TD 学習による 2048 プレイヤの評価値の分析

1200331　立石　孝幸　【 高度プログラミング研究室 】

1　はじめに

「2048」は G.Cirulli が 2014 年に公開した不完全情報ゲームである．このゲームで高得点を出すための手法の 1 つに,N タプルネットワークを利用した,TD 学習 (Temporal Difference Learning) が先行研究として存在する [1]．この手法によって,100 億盤面学習させたプレイヤが, 平均的に 14 万点以上を出せるようになった [2][1]．しかし,10 億盤面学習した際の平均点は 9 万点前後であるため, 学習した量に対して得点が伸びていないと言える．

そこで本稿では,TD 学習に利用される [1][2] ことの多い N タプルネットワークから構成された盤面評価値を解析することで, 学習後の動きに明確な誤りがないかを検証する.「2048」のゲームの性質上, 盤面上の最大値は上下左右の 4 隅に置いておくことが, 高得点を取るための基本方針である．そのため,「4096」以上の値を, 端以外の場所に動かす場合があるのかを検証する．

2　既存手法

2.1　N タプルネットワーク

「2048」における盤面の評価手法として,N タプルを用いたものがある．これは, 盤面の一部分 (N タイル分) の値の組み合わせから, 盤面全体の評価値とする手法である．この評価値が特徴量として, 評価値テーブルに格納される．この評価値は, ゲームが終了したときの最終得点の予測値となる [1]．

2.2　TD 学習

TD 学習は, 現在時刻を t としたアクション前の状態 S'_{t-1} の評価値を $V(S'_{t-1})$ とした盤面から, アクション後の盤面 S'_t の評価値 $V(S'_t)$ とアクション後の報酬 r_t を含めた値の差を用いて, 正しい盤面評価値を学習していく．今回の実験では, 選ばれるアクションは最も評価値の高いアクションを選ぶ．

3　解析結果

解析にあたって,10 億盤面分を学習させた平均 89828 点のプレイヤからゲームを開始させて誤った動きがないかを解析した．解析した結果, 図 1 のように大きなタイルを, 動かさなくても良い場面で動かすアクションを行っていることが複数確認された．

図 1 における盤面 $V(S'_{t-1})$ のアクションの評価値となる, また, 図 2 のタプルにおいて部分盤面評価値が,−833 という値を出した．ゲーム中盤 (最高値が 4096) においても誤った動きをする場合があり, その場面ではタプルの評価に極端な偏りがあることも判明した．この評価の偏りが学習不足なのか, また偶然なのかをさらに検証するため, 同様の方法で学習した他の評価値テーブルを用いて図 2 のような問題のある動きを行なった盤面のタプルの評価がどのように変化するのかを検証する．

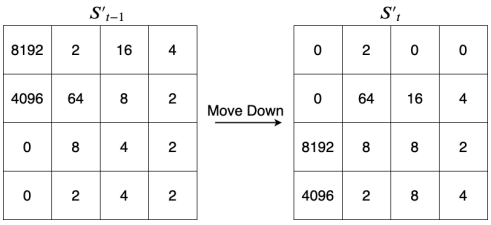

図 1　不合理な動き

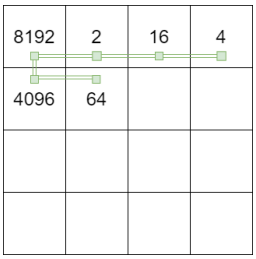

図 2　負の評価値となるタプル

アクション	アクション後の盤面評価値
Up	52499
Down	53312
Left	52591

表 1　S'_{t-1} から選択できる各アクションの評価値

4　まとめ

10 億盤面分学習させたプレイヤにおいて, 大きなタイルを不必要に動かす場面があることが判明した．また, 誤った動きをする盤面において, タプルの評価に偏りがあることも判明した．

参考文献

[1] Marcin Szubert, Wojciech Jaskowski Temporal Difference Learning of N-Tuple Networks for the Game 2048 2014 IEEE Conference on Computational Intelligence and Games, pp. 1-8, 2014

[2] Kiminori Matsuzaki Developing 2048 Player with Backward Temporal Coherence Learning and Restart,15th International Conference on Advances in Computer Games,2017.