粗い離散値行動の重ね合せによる行動の細粒化

6Q－07

土谷 千加夫　入沢 達矢　乾 伸雄　小谷 善行
東京農工大学工学部情報コミュニケーション工学科

1. はじめに

Q-learning[1]やSarsa[2]などの一般的な強化学習アルゴリズムは，量子化された状態・行動空間を必要とする．そのため，状態数と行動数の増加に対して評価すべき状態行動対が指数関数的に増加し学習を困難にするという**組合せの爆発**が起きる．特に，連続値状態や連続値行動を扱う問題ではこの影響が顕著になる．

本研究では，複数の学習器が行動空間中の粗い特徴点について学習を行い，その結果を重ね合せることで粒度の細かい行動の価値を導出する手法を提案する．この手法は行動数の増加に対して線形オーダでしか特徴点が増加しないので，行動空間の大きな問題においても組合せの爆発の影響を回避できると考えられる．

2. 表形式強化学習の問題点

各状態行動対に対応する行動価値を近似する強化学習手法を表形式の強化学習と呼ぶことにする．

連続値状態や連続値行動を扱う問題では，表形式の強化学習を適用するために，連続値を量子化する必要がある．量子化が粗すぎると近似精度が低下し，量子化が細かすぎると状態行動対の増加により学習が困難になる．

このように，表形式の強化学習は状態空間や行動空間が非常に大きい問題には単純に適用できないという問題点がある．

本研究では，連続値行動空間の扱い方に焦点を当て，従来の強化学習アルゴリズムをより大きな行動空間へ適用可能にする手法を提案する．

3. 提案手法

提案手法は状態空間の離散化手法として知られているtile-coding[2]を行動空間に応用したものである．以下，提案手法を順を追って説明する．

あらかじめ行動空間をいくつかの領域に分割しておく．この分割された領域をタイルと呼び，タイルの集合をタイリングと呼ぶことにする．次に，最初に設定したタイリングを少しずらし，新たなタイリングを作る．以降，これと同じことを繰り返して，相異なる分割を持つタイリングをn枚用意する．タイリングの各領域には，そこに含まれる行動の価値が関連付けられているものとする．

最後に，すべてのタイリングを重ね合せて，行動空間の分割を細かくする(図1)．細かく分割された領域の価値は，そこに重なり合っているすべてのタイルの価値の合計とする．

行動の選択は，タイリングを重ね合せた後の細かい行動価値に対して任意の行動選択規則を適用することで行う．

行動価値の更新は，実際にとった行動を含むすべてのタイルに対して，Q-learningなどの従来の強化学習アルゴリズムで行う．その際に，環境から得た報酬をタイリングの枚数nで等分し，各タイルに分配する．

この手法をアルゴリズムの形で図2に示す．

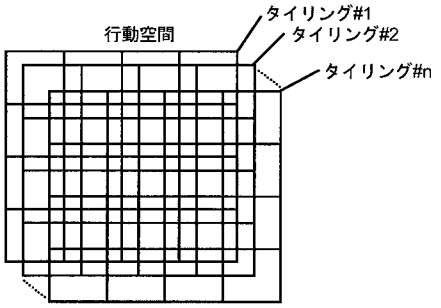

図1　2次元行動空間におけるタイリングの重なり

For all $i \in$ **tile**
　For all $j \in$ **TILE**
　　if i is_constracted_by j
　　$Q(i) \leftarrow Q(i)+\theta(j)$
select_action_from **Q**
receive_reward
For all $j \in$ **TILE**
　if i is_constracted_from j
　　update $\theta(j)$ with usual RL algorithm

tile：重ね合せで細かくなったタイルの集合

TILE：重ね合せる前のタイルの集合

Q：**tile** に関連付けられた行動価値の集合

θ：**TILE** に関連付けられた行動価値の集合

図2　アルゴリズム

図1のようにタイリングを単純に重ね合せた場合，行動空間の周辺にすべてのタイリングが重なり合わな

Making Actions of Fine Granularity from Rough Discrete Actions.
Tsuchiya Chikao, Irisawa Tatsuya, Inui Nobuo, Kotani Yoshiyuki.
Tokyo University of Agriculture and Technology. 184-8588.

い部分が生じる．この部分の行動価値は正しく評価されないため，この部分に最適行動が含まれる問題では，提案手法は正しく機能しないと考えられる．

これに対処するためには，最適行動がタイリングの十分内側に入るように，タイリングを十分に広くとっておく必要がある．

4. 比較実験

提案手法とタイリングの重ね合せを行わない従来法を RoboCup サッカーシミュレーションのパスの学習に適用し，それらの学習効率を比較する実験を行う．

実験では提案手法を用いたエージェントを 2 種類用意する．ひとつはタイリングを 2 枚重ね合せて，ボールを蹴る角度を receiver に対して±30°，±20°，±10°，0°，蹴る強さを＋30，＋60，＋90 とした合計 21 通りの行動をもつ[propose-1]である．このときのタイリング 1 枚あたりの行動は，ボールを蹴る角度が 4 通り，蹴る強さが 2 通りの合計 8 通りである．

もうひとつは，このタイリングを 4 枚重ね合せた[propose-2]である．

比較対象となるエージェントは，[propose-1]の行動と同じ数の行動を 1 枚のタイリングで表現する[fine]と[propose-1]と同じタイリングを 1 枚用いる[rough]を用意する．

実験で使用する局面を図 3 に示す．passer とボールを 1m離して配置し，receiver を 10m以上 20m 以下だけ離して配置する．パスを阻害する enemy は passer から receiver を見たときに，−40°以上＋40°以下の角度，5m以上 20m 以下の距離になるようにランダムに配置する．

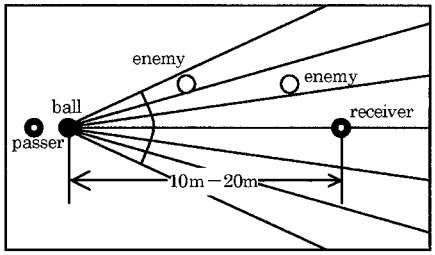

図 3　初期配置と状態空間の分割

実験での 1 試行は passer がボールを蹴ることで開始し，次のいずれかの条件が満足されたときに終了する．

[条件 1] receiver がボールをとる　(成功)
[条件 2] enemy がボールをとる　(失敗)
[条件 3] 40 simulation step が経過する　(失敗)

行動価値の更新は，条件 1 で終了した場合に＋1，それ以外の条件で終了した場合に−1 の報酬を与え，この報酬をとった行動の価値に加算することで行う．

10000 試行の学習を 10 回行ったときの 100 