The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

拡張されうる行動空間での特徴の表現学習を伴う価値関数の近似 ～“逆転オセロニア” を例に～

Approximation of Value Function with Feature Representation Learning to Deal with Extendable Action Space: Taking “Gyakuten Othellonia” as an Example

甲野 佑 [*1] 田中 一樹 [*1] 奥村 エルネスト 純 [*1]
Yu Kono Ikki Tanaka Jun Ernesto Okumura

[*1]株式会社 ディー・エヌ・エー
DeNA Co., Ltd.

In a general decision-making task, the options of action are expanded indefinitely due to the change of environment or the discovery of new action by agent. In a situation that the number of options increase, it is necessary for an agent to acquire an abstracted expression of actions autonomously. Here we propose a learning framework that solve this issue. In the proposed method, value function is approximated with embedded behavioral representations, which generalize the expression of actions, using state-tradition trajectories. We confirmed the efficiency of the framework using the mobile game “Gyakuten Othellonia”. This game is a mixture of board game and trading card game and characters are added to the environment frequently, which is a good testbed to realize expandable action space. Finally, we show that, with the proposed framework, an agent can learn character’s representation and utilize it to learn optimal strategies in the game.

1. はじめに

近年, 非常に取りうる状態数が多く, 様々な戦略を相手取る囲碁などの対戦ゲームにおいて木探索, 深層学習, 強化学習の知見の融合により人間のプロプレーヤーに勝る行動選択が可能になった [Silver, 2017]. このようなゲーム課題は運動制御における各関節の調整などの低次の行動を扱った学習と異なり, 行動の選択肢が記号化されている前提での高次の意思決定を扱っている. 既存の深層学習, 強化学習における行動の学習手法の多くは, 人工ニューラルネットワーク (NN) 等による関数近似が用いられるが, 出力はあらかじめ固定の行動選択肢数で学習される. しかし現実の離散的な高次の行動選択肢の量と質は, ある種のプランニングや経験によって新たに発見・変更されるものであり, 前述した形式では無際限に増加・変更されゆくケースに対応できない. そもそも潜在的な行動種類数が膨大である場合も出力が肥大化して学習が困難になる. クラスタリングにより膨大な数の行動を抽象化する手法は存在するが, 行動そのものに特徴が付加されていなければ対処できない.

そこで本研究では, 行動選択肢を状態遷移軌跡から任意の表現をベクトルに埋め込む表現学習手法を提案する. 本提案手法で前述した膨大かつ拡張されうる行動選択肢数が表現ベクトルとして有限長の連続値に変換されることで, 評価関数の関数近似器の入力として扱えるようになり, クラスタリング無しで行動数の膨大さと無際限な拡張に対処できるようになる. しかし行動選択肢の発見等は本来, 行動学習 (例えば強化学習) 以外のシステムによって提供されるため, ATARI など従来課題との対応付けが難しい. そこで本研究は “逆転オセロニア” という行動選択肢となるキャラクター (コマ) 種類数が増大することを前提とした, カードゲームとボードゲームの要素を組み合わせた課題において, 提案手法の有効性を示す.

2. 拡張意思決定課題の想定

“逆転オセロニア” は株式会社ディー・エヌ・エーがサービスを提供するオンラインゲームである. オセロと同様の法則下で設置可能なマスが決まる 6×6 の盤面にそれぞれ効果の異なるコマを交互に置いて, コマの設置効果により発生するダメージで相手の体力を早くゼロにした方が勝利するというのが “逆転オセロニア” の基本ルールになる. プレイヤーは自身が保有するコマの中から 16 種の組み合わせによってデッキと呼ばれるコマの集合を作成し, ゲームを開始する. ゲーム開始時に両ユーザーには, 16 コマの中から “リーダー” と呼ばれる属性に予め選んでおいた 1 コマと, 残り 15 コマからランダムに選ばれた 3 コマの計 4 コマが与えられ, その後もコマを設置したら自身のターン開始時に新たなコマが 1 つデッキから与えられる. コマの種類は 1,000 キャラ以上存在し, 度々新たにリリースされる. ルールの詳細は参考サイト [Othellonia 16][Othellonia wiki] を参照のこと. ゲーム課題としては以下の特徴を有する.

