6P-02

強化学習の価値関数近似器として SDNN を用いた格闘ゲーム AI

小川 拓実† 阿久津 光範† 金 致中† 山根 健†
帝京大学大学院理工学研究科†

1. はじめに

ビデオゲームにおいて，人が操作しない自律的なキャラクタ（以下，エージェント）は重要な要素であり，その行動設計が課題である．特に，格闘ゲームでは強さの適切な設定によりプレイヤが楽しさを感じる[1]．しかし，従来方法（例えば[2]など）では，強い相手への対応能力や計算効率などにおいて課題が残っていた．

試行錯誤的に行動学習する強化学習に関して，新保らは選択的不感化ニューラルネット（以下，SDNN）を価値関数近似器として用いて効率的に学習できることを示した[3]．本方法を用いることで，格闘ゲームにおいても相手の戦略に素早く適応するエージェントを開発できる可能性がある．しかし，格闘ゲームのような多数の行動を扱うタスクへの適用例はない．

そこで本研究では，分散表現を用いて多数の行動を扱えるように新保らの方法を拡張して，強化学習を用いて格闘ゲームをプレイするエージェントを設計する方法を提案する．また，実際にエージェントを構築して性能を評価する．

2. エージェントの構成方法

2.1 出力層における行動価値の表現

提案エージェントの構成を図 1 に示す．SDNN を用いた行動価値関数近似[3,4]では，ある時刻 t において入力層に状態 s_t が入力されると，出力層に行動 a_t に対応した行動価値 Q_t が分散表現される．本研究では，多数の離散的な行動に対応できるように出力層にすべての行動 a に対する Q を分散表現することを考える．

最も単純な方法は，出力層の n 個の素子を行動数 m で分割して，それぞれの行動価値を n/m 個の素子で表現する方法である．また，別法として，行動に関する情報を入力層に他の状態変数と同様に入力する，あるいは中間層に多重不感化によって修飾させる方法が考えられる．しかし，これらには，Q 値の分解能，計算量，行動方向の汎化性能などの点から課題がある．

そこで本研究では，図 2 に示す方法を採用する．出力層の n 個の素子から行動毎に k 個を選択して

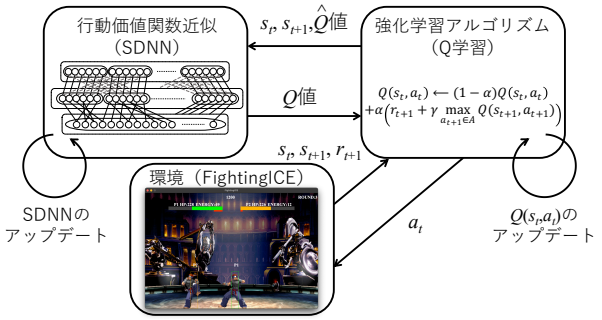

図 1 提案エージェントの構成

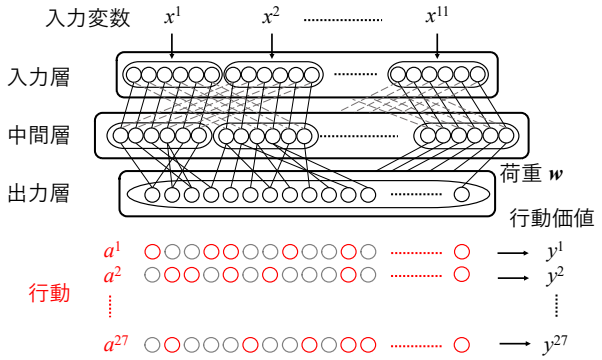

図 2 多数の行動に対応する方法

行動価値を表現する（ただし，$n>k$）．つまり，k 個の素子を選択するパターンが行動情報を表す．これにより，行動毎に異なる価値を表現でき，選択される素子の一部が重なるため，行動方向でも分散表現の利点を活かせる可能性がある．

行動間には何かしらの類似性があると考えられるので，発展的には類似する行動の表現において重なりを大きくするなど選択方法を工夫できる．ただし，これは今後の課題とする．

2.2 提案エージェントにおける強化学習

提案エージェントでは，強化学習のアルゴリズムとして Q 学習を用いる．そして，新しい Q_t が計算される毎に，状態 s_t に対する Q_t の対応関係を SDNN において 1 回だけ誤り訂正学習する．状態 s_t については，相手との距離，自分の行動，自分の HP など入力変数 x^1〜x^{11} で構成する．

行動に関しては，移動系 8 種類，防御系 2 種類，攻撃系 17 種類，合計 27 種類の行動 a^1〜a^{27}（対応する行動価値 y^1〜y^{27}）とする．ただし，ジャンプが伴う AIR 系については明示的には扱わない．

強化学習においては，報酬関数の設計が重要である．本研究では，行動選択したフレームから，行動が実行あるいはキャンセルされて次の行動を受け付け可能になるフレームまでの自分

Fighting Game AI Using SDNN as a Value Function Approximator for Reinforcement Learning
† Takumi Ogawa, Mitsunori Akutsu, Zhizhong Jin, Ken Yamane, Graduate School of Science and Engineering, Teikyo Univesity

の HP の変化量と相手の HP の変化量に注目して，**報酬 = (自分の HP の変化量) - (相手の HP の変化量)** $+\delta$ として，次の行動を選択する直前に報酬を与える．行動から報酬までの時間が変動するため，一般的に難しい学習となる．なお，δ については，予め設定した望ましくない行動の時に小さな負の報酬が入る．また，ラウンドの勝利時に+100，敗北時に-100 の値を加算する．

