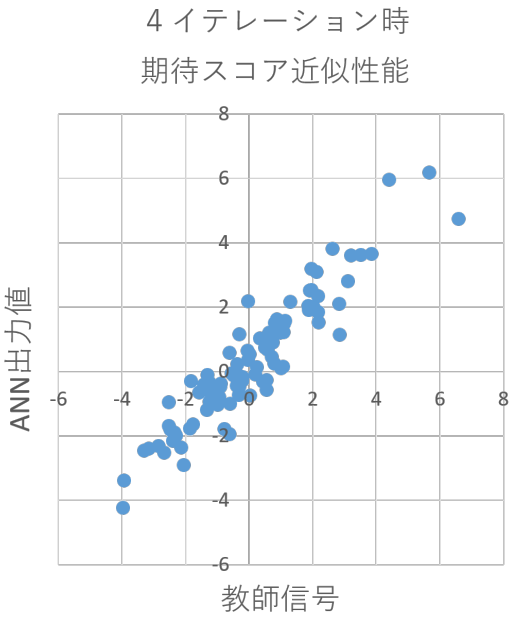

図 6 4 イテレーション後の ANN による局面結果スコア予測

れないが，これは 1 イテレーション目で既にほとんど限界性能に達したためと我々は考えている．

またイテレーションごとに ANN が局面の結果スコアを正しく予測できているかを確かめる予備実験を行った．訓練データとテストデータからランダムに，1，3，5，7，9，11 ターン目の局面を 15 件ずつ取り出して教師信号と ANN 出力値が近い値になっているか比較した．結果を図 5，6 に示す．図中のプロットはおおむね右上がりの直線状に分布しており，学習はおおむねうまくいっているように見え

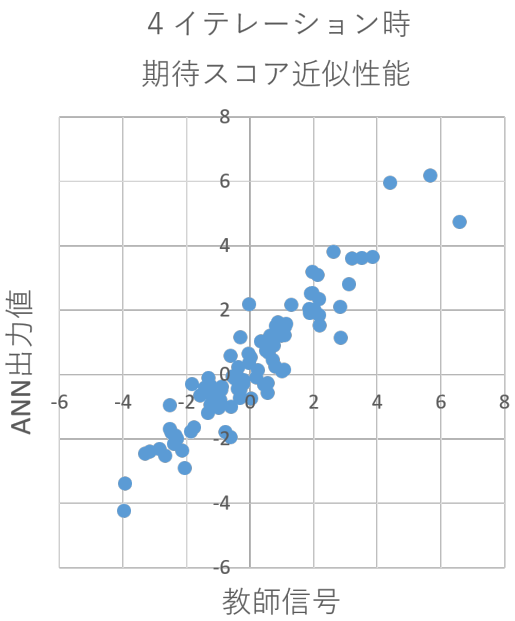

図 7 3 層 ANN による局面結果スコア予測

図 8 NFQ プレイヤがルールベースより優れた判断を行った局面の例．この場合，光札 4 枚による 8 点役の完成を目指しても，敵に 5 枚中 2 枚の光札を既に取られているため完成できない．よって，高得点役の札を取るよりもカスやタネの低得点役（1 点）の札を地道に集める判断が正解となる．

る．サンプルの教師信号値と ANN 出力値の分布が 4 イテレーション目で全体的に大きくなっているのは提案プレイヤの性能が向上して全体的に獲得スコアが向上したためと考えられる．具体的な平均 2 乗誤差の値は，テストデータ 6,000 件に対して 1 イテレーション目と 4 イテレーション目の学習後 ANN でそれぞれ 1.0，0.74 である．また図 7 は文献 [11] で我々が隠れ層 1 層のみの ANN で同様の実験を行った結果である．図 5，6 どちらの場合も隠れ層 1 層のみの時に比べてスコア近似の良さがあきらかに改善している事が見て取れる．

