格闘ゲームにおける模倣 AI
Imitation AI for a Fighting Game

岩島　功太郎
Kotaro Iwashima
法政大学情報科学部コンピュータ科学科
E-mail: kotaro.iwashima.3t@stu.hosei.ac.jp

Abstract

In fighting games, computer-controlled players usually have few patterns of their behaviors. Therefore, human players often get tired of the games by remembering the behavior patterns of the computer-controlled players while playing the games repeatedly. Also, players may want to fight against specific players such as champions of tournaments and opponents against whom they are not good at fighting. This paper proposes imitation AI that imitates the behavior of a player. This AI combines Q-learning and a random forest. It records what actions the imitation target player took in various situations of a game so that the AI can learn to imitate the optimal behavior in a specific situation. The proposed AI can imitate a target player with Q-learning by using less situation data than the rule-based AI. The proposed AI can well treat a situation that is not observed in the imitation target player, because the reward necessary for Q-learning is decided by a random forest. This paper presents an experiment to evaluate the proposed AI by comparing it with the rule-based AI and the AI of Q-learning only. Also, the effectiveness of the random forest is verified by reducing the situation data for the AI to learn.

1. はじめに

格闘ゲームには対人戦と対 COM 戦のモードがある．対人戦は人間同士で戦うため，相手によって行動パターンが異なる．それに対し，対 COM 戦はコンピュータが制御するキャラクタと戦うため，行動パターンの変化に乏しい．そのため，プレイヤは繰り返しプレイすることで COM の行動パターンを憶え，飽きてしまう．対 COM 戦でも様々な行動パターンをもつ相手と戦いたい，苦手な相手と練習したい，大会のチャンピオンと戦いたいというニーズが考えられる．

本論文ではコンピュータがプログラムで操作するキャラクタを AI と呼ぶこととする．本研究では Q 学習とランダムフォレストを用いて，プレイヤの行動を模倣する AI を実現させる．Q 学習とは強化学習の一種であり [1]，自身の行動に対する報酬を受け取りながら，試行錯誤を重ねることで最適な行動を自ら学習するものである．対戦中の状況と，その状況でプレイヤはどのような行動を選択したかを記録したデータ（以下「状況データ」という）を Q 学習に用いることで，その状況での最適な行動を推定する．また，非観測な状況でも最適な行動が選択できるように，決定木のアルゴリズムの一種であるランダムフォレストを用いる [2]．決定木とは機械学習の一種であり，木構造を用い，分類や回帰を行うものである．模倣対象者の行動そのものを学習させるため，学習後の AI は対象者の行動を模倣するようになると考えられる．この手法により，AI を状況データから自律的に学習させる．ゲームは，AI 研究のための対戦格闘ゲーム FightingICE を用いる [3]．

2. 関連研究

模倣学習する AI の研究には様々なものがあるが，エキスパートの行動を観測することで状況に対する適切な行動を学習させるものが代表的である [4]．エキスパートから適切な行動を模倣することを目的とするため，ある状況に対して 1 つの適切な行動が選択される．しかし，格闘ゲームのように 1 つの状況で複数の行動が考えられるような場合を再現することはできない．一方，格闘ゲームにおいて模倣対象であるプレイヤのプレイデータから行動を模倣する研究がある [5]．この研究では，プレイヤの誤選択や癖，弱点などを含めて対象を忠実に模倣することを目的とし，ある状況で複数の行動をとる場合でも模倣できる．しかし，この手法はルールベースでの実装のため，膨大なプレイデータが必要であることや，非観測な状況ではうまく模倣できないという問題点がある．

非観測な状況で学習させる方法として，Q 学習がある [1]．Q 学習を用いてゲーム AI を作る研究では，カードゲームの Hearts をテーマとしたものがある [6]．手法として，状態価値関数を用いることで現在の局面を評価することや，モンテカルロ法を用いて非観測な状態を推定することを提案している．この手法で学習させた AI と，ルールベースで実装した AI を対戦させる実験で，学習 AI はルールベース AI よりも優れた戦略をとることに成功した．また，Q 学習を用いて人間らしいゲーム AI を作る研究では，生物学的制約に基づき人間的な振る舞いを自動獲得するものがある [7]．生物学的制約とは，操作ミスや見間違い，ゲーム内の状況を認識してからの操作の遅れなど，人間がプレイするときに必ず生じるものである．この制約を Q 学習に導入し，ビデオゲームの Infinite Mario Bros で AI を学習させた．実験の結果から，生物学的制約の導入により，人間的な振る舞いの獲得に成功している．