1. 2 人零和不完全情報ゲーム (本研究では課題の簡略化のため完全情報に変更)
2. ターン制かつ 1 ターンにつき 1 回の行動選択
3. 可能な行動の集合が現在の手札, 場のコマ配置で決定
4. 潜在的な行動選択肢数はコマ種類数 × マス数 (6×6) であり, コマ種類数の増加によって増えていく可能性がある
5. コマが盤面に留まるため, 全てのコマを離散的に定義すると, 行動種類数 (コマ種) の増加に対して指数的に状態空間も拡張される
6. 手札としてのコマの出現順番が確率的かつ予測不能であるため, 木探索が有効ではない

本研究ではある時点の状態で可能な行動集合は “逆転オセロニア” のゲームルールを提供するシミュレータ (環境) から付与されるとし, 問題の範囲を制限させている. “逆転オセロニア” 自体の複雑性の影響は大きいものの, 本研究は飽くまで潜在的な行動種類数が膨大かつ増大しうる前提での非常に複雑な行動選択への対処を目的としている.

3. 行動種類数が可変な行動評価値の近似

本研究においてエージェントは現在実行可能な行動集合の中から, 学習により付加した各々の行動評価値を参照して行動選択する. そしてある行動選択肢の評価値は関数近似器による近似を前提とする. 関数近似器は状態特徴 s, 行動特徴 a

連絡先: 甲野 佑, 株式会社ディー・エヌ・エー, 150–8510, 東京都渋谷区渋谷 2–11–1 渋谷ヒカリエ, yu.kono@dena.com

を入力とし，スカラー値である評価 $f(s,a;\theta)$ を出力とする．それらの関数近似は近似対象である行動の評価関数 y の変更により，y の予測器 f のパラメータ θ の学習方法は，教師あり学習 (Supervised Learning, SL)，強化学習 (Reinforcement Learning, RL) の双方に対応できる．本研究では行動評価値の関数近似器 $f(s,a;\theta)$ に任意の多層 NN を用いる．

3.1 ユーザーログに対する教師あり学習

“逆転オセロニア” は対戦ログを用いて，実際にユーザーがどの局面 (状態 s) で手札中のどのコマを，設置可能なマス中のどの座標に設置したか意味する行動 a の選択について，ユーザーログを用いた教師あり学習ができる．しかし潜在的な行動種類数が膨大かつ，新コマのリリースで増大する可能性があり (ゲームの性質上，伴って状態空間も指数的に拡張される)，離散的に定義された全行動種類 A_{all} ($|A_{\mathrm{all}}|$ = コマ種類数 × マス数) に対する行動選択確率 $P_{\mathrm{all}}(a \in A_{\mathrm{all}}|s)$ は計算不可能である．

そこで本研究では，ある状態において選択可能な任意の行動が選ばれた (put) か否か $(\mathrm{not\ put})$ の可能性 $P_{\mathrm{put}}(y \in \{\mathrm{put},\ \mathrm{not\ put}\}|a,s)$ をユーザーの主観的な行動評価値として学習する．これは入力である状態特徴 s，行動特徴 a が実際にユーザーログ上で選ばれたか否かの 2 値分類タスクになるため，近似関数の出力 $y^{\mathrm{pred}} = f_{\mathrm{put}}(s,a;\theta)$ を棋譜上の教師信号 y^{user} に対する cross–entropy 損失関数で学習可能になる．また 2 値分類であるため，手札中の選択されなかったコマ，設置可能だが選択されなかったマスも学習データに用いる．