行動選択方法として，ε-Greedy 法を採用する．ランダム行動率を ε として，過去 100 ラウンドの勝率が 0.3 以下であれば $\varepsilon = 0.1$，そうではなくて 0.5 以下であれば $\varepsilon = 0.05$，それ以外は $\varepsilon = 0.0$ とする．また，強化学習のパラメータである学習率 $\alpha = 0.1$，割引率 $\gamma = 0.9$ とする．

複雑なタスクにおいて知識がない状態から強化学習すると，正の報酬を得る経験ができず学習が進みにくい．これに関して，生物では学習率を適切に変化させている可能性が示唆されている[5]．そこで，ラウンド終了時の累積報酬が負である場合，次のラウンドにおいて，正の報酬に対しては勝率に応じて 0.1～0.9 の間で α を変化させることとした．しかし，一般的に，このような操作により学習が不安定になりやすい．

3. 実験

提案方法の性能を調べるため，対戦型格闘ゲーム FightingICE（https://www.ice.ci.ritsumei.ac.jp/~ftgaic/）の実行環境を用いてサンプルエージェント 6 体と対戦した．総当たりによる結果，強い方から順に，ReiwaThunder > Dora > Toothless > JayBot_GM > BCP > TOVOR であった．

SDNN のハイパパラメータについて，入力層の素子数 11×80 個，中間層の素子数 11×10×80 個，出力層の素子数 $n = 4000$ 個，$k = 1000$ 個とした．また，図 1 左上の価値関数近似器のみルックアップテーブル（LUT）を用いるエージェントを比較モデルとして用意した．

実験の結果，提案エージェントでは実時間で安定して学習することができ，すべての相手に対して効率よく平均勝率を上げることができた．6 体のエージェントの中で最も強い ReiwaThunder との対戦における学習過程を図 3 に示す．なお，図中の「完全」とは 1 ラウンド 60 秒以内にどちらかの HP が 0 になり勝敗が決した場合の割合を表す．初期状態から 222 ラウンドで平均勝率が 0.5，1147 ラウンドで 0.9 に到達した．これに対して，比較モデルの結果を図 4 に示す．ラウンドを重ねても勝率が上がらずに 0.1 未満であった．

これらの結果から，提案エージェントは簡単に構成でき，様々な相手に素早く学習して対応できる大きな可能性があることがわかった．

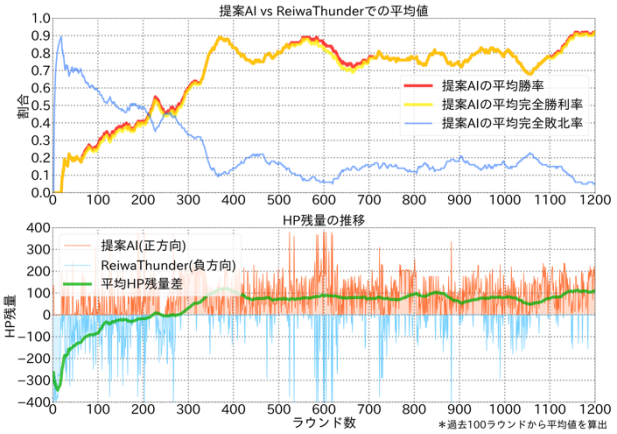

図 3 提案エージェントの学習過程

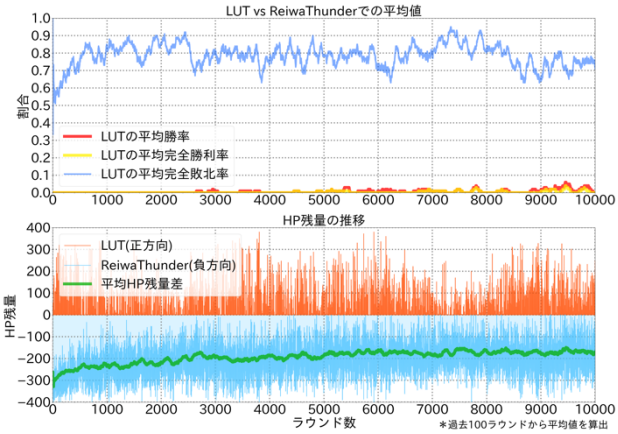

図 4 テーブルを用いた学習過程（比較モデル）

4. まとめ

本研究では，多数の行動を扱えるよう SDNN を拡張して格闘ゲーム AI を構築した．その結果，相手に素早く対応でき，学習収束時の勝率も高かった．今後の課題として，詳しい性能評価，複数の相手や未知の相手への対応などがある．

参考文献

[1] 石原誠ら，“対戦格闘ゲームにおけるゲーム AI や操作法の違いがプレイヤーの感じる面白さに与える影響の分析，”情報処理学会論文誌，vol.57，no.11，pp.2414-2425，2016.

[2] 邓士达ら，“動的な難易度調整により対戦して楽しい格闘ゲーム AI，”ゲームプログラミングワークショップ 2020 論文集，pp.58-61，2020.

[3] 新保智之ら，“選択的不感化ニューラルネットを用いた強化学習の価値関数近似，”信学誌 D，vol.J93-D，No.6，pp.837-847，2010.

[4] 小林高彰ら，“選択的不感化ニューラルネットを用いた連続状態行動空間における Q 学習，”信学誌 D，vol.J98-D，no.2，pp.287-299，2015.

[5] Ohta，H. et al.，“The asymmetric learning rates of murine exploratory behavior in sparse reward environments，” Neural Netw.，143，pp.218-229，2022.