FightingICE を題材にした研究として，相手の行動を予測するものがある [8]．この研究では，相手の位置を線形補外によって予測し，相手の行動を k 近傍法によって予測する．これらを組み合わせることによって相手の未来の情報を予測することができ，手法の有効性が証明された．さらに，決定木分析により相手の行動を推測する研究がある [9]．各フレームで取得できるキャラクタの座標や HP などの情報から分析している．しかし，行動ルー

Supervisor: Prof. Hiroshi Hosobe

ルを複数持つ相手や，こちらの行動を学習によって予測する相手に対しては有効ではなかった．

3. 格闘ゲーム FightingICE

格闘ゲームとはプレイヤがキャラクタを操作し，主に1対1で戦う対戦型ゲームである．FightingICE (図1) は AI の研究目的のために作られた格闘ゲームである [3]．Java 言語によって AI を実装できる．1ラウンド 60 秒であり，1/60秒を 1 フレームと呼ぶ．AI は毎フレームでお互いの行動や距離などの情報を取得できる．

キャラクタに設定されるパラメータは HP とエネルギーがある．HP は相手から受けたダメージの量を表わす．HP の初期値は 0 であり，相手の攻撃を受けると負の値に減少していく．最終的に HP が高い方が勝利となる．エネルギーは技を出すために使用する値である．エネルギーの初期値は 0 であり，相手からダメージを受けたり，自分が相手にダメージを与えたりすることで増加し，自分が技を使うことで減少する．攻撃には様々な技があり，技によってエネルギーの使用量が異なる．強力な技であるほどエネルギーの消費量が大きいが，自分のエネルギーが技の使用量未満の場合は，その技を出せない．

また，このゲームは AI の認識能力が制限されるという特徴がある．人間の認知能力の限界を模倣するために，AI は相手の情報(位置や行動など)について 15 フレーム前のものしか認知できないという制約が与えられている．そのため，AI は人間離れした行動をとれないようになっている．ゲームには、4 人のキャラクタがいるが，本研究では ZEN というキャラクタのみを使用する．

図 1　FightingICE のゲーム画面

4. 準備

4.1. 決定木とランダムフォレスト

決定木とは，機械学習の一種であり，分類や回帰のルールを木で表現する．親から順番に条件分岐によって分割することで，同じような属性で構成されたグループを分類し，結果を予測する手法である．条件分岐に使う条件を説明変数という．情報利得IGが最大になるようにデータを分類する．情報利得IGは以下の式で表される．

$$IG(D_{\mathrm{p}}, f) = I(D_{\mathrm{p}}) - \frac{N_{\mathrm{left}}}{N_{\mathrm{p}}} I(D_{\mathrm{left}}) - \frac{N_{\mathrm{right}}}{N_{\mathrm{p}}} I(D_{\mathrm{right}})$$

D_{p}は親のデータセット，fは説明変数の数(特徴量)，Nはノード，jは注目しているデータ，mは木を分割するノード数を表わす．Iは不純度という指標であり，分類したグループに偏りがあるほど大きな値になる．データの分割はこの不純度をもとにして行う．不純度はエントロピーやジニ係数を用いて計算する．

不純度はジニ係数を用いて以下の式で計算する．

$$I_{\mathrm{G}}(t) = 1 - \sum_{i=1}^{c} p\,(i|t)^2$$

cはグループの数，tは現在のノード，iは各グループ，cはノードtに存在するグループの数，pは割合を表現する．特定のノードに存在するグループが 1 つだけの場合，不純度は 0 になる．

ランダムフォレスト(RF)とは決定木を大量に作って識別するものである [2]．決定木を複数組み合わせて，各決定木の予測結果を多数決するため，1 つの決定木のみで分類するよりも精度が高くなりやすい．