提案プレイヤがルールベースプレイヤよりも優れた判断を行った局面例を図 8 に示す．この局面でプレイヤの有力な着手は 2 通りある．柳の札を出して光札とタネ札を 1 枚ずつ取得するか，荻の札を出してカス札とタネ札を 1 枚ず

© 2017 Information Processing Society of Japan

つ取得するかである．1 点役のカスやタネと違って，光札は 4 枚集めると 8 点もの得点がもらえるが，現局面で既に 5 枚中 2 枚の光札を敵に取られてしまっているため光役の完成は不可能である．我々のルールベースプレイヤは札の一般的な強さだけ考慮した点数付けをするため，ここで光札を取りにいって失敗するが，NFQ プレイヤはより細やかな状況判断によって得点期待値の高い着手を選ぶ事ができる．

7. まとめと今後の課題

我々は花札の「こいこい」ゲームを対象に，方策勾配法による人工プレイヤ実装を試みた．方策勾配法では設定のためもあるのかもしれながルールベースプレイヤより強いプレイヤは得られなかったが，NFQ 手法ではルールベースプレイヤに有意に勝ち越す人工プレイヤの獲得に成功した．

今後の課題として，「こいこい」のより複雑なルールを適用した場合の性能の確認，およびそうしたルール上で動作する既存の花札人工プレイヤの対戦などがある．

参考文献

[1] 水上直紀, 中張遼太郎, 浦晃, 三輪誠, 鶴岡慶雅, 近山隆. 降りるべき局面の認識による 1 人麻雀プレイヤの 4 人麻雀への適用. The 18th Game Programming Workshop 2013，pp.1-7 (2013).

[2] Martin Zinkevich, Michael Bowling, Michael Johanson, and Carmelo Piccione . Regret Minimization in Games with Incomplete Information. Advances in neural information processing systems 2007, pp.1729-1736 (2007).

[3] Tammelin Oskari. Solving large imperfect information games using CFR+. arXiv preprint arXiv:1407.5042 (2014).

[4] Risk Nick Abou, Szfron Duane. Using counterfactual regret minimization to create competitive multiplayer poker agents. Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1. International Foundation for Autonomous Agents and Multiagent Systems 2010. p. 159-166 (2010).

[5] GANZFRIED Sam, SANDHOLM Tuomas. Game theory-based opponent modeling in large imperfect-information games. The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2. International Foundation for Autonomous Agents and Multiagent Systems, pp.533-540 (2011).

[6] 水上 直紀, 鶴岡 慶雅．牌譜を用いた対戦相手のモデル化とモンテカルロ法によるコンピュータ麻雀プレイヤの構築，The 19th Game Programming Workshop 2014，pp.48-55 (2014).

[7] 「麻雀サーバーの紹介」, http://www.logos.ic.i.u-tokyo.ac.jp/m̃izukami/slide/majong_server.pdf (accessed 2017-06-22).

[8] Williams Ronald J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning.8(3-4) pp.229-256 (1992).

[9] 五十嵐治一, 石原聖司, 木村昌臣. 非マルコフ決定過程における強化学習—特徴的適正度の統計的性質—. 電子情報通信学会論文誌 D 90.9 pp.2271-2280 (2007).

[10] Reinforcement Learning: An Introduction Second edition, in progress ****Draft****. http://ufal.mff.cuni.cz-/ straka/courses/npfl114/2016/sutton-bookdraft2016sep.pdf (accessed 2017-06-22).

[11] 佐藤 直之，池田 心，上原 隆平．花札の「こいこい」ゲームの強化学習によるコンピュータプレイヤ，第 38 回ゲーム情報学 (GI) 研究発表会，2017-07-15.

[12] Martin Riedmiller. Neural fitted Q iteration-first experiences with a data efficient neural reinforcement learning method. Lecture Notes in Computer Science: European Conference on Machine Learning, pp. 317328(2005).

[13] Mnih Volodymyr, et al. Human-level control through deep reinforcement learning. Nature 518 pp.529-533 (2015).

© 2017 Information Processing Society of Japan