強化学習におけるタイルコーディングの性能評価

太田　健二†　　尾関　智子††
† 東海大学院工学研究科情報理工学専攻　†† 東海大学情報理工学部

1. はじめに

強化学習とは，コンピュータが試行錯誤することで目的を達成するための行動を獲得する学習方法である[1]．コンピュータの探索行動により，環境に適応できる特徴がある．しかし，学習空間が連続的な場合では強化学習の導入は困難である．本稿では，連続的な学習空間への代表的なアプローチ手法であるタイルコーディングについて，その性能評価を行う．

2. Q 学習

本研究では学習空間に対して各状態における行動に価値を持たせる Q 学習を用いる．以下にQ学習の更新式を示す．

$$\delta(t) \leftarrow r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \quad \cdots(1)$$

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\delta(t) \quad \cdots(2)$$

ここで，Q は状態行動価値，δ は TD 誤差，r は報酬，s は状態，t は価値更新時刻，a は行動である．α と γ はそれぞれ学習率と割引率を表す．本研究では行動選択手法に ε =0.1 の ε −greedy 手法を用いる．

3. タイルコーディング

タイルコーディングは強化学習において連続的な学習空間を持つ問題への代表的なアプローチ手法である．タイルコーディングは学習空間を任意に分割するタイルの集合であるタイリングからなる．タイリングを学習空間全体に覆うよう実施することで学習空間を擬似的に離散化する．

4. 実験内容

本稿ではタイルコーディングにおいて，タイリングの枚数，タイリングの分割数の違いによって生じる学習速度の変化を空間探索問題において検証する．また，既存手法であるタイムシフトタイルコーディング[2]についても同様に検証を行う．シミュレーション実験を行うにあたり学習環境は図 1 のような$(0.0 \leqq x \leqq 4.0)$, $(0.0 \leqq y \leqq 4.0)$の連続的な空間を用意する．エージェントの初期座標は(0.0, 0.0)でゴールは$(3.9 \leqq x \leqq 4.0) \wedge (3.9 \leqq y \leqq 4.0)$の領域である．エージェントは上下左右の 4 方向に幅 0.05 の間隔で移動する．しかし，移動の幅には毎回の行動毎に標準偏差=0.01 の正規乱数によるブレを生じさせる．環境からエージェントが得られる報酬は行動時に−1，ゴール到着時に 10000000 とする．エージェントがスタートからゴールに到着するまでを 1 エピソードとして，エージェントが 10 回エピソード達成するまでに処理にかかった時間の検証を行う．図 2 に各タイリング枚数における分割数毎の処理時間を記したグラフを示す．縦軸に 10 エピソードまでかかった時間(秒)，横軸にタイリングの分割数を示す．図 2 より本稿で用意した環境において最も学習が速いものは 2 枚タイリングを用いたものであった．1 枚タイリングと 2 枚タイリングを組み合わせた手法であるタイムシフトタイルコーディングはこの環境設定において学習の高速化について有力な結果が得られなかった．

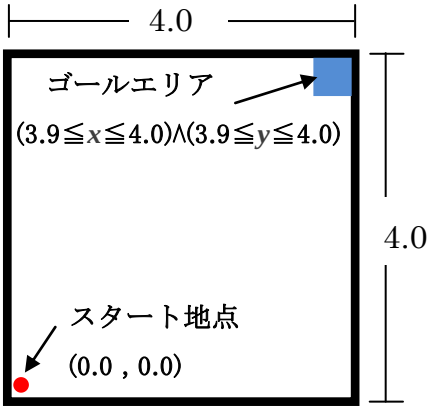

図1. 学習環境

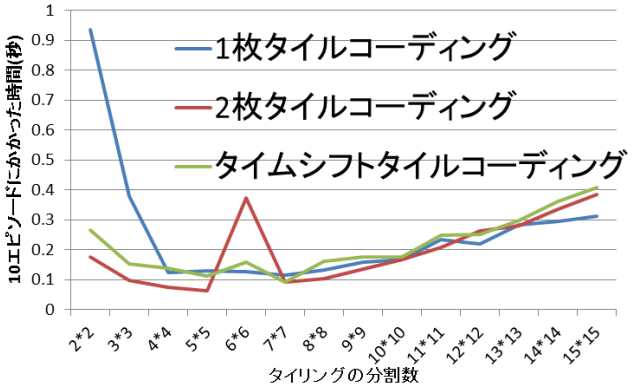

図 2. 各タイリング枚数における分割数毎の処理時間

4. まとめ

空間探索問題において様々なタイルコーディングの学習速度について検証を行った．タイムシフト法について有力な結果を得られかった理由は，本研究と先行研究に学習パラメータや環境の違いがあったためと考えられる．

参考文献

[1] R.S.Sutton, A.G.Barto 著，三上貞芳，皆川雅章共訳，“強化学習”，森北出版 2000.

[2] 梶本洋平，安達雅春 著，”タイムシフトタイルコーディングによる強化学習の高速化の試み”，信学技報，IEICE Technical Report, NLP2008-22(2008-7).