4.2. Q 学習

Q 学習とは強化学習の手法の一種である．ある状態において AI がある行動をとった時に報酬が与えられ，その報酬を最大化するために試行錯誤を重ねて学習する方法である．強化学習では，AI は即時に得られる目先の報酬ではなく，将来に得られる価値を最大化させるように学習する．そのために Q 値を試行錯誤によって更新していくことが必要である．Q 値は状態行動価値と呼ばれ，ある状態で，ある行動をとったときの価値のことである．Q 値は行動価値関数$Q(s,a)$で表される．sはその時の状態であり，aはその時にとることができる行動である．状態sにおける行動aが最適な行動ならば，$Q(s,a)$の値が高くなり，その行動が選択されやすくなる．AI が行動を行うたびに報酬rが得られる．報酬rに基づいて以下の式によって状態行動価値Q(s,a)を更新する．

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left\{r + \gamma \max_{a' \in A(s')} Q(s',a') - Q(s,a)\right\}$$

αは学習率と呼ばれ，Q 値をどれだけ大幅に更新するかを決める値である．$0 < \alpha \leq 1$の範囲で設定される．γは割引率と呼ばれ，将来の価値をどれだけ重視するかを決める値である．$0 < \gamma \leq 1$の範囲で，一般的に 0.90~0.99 程度にすることが多い．$\max_{a' \in A(s')} Q(s',a')$は行動$a'$の価値である．行動$a'$は集合$A(s')$の中で最大の価値を持っている．集合$A(s')$は状態$s$で行動$a$をとった後の状態$s'$のときにとることができる行動の集合である．

また，行動選択にはε-greedy 法がある．AI は Q 値の大きい行動を優先して選択するが，現在の Q 値が最も大きい行動が最適な行動でない場合，常にその行動が選択され続けてしまため，最適な行動を見つけられない．そこで，ε -greedy 法を用い，確率εで Q 値の大きい行動を選択し，確率$(1-\varepsilon)$でランダムに行動を選択する．

5. 提案手法

本研究では，Q 学習と RF を組み合わせた模倣 AI を提案する．最初に模倣対象者がランダムに行動する AI と戦い，状況データリストを記録する．AI はその状況データリストから Q 学習を用いて，ある状況での最適な行動を学習させる．また，状況データリストから最適な行動を予測する手段として，RF を用いた．RF によって，非観

測な状況でも最適な行動を予測できるため，学習精度が高まると考えられる．少ない状況データリストでも模倣できるということがこの手法の利点である．

5.1. 状況データ

状況データは，ある状況でプレイヤがどのような行動を選択したかを記録したものである．取得する情報は，自分の行動，自分の前回の行動，相手の行動，自分の HP，相手の HP，自分のエネルギー，相手のエネルギー，自分の Y 座標，相手の Y 座標，互いの距離，残り時間の 11 項目である．自分の行動以外の 10 項目がその時の状況であり，自分の行動がその状況で選択した行動である．

5.1.1. プレイヤの状況データリスト

状況データを毎フレーム記録して，時系列順に記録したものを状況データリストと呼ぶ．ここで，行動は 1 フレームでは終了せず，複数のフレームをかけて行われる．よって，行動を開始した 1 フレーム目とその次の 2 フレーム目での行動は等しい．しかし，2 フレーム目の行動は，1 フレーム目に選択した行動の途中の状態であるため，2 フレーム目は行動を選択したことにはならない．そのため，行動開始時以外の状況データは無視する．

5.1.2. 状況データの区分

HP などの連続値をそのまま記録すると，状態数が膨大になる．そのため，AI が状況データを扱う際，それぞれの項目をある範囲で区分する．また，プログラムで扱いやすくするために，行動には番号をつける．さらに，相手の攻撃を受けて倒れる(DOWN）や空中から着地した瞬間(LANDING）のような復帰を表わす行動は，自分の意志で行っていないため，受動的な行動である．このような受動的行動はすべて無視し，能動的な行動のみを対象とする．何もしない(STAND)は対象とする．

5.2. RF による行動予測

模倣対象者から取得した状況データリストから RF を用いて最適行動を予測する．状況データの中で予測する項目は自分の行動である．決定木は項目ごとに条件分岐を行い分割するが，行動の類は条件分岐ができない．そのため，相手の行動ごとに RF を分ける．つまり，ある相手の行動のとき，最適な自分の行動を予測する RF が生成される．相手の行動ごとに決定木を 10 本生成した RF を記録する．また，簡略化のために RF では自分の前回の行動を考慮しない．

5.3. 行動の Q 学習

状態sにおいて行動aを行い，状態s'に遷移したとき，その行動の価値を行動価値関数$Q(s,a)$で評価していく．状態sを状況データの状況，行動aを AI の行動とする．次に報酬を与える行動を決めるため，現在の状況での最適行動を決める．最適行動の決め方は，状況データリストから決める方法と，RF から決める方法の 2 種類がある．