3.2 ゲームシミュレータに対する強化学習

本研究では “逆転オセロニア” のゲームシミュレータを用いて行動選択の強化学習も行った．近似関数器 $f(s,a;\theta)$ の近似対象として報酬 r_t の累積値である収益の予測値である行動価値関数 $Q(s,a)$ を学習する．行動価値関数の近似関数 $f_{\mathrm{Q}}(s,a;\theta)$ は Double Q–Learning [VanHasselt 15] を用いた TD 学習の更新則に従い，ある時刻 t に対する当該の状態行動対 (s_t,a_t) については以下の損失関数 (2) を最小化するよう Prioritized Experience Replay (PER)[Schaul 15] を用いて学習した．

$$a_{\mathrm{back}} \leftarrow \underset{a_i \in A_{t+1}}{\arg\max}\, f_{\mathrm{Q}}(s_{t+1},a_i;\theta) \tag{1}$$

$$\mathcal{L}_{\mathrm{Q}} \leftarrow \left(r_{t+1} + \gamma f_{\mathrm{Q}}(s_{t+1},a_{\mathrm{back}};\theta^-) - f_{\mathrm{Q}}(s_t,a_t;\theta)\right)^2 \tag{2}$$

ここで A_t, $r_t = \{0,1\}$ はそれぞれ時刻 t で実行可能な行動集合と得られた報酬を意味し，本課題で報酬はゲームの勝敗が決した時のみ与えられる (勝利時に $r_t = 1$, 敗北時に $r_t = 0$)．またターゲットパラメータ θ^- はそれ以前の近似関数 f_{Q} のパラメータ θ であり，一定間隔の間 θ^- を固定することで関数の更新を安定させる Target–net という手法である．DQN [Mnih 15] などの手法と異なり，近似関数 $f_{\mathrm{Q}}(s_t,a_t;\theta)$ の出力は常に一つのスカラー値であることに留意されたい．

3.3 行動特徴

本研究で近似関数 $f(s,a;\theta)$ に入力される行動特徴を定義する．多くのラジオゲームにおける行動は離散的に定義されたコントローラのスイッチ種類数に対応する on/off のバイナリベクトルで定義できる．“逆転オセロニア” 等のカードゲーム要素を持つ場合は潜在的に存在して選択されうるコマ (カード，コマンド) が離散的に定義された行動に該当し，コマ種類数の長さを有する one-hot ベクトルで行動特徴が定義される．“逆転オセロニア” の場合，盤面空間 6×6 の one-hot 行列，コマ毎が当該の座標に置くことで効果 (スキル・コンボ，詳細は参考サイト [Othellonia 16][Othellonia wiki] を参照のこと) が発動するかどうかの 2 値フラグ，そのコマの元々の攻撃力なども行動特徴に該当する．盤面座標，効果フラグや攻撃力は拡張される概念ではなく，それ自体が相対的な特徴量であるため，後述する表現学習の対象や表現ベクトルへの置換対象は，この中で非相対的な情報であるコマ種ベクトルに限る．

3.4 実ゲームにおける行動選択の方法

ゲームのプレイ時の行動はある時点での状態 s_t において実行可能な行動集合 A_t の要素を全て評価し，その中で最も評価値の高い行動選択肢 $(\arg\max_{a_i \in A_t} f(s_t,a_i;\theta))$ が選ばれる．それ以外にも行動選択には強化学習時の探索や，評価時に行動バリエーションのため $f(s,a;\theta)$ に基づく softmx 関数や，ϵ-greedy 手法などで確率性を与える場合もある．

4. 部分遷移要因の表現学習

本研究では自然言語処理技術を参考に，近似関数 $f(s,a;\theta)$ の入力であり，拡張されうる行動特徴 a を固定長の実数値ベクトルとして暗黙的に学習する手法を提案する．具体的には話者特徴を表現ベクトルとして埋め込むペルソナモデル [Le 16] を元に，状態行動対中の離散的な行動要素等，ある部分集合へ状態遷移の要因を表現ベクトルとして埋め込む形式に考案した．

