ISS-P-52

TD 学習を用いたテトリス解法アルゴリズム

中山 亮士† 平原 誠†
† 法政大学大学院理工学研究科応用情報工学専攻

1 はじめに

テトリスとは, 無駄なスペースを発生させず 2 次元平面上にブロックを効率よく配置させていくゲームと捉えることができる. この最適化問題を解決することにより, 多様な場面での最適化問題に応用が可能であり, 幅広く展開が可能であると考えている. その例として, 駐車場の車両配置の最適化やトラックの荷物詰め込み作業の効率化等が挙げられる.

本研究の目的は，TD 学習を用いたアルゴリズムによる新たなテトリス解法を提案することにある.

2 従来研究

従来研究として，GA とニューラルネットワーク（以下 NN）を組み合わせたものがある[1]. NN の入力には盤面の特徴をとらえた特徴量 8 個を使用している. 特徴量のいくつかとして, テトリスフィールド内にある上得共にブロックに挟まれた穴の数や, 穴の上に詰まれているブロックの数等がある. NN の出力はテトリスの盤面の評価値である. また通常の NN とは異なり，1 つの NN を 1 個体とし，GA によって NN の結合荷重を最適化する手法を取っている. 個体の適応度は, 個体によって表現される評価関数を用いて 100 回のテトリスゲームを行った時の平均消去段数としている. 世代数 500，個体数 100 とし，淘汰，交叉，突然変異を繰り返し，最適化を行ったところ，結果として，約 6000 万消去段数の性能を示した. 単純計算で 5 百万回のテトリスゲームの試行を行わなければならず，学習時間が長いのが欠点といえる.

3 TD(λ)と NN を用いた提案手法

本研究では先行研究[1]よりも学習時間を短縮したアルゴリズムを目指す. そのために，強化学習の 1 手法である TD 学習と NN を用いて, オンライン学習により最適化を行う.

3.1 TD 学習

強化学習とは，探索を行うエージェントが方策（与えられた情報下での行動選択）と価値関数（その時の状態や行動の価値を求める関数）をもとに，エージェントが得られる報酬を最大にする手法である. TD 法では最終結果を待たずに, 他の状態での価値関数の結果をもとに価値関数を変更することができる. また，一定ステップ先の学習結果を価値関数の推定値の更新として利用できるアルゴリズムは TD(λ)と呼ばれる. λ=0 であれば即時報酬のみを得るために動き，λ=1 であれば，最終結果までの報酬を考えたアルゴリズムとなる(0≦λ<1). TD 学習は Temporal Difference Learning（誤差伝搬学習）の名の通り，n ステップ先の価値と現在の価値の誤差を計算し，未来から現在の価値を伝搬させるようなアルゴリズムとなっている.

3.2 実験

本研究で使用する NN の入力は，従来研究[1]で使用された 6 つの盤面の特徴量と, 筆者自身が導入した盤面の特徴量の 9 つを合わせた入力層ニューロン 9 個を採用している. さらに TD-Gamon[2]を参考に，盤面を直接的にとらえた入力（フィールド横 10 マス×縦 20 マス計 200）を採用し，合計 209 個の入力ニューロンを作成した. また，中間層ニューロンは 50 個，出力層ニューロンは 1 個としている. この NN の最適化に TD(λ)学習を用いる.

通常 TD 学習では報酬を最大にすることを目的とするが, 本研究では盤面のゲームオーバーになる可能性を算出し, 値が最も低くなった盤面にテトリミノを配置する. ゲームオーバーになった盤面の価値を 1 とし，その一手前に伝搬し続けることにより，常にゲームオーバーになりにくい盤面を作成できるのではないかと考えた.

4 今後の予定

今現在では望ましい学習結果は出せていない. おそらくテトリスの盤面では状態数が非常に大きく，NN の結合荷重が収束していないのが原因と考えられる. また，λの値や中間層ニューロンの数，入力に用いる盤面の特徴やその他のパラメータの値が重要となるため，その最適な値を，実験を重ねて発見しなければならない.

参考文献

[1] 荒川正幹，宮崎真奈実(2012) ，ニューラルネットワークと遺伝的アルゴリズムを用いたテトリスコントローラの開発, 情報処理学会　第 74 回全国大会講演論文，pp. 539-540.

[2] Gerald Tesauro(1995) ,Temporal Difference Learning and TD-Gammon ,Communication of the ACM,vol.38,pp.58-67.