まず，状況データリストから最適行動を決める方法を説明する．あるフレームのとき，状況データリストの中から，そのフレームでの AI の状況データに最も近い状況データを取得する．その取得した状況データでの自分の行動がそのフレームでの最適行動である．AI がそのフレームで選択した行動が，最適行動と一致していたならば，報酬を与える．取得した状況データが AI の状況データと大きく異なっていた場合，罰を与える．この方法でうまく学習を進めるためには多くの状況データリストが必要であるが，RF を用いれば問題ない．

次に RF から最適行動を決める方法を説明する．あるフレームでの AI の状況データから相手の行動に対応する RF を使用する．AI の状況データの行動以外の項目を使い，決定木によって分類することで行動を予測する．10 本の決定木で予測した行動から多数決をとることで最適な行動を決める．AI の状況データの自分の行動と最適な行動が一致するならば報酬を与える．

6. 実装

本研究では，FightingICE で動作する模倣 AI を Java 言語で実装した．3 種類の手法と 2 種類の状況データリストを用いて，表 1 のような計 6 種類の AI を実装した．

表 1　異なる 6 種類の模倣 AI

	模倣 AI	模倣者のプレイ回数	AI の学習試行回数
[a]	Q 学習+RF	100	1000
[b]	Q 学習+RF	50	500
[c]	Q 学習	100	1000
[d]	Q 学習	50	500
[e]	ルールベース	100	-
[f]	ルールベース	50	-

[a][b]は Q 学習と RF を組み合わせて学習させた AI である．[c][d]は Q 学習のみで学習させた AI である．[e][f]はルールベースで実装した AI である．模倣対象者は著者とし，ランダムに行動する AI と 100 ラウンド対戦を行った．その対戦の状況データリストを記録し，ファイルに出力した．ここで，状況データリストの量による AI の性能の違いを検証するため，半分の 50 ラウンド時点の状況データリストを別のファイルに出力した．この状況データリストを読み込み，相手の行動ごとの RF を作成し，ファイルに出力した．[a][b]は状況データリストと RF のファイルを，[c][d][e][f]は状況データリストのファイルのみを読み込ませた．また，Q 学習の試行回数は，100 ラウンドの状況データリストを用いる場合は 1000 回，50 ラウンドの場合は 500 回とした．学習率を 0.1，割引率を 0.9，ε-greedy 法のεを 0.3 とした．

7. 実験

7.1. 被験者と実験方法

被験者は，21～24 歳の男女 7 人に対して行った．異なる 6 種類の AI の対戦動画を被験者に見せ，アンケートをとる．アンケート項目は以下の 3 つである．

1. AI は対象者を模倣できていると感じたか
2. AI の行動は人間的だと感じたか
3. AI についてどんな印象を持ったか

1 と 2 の質問は 5 段階で評価してもらう(5 が最も良い評価)．3 は AI の動きの特徴について感じたことを記述してもらう．

7.2. 結果

アンケート結果を図 2 と図 3 に示す．実験の結果，QL と RL を組み合わせた AI がどちらも最も模倣できていた．人間的な行動に関しても高評価が得られた．QL のみの AI は，模倣に関してはどちらもまずまずの評価であるが，人間的な行動に関して，[d]は若干低評価が見られる．ルールベースの AI は，模倣に関して，[e]と[f]は状況データリストの量による違いが見られた．特に[f]は低評価であった．人間的な行動に関しては，どちらも低評価が多く見られた．