4.1 学習方法

ある時点での状態 s_t とその次状態 s_{t+1} が与えられたとき，その状態遷移軌跡が状態行動対特徴 $X_t = (s_t,a_t)$ 中の任意の部分集合 x_t^{tar} とそれ以外の要素 x_t^- に引き起こされたと仮定する．この任意の要素 x_t^{tar} を本研究では部分遷移要因と呼ぶ．そして状態行動対 X_t から部分遷移要因 x_t^{tar} をマスクされた入力信号 X_t^- を定義する．議論の簡略化のため部分遷移要因をその時点での行動 $(x_t^{\mathrm{tar}} = a_t)$ だとした場合，いずれの要素であるか特定不明という意味の値 x^{u} でマスク $(a_t \leftarrow x^{\mathrm{u}})$ された入力信号は $X_t^- = (s_t, x^{\mathrm{u}})$ と表すことができる．このマスクされた入力信号 X_t^- を用いて次状態 s_{t+1} の予測を任意の予測器 f_{rep} で学習すると，部分遷移要因 x_t^{tar} に引き起こされる状態遷移の成分が行き場を失う．

そこで入力としてマスクされた入力信号 X_t^- の他に x_t^{tar} に対応するランダムに初期化されたユニークな特徴変数ベクトル c_t^{tar} を与えて，次状態 s_{t+1} の予測学習を行わせる．これを多くの状態行動対 X に対して行うことで同様，あるいは類似した幾つかのマスクされた入力 X_t^- に，それぞれ異なる結果 s_{t+1} が起きたとき，その差異の要因となる成分が特徴ベクトル c_t^{tar} に吸収されて，部分遷移要因 x_t^{tar} の状態遷移に寄与する表現が暗黙的に学習されていく．部分遷移要因が行動要素である場合は行動 a_t の入力表現 $a_t^{\mathrm{rep}} = c_t^{\mathrm{tar}}$ あるいはその要素の一部に関する表現ベクトルとして行動の学習に再利用される．

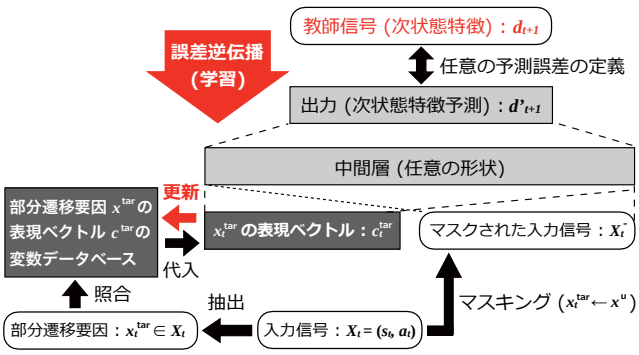

図 1: 部分遷移要因の表現学習アーキテクチャ

学習に用いる状態行動対 X_t と次状態 s_{t+1} の状態遷移軌跡は環境モデル上でありえる遷移であれば良い. しかし良い行動の評価関数を作るため, 実用上は一定のリテラシーを持ったエージェント, あるいはユーザーログから得た状態遷移軌跡が望ましいと考えられる.

4.2 次状態の代替えと損失関数

現実的には s_{t+1} の直接的な予測は困難である. 特に “逆転オセロニア” は行動特徴が手札や盤面上の情報として状態特徴にも現れるため, 長大で延長を前提としたコマ種の one-hot ベクトルを含む s_{t+1} の推定は不可能である. そこで十分に次状態 s_{t+1} の代替えとなりうる, 状態要素の部分集合や, 行動 a_t を実行した後に発生する観測可能な事象を教師信号とすることも想定する. 次状態信号そのもの, その代替信号も含めて本表現学習手法の教師信号を次状態特徴 d_{t+1} と呼ぶ. この次状態特徴とその損失関数の設計は次状態の予測の困難さや観測しうる特徴量等に依存するため, 損失関数であれば連続量は正規化して回帰, 排中律を有する離散事象の集合部のみ cross–entropy にするなど, ゲームドメイン毎に行う必要がある. しかしラジオゲーム (画像・動画から) の行動選択の学習に本手法を用いる場合, 次状態である画面の各ピクセルごとの各要素の発生確率の予測誤差を最小化するように学習することが想定される. 以上に記した表現学習器の基本的な構造は図 1 に示す.

4.3 利点

以下は表現ベクトルの事前学習による利点の推定である.

1. 種類数が潜在的に膨大, あるいは増大しうる行動要素表現の削減と固定長化
2. 利点 1 に伴うユニット数の節約と学習時間の削減
3. 表現空間上での類似行動の汎化による学習の効率化
4. 利点 3 に伴ってあまり選択されていない類似行動コマの使い方を習得可能
5. 行動の使い方のような概念＝表現ベクトルを事前学習できるため, 行動評価の関数近似時のハイパーパラメータ等の試行錯誤が比較的容易になる

根本的に扱えなかった行動概念に対処できる利点 1 が本研究の最大の効果である. 利点 2, 4 は仮説であり, 困難ではあるが実験により幾分か定量化できる. しかし利点 5 は実運用上の恩恵であり, 定量化はできない. 前述した通り, “逆転オセロニア” はコマの増加により状態空間が指数的に広がるため, 利点 2 の効果が大きいと思われる. また様々なコマを学習する都合上, 利点 4 は有益な性質であるが本研究では検証していない.

5. 実験

本研究では利点 2 の仮説について, それぞれ表現学習と教師あり学習, 強化学習による行動評価値の学習とを組み合わせた際の実験により定量評価を試みた.

5.1 逆転オセロニアでの行動学習の共有設定

後述する実験 1, 2 ともに行動特徴の中の “膨大かつ拡張されうる離散的要素” であるコマ特徴の表現ベクトルは共通のものを用いた. その表現学習には 2017 年 1 月に使用できた 916 コマ種 (+ キャラクタの付加されていない無地ゴマ 1 種) を対象に, ユーザーランク 76〜200 同士の対人戦データから得られる状態遷移軌跡を用いた. また実験 1 の教師あり学習, 実験 2 の強化学習には同様の中間層の構造を用いる. 各種学習器の構造やハイパーパラメータは表 1 を参照のこと.

表 1: 各種パラメータ

パラメータ名	表現学習	行動学習
学習率	1.0×10^{-3}	1.0×10^{-5}(SL, RL)
minibatch size	2^{10}	2^{10}(SL), 2^5(RL)
Dropout 率	0.0	0.0(SL), 0.0(RL)
L_2 正則化係数	1.0×10^{-8}	1.0×10^{-8}(SL), 0.0(RL)
入力 Unit 数	55663	実験毎に変わる
中間 Unit 数	(2^{12}, 2^{11}, 2^{10})	
出力 Unit 数	63	1
活性化関数	ELU (出力層以外)	
Optimizer	Adam	
表現ベクトル	キャラ数 : $916 + 1 \rightarrow$ ベクトル長 : 30	

5.1.1 入力特徴

各学習器の入力特徴は状態 s_t としてターン数や選択側の色 (白・黒), 自分・相手の残り体力, 手札, デッキ, 盤面などを, 行動 a_t には任意の選択可能なコマや設置可能なマス座標, スキルやコンボ等の特殊効果の発動の可否を用いた. 表現ベクトルを使用しない場合は one-hot ベクトルを手札, デッキ, 盤面のコマの表現として用い, コマの表現ベクトルを使用する場合は, その全てを上述したベクトル長 30 の表現ベクトルに置き換えた. そのため両実験とも比較対象である表現ベクトルを使うか否かで第一層のみ入力数とパラメータの数が異なる. ターン数を対数にした値など入力の特徴量エンジニアリングも行われているが, 入力特徴や表現学習時の教師信号, 損失関数は実サービスのゲームを用いている都合上, 詳細な言及は避ける.

5.1.2 勝率の定義

“逆転オセロニア” では非対人対戦イベント (クエスト等) や通信が切れた際の代打ちとして, ルールベース AI が実装されており, 強さが固定であることから本研究の勝率の定義にはルールベース AI との戦績を用いた. 実験 2 では勝率のベースラインとして学習エージェントと同デッキでの, ランダムな意思決定でルールベースと戦った場合, ルールベース同士で戦った場合の勝率を示す. 勝率は, 各試合で各々異なるシードでデッキのシャッフルと先攻後攻の決定した 1,000 試合中何勝したかで評価し, 試合時には学習された行動評価の近似関数の出力に対して greedy な行動選択が行われる.