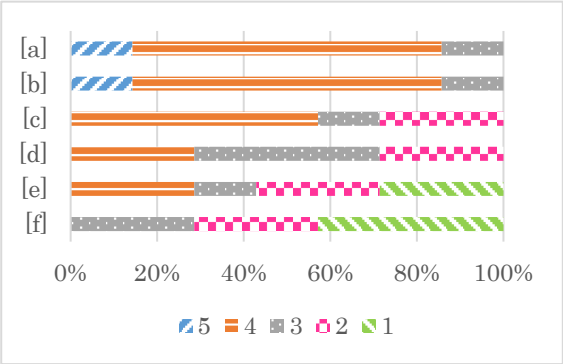

図 2　模倣できていると感じたか

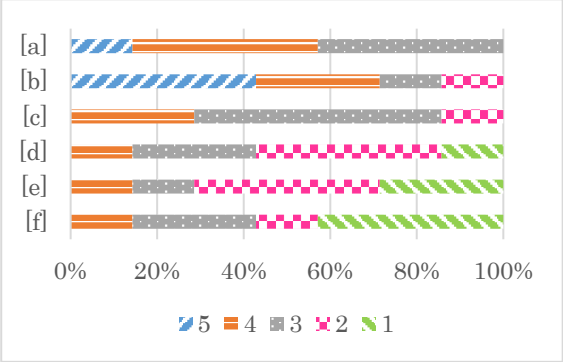

図 3　行動は人間的だと感じたか

8. 議論

Q 学習と RF を組み合わせた AI は，状況データリストと学習回数の違いによって，模倣の出来や，人間的な行動には差がなかった．これは，RF によって少ない状況データリストでも最適な行動が予測できているためと考えられる．よって，少ない学習回数で模倣できる．

Q 学習のみの AI は，状況データリストの量と学習回数によって，模倣に関して大差はなかったが，学習回数がさらに多くなれば，より模倣できるだろう．人間的な行動に関しては違いがあった．[d]は状況データリストが少ないため，非観測な状況では最適な行動がうまく見つけられない．そのため，数少ない行動に報酬が与えられ，動きが単調になっていると考えられる．

ルールベースの AI はどちらも行動が人間的ではなかった．これは，対象者の非観測な状況では，それに近い状況から行動を選択したため，行動に多様性がなく，単調であったと考えられる．

9. おわりに

本研究では，Q 学習と RF を組み合わせた模倣 AI を提案した．この手法の有効性を検証するために，Q 学習と RF を組み合わせて学習させた AI と Q 学習のみをさせた AI とルールベースの AI を比較する実験を行った．実験の結果，Q 学習と RF を組み合わせた AI が最も対象者を模倣できていた．さらに，状況データリストと学習の試行回数が少ない場合でも模倣できていたことから，RF が有効であることが分かった．

本研究では，Q 学習を行う際，すべての状況での Q 値を 0 にして行った．ここで，状況データリストで観測できた状況での行動の Q 値を予め高く設定すれば，学習速度が上がるのではないかと考える．また，学習率や割引率，ε-greedy 法のεを固定して行ったが，どのような値が最適であるか検証すべきである．また，状況データの区分を大まかに行ったが，どれほど細かく区分すべきか議論する余地がある．

文　献

[1] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, MIT Press, 2017.

[2] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5-32, 2001.

[3] F. Lu, K. Yamamoto, L. H. Nomura, S. Mizuno, Y. Lee and R. Thawonmas, "Fighting Game Artificial Intelligence Competition," *Proc. IEEE Conf. Consumer Electronics*, pp. 320-323, 2013.

[4] M. Bain and C. Sammut, "A framework for behavioral cloning," *Machine Intelligence*, vol. 15, pp. 103-129, 1995.

[5] 服部裕介, 田中彰人, 星野准一, "対戦型アクションゲームにおけるプレイヤの模倣行動の生成," 情報処理学会研究報告：ゲーム情報学, vol. 17, no. 20, pp. 1-8, 2007.

[6] 藤田肇, 石井信, "部分観測カードゲームのためのモデル同定型強化学習," 電子情報通信学会論文誌, vol. J88-D-2, no. 11, pp. 2277-2287, 2005.

[7] 藤井叙人, 佐藤祐一, 中嶌洋輔, 若間弘典, 風井浩志, 片寄晴弘, "生物学的制約の導入によるビデオゲームエージェントの「人間らしい」振舞いの自動獲得," 情報処理学会論文誌, vol. 55, no. 7, pp. 1655-1664, 2014.

[8] 浅山和宣, 森山甲一, 福井健一, 沼尾正行, "線形補外と k 近傍法を用いた格闘ゲームにおける敵の位置と行動の予測," 人工知能学会全国大会論文集, no. 1F2-4, pp. 1-4, 2015.

[9] 酒井賢人, 森山甲一, 武藤敦子, 犬塚信博, "決定木学習を利用した格闘ゲームにおける対戦相手の行動予測に基づく行動選択," 情報処理学会全国大会講演論文集, vol. 1, pp. 1015-1016, 2017.