5.2 実験 1: 教師あり学習と表現ベクトル

実験 1 では利点 2, 表現ベクトルを用いた場合とそうでない場合での学習効率の程度を示す. ここでいう効率とは計算時間に対する勝率の向上速度や, 最終的な到達勝率の高さを意味する. 勝率は様々なデッキの組み合わせによって測るべきだが, 現実的にあらゆるデッキの組み合わせで評価するのは困難であるため, ここでは代表として 2017 年 1 月の時点でよく使われていたデッキバリエーションである 4 種を用いた. 限定されたコマ種類数での勝率評価であるため, 表現ベクトルの有無で大きな差が現れないことが予想される. そのため学習回数に対する勝率以外に, 同条件で学習に掛かった経過時間を提示する.

5.2.1 実験設定

学習には 2017 年 1 月に集計されたユーザーランクが 76〜200 同士の対戦ログを使用した. 勝率評価に 4 種の内訳はデッキ内のコマの属性を, 竜 (攻撃力が高い), 神 (耐久値が高い), 魔 (トリッキーな戦術) で構成した 3 種とそのバランス的な組み合わせを用いた. 学習と評価に使用したデッキの構築はアソシエーション分析と階層的クラスタリング手法の一種であるウォード法と k–means 法を組み合わせたクラスタリングに

表 2: 表現ベクトル ＋ 教師あり学習の勝率推移と経過時間

	表現ベクトルあり		表現ベクトルなし	
Step 数	勝率 (%)	時間 (min)	勝率 (%)	時間 (min)
1×10	23.72	1	11.60	1
1×10^2	42.59	26	59.85	27
1×10^3	77.16	47	77.60	73
1×10^4	82.97	92	82.27	264
1×10^5	85.23	406	85.28	1,947
2×10^5	85.56	746	85.64	3,808
3×10^5	85.06	1,103	85.96	6,623
4×10^5	85.24	1,444	86.32	8,426
5×10^5	85.78	1,839	86.20	10,242

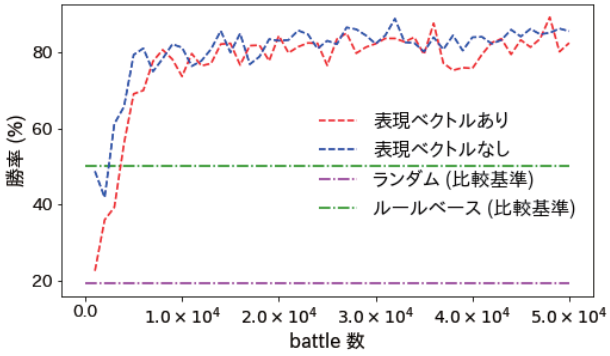

図 2: 表現ベクトル ＋ 強化学習モデルの勝率推移

より抽出した頻出するコマの組み合わせから, 任意の組み合わせによるデッキを自動生成した. 行動学習ネットワークの表現ベクトルを使用する場合の入力サイズは 5,649 になった. それに対して表現ベクトルを使用しない場合, 2017 年 1 月時点で使用されたコマ種類数 (916 種 ＋ 無地コマ 1 種) を直接扱えなければならないため 74,570 にもなり, 表現ベクトルを使用した場合の約 13 倍となった.

5.2.2 結果及び考察

表 2 に各 step （minibatch で学習した回数） での勝率と, 同条件の GPU で学習させた場合の経過時間を示す. 毎 step の勝率はほぼ等しいが, 50 万 step 時の経過時間が約 5.6 倍になった. これは 916 のコマ種類数を想定したものであり, 現在はさらにコマ種類数が増加しており, 学習コストはコマの増加に伴い更に大きくなる. 本研究では生成されたメジャーなデッキ構成を用いたため, マイナーなコマの学習などに影響を評価できていないものの, コマ表現ベクトルが計算的な時間削減に寄与し, 成績に影響を及ぼさない示唆が得られた.

5.3 実験 2: 強化学習と表現ベクトル

強化学習でも表現ベクトルの使用に対して成績に変化が表れるか実験を行なった. 教師あり学習と異なり, 自・敵デッキは単一デッキに固定して強化学習を行なった.

5.3.1 実験設定

デッキ構成は実験 1 と同じ手法で自動生成した. 対戦相手には初期 1,000 対戦はランダムで, その後 1,000 対戦毎に保存される過去の近似関数を対戦毎にランダムに読み込み, 対戦相手の行動選択に用いた. アーキテクチャには表現ベクトルのあり, なしをぞれぞれ学習し, 勝率を比較した. 実験 1 の教師あり学習と同じく, 916 コマ種 ＋ 無地駒 1 種について事前に学習したベクトル長 30 であるため表現ベクトルを使用する場合の近似関数の入力サイズは 5,649 になった. 対して表現ベクトルを使用しない場合, コマ 16 種 ＋ 無地駒 1 種で計 18 の長さの one-hot ベクトルを使い, 入力サイズは (16 種しか使用しないゆえに) 3964 になった . 学習は PER により抽出された minibatch を 1 学習として 2 対戦ごとに 32 回学習を繰り返し行った. この時に損失関数に用いる割引率 $\gamma = 1.0$ で Replay Buffer のサイズは 1 回の状態行動対を一つの単位として 2,000,000 とし, Target-net は 4,000 回の学習ごとに同期した.

5.3.2 結果及び考察

図 2 に battle 数に対する勝率の推移を示す. 毎 battle の勝率はほぼ等しく, 強化学習でも表現ベクトルの使用により, 成績に悪影響を及ぼさない示唆が得られた. ただし現時点では単一のデッキに対する結果でしかなく, 今後の大規模で多種なデッキに対する検証を必要とする.

6. 結論

コレクションカードゲームなど, 離散化された潜在的な行動選択肢がトランプの枚数とは比較にならない種類数で存在する意思決定課題に対処するためには, 行動概念の汎化が重要になる. 本研究では提案手法により事前に獲得した表現ベクトルを用いる事でクラスタリングなしで拡張されうる膨大な行動を扱える事を示した. 表現ベクトルを使用する際の他の利点に関する検証, 定量化や, 提案アーキテクチャ構造を前提とした学習の効率化は必要であるが, 本研究により複雑なゲームへの機械学習, 強化学習の応用範囲を広めることができたと言える. 今後, 過去の経験や知識に基づく低次の行動から高次行動の発見と汎化が前提となることが予想される. その場合, 本研究のように拡張されうる行動空間を扱えることが重要になると考えられる.

参考文献

[Othellonia 16] 株式会社ディー・エヌ・エー: 逆転オセロニア. https://www.othellonia.com/

[Othellonia wiki] 株式会社ディー・エヌ・エー: 逆転オセロニア最速攻略 wiki. https://オセロニア攻略.gamematome.jp/

[Mnih 15] Mnih, V., Kavukcuoglu, K., Silver, D., Hassabis, D., et al.: Human-level control through deep reinforcement learning. Nature, 518(7540), p.529-533. (2015).

[VanHasselt 15] Van Hasselt, H., Guez, A. and Silver, D.: Deep reinforcement learning with double q-learning. arXiv preprint arXiv:1509.06461. (2015).

[Schaul 15] Schaul, T., Quan, J., Antonoglou, I. and Silver, D.: Prioritized experience replay. arXiv preprint arXiv:1511.05952. (2015).

[Silver, 2017] Silver, D., Hassabis, D., et al.: Mastering the game of go without human knowledge. Nature 550(7676), pp.354–359. (2017).

[Le 16] Li, J., Galley, M. Brockett, C., Gao,J. and Dolan. B.: A persona-based neural conversation model. ACL2016. (2016).

[濱田 17] 濱田 晃一, 藤川 和樹, 小林 颯介, 菊池 悠太, 海野 裕也, 土田 正明: 対話返答生成における個性の追加反映, 研究報告自然言語処理（NL）, 2017-NL-232(12), p.1–7, 2188-8779 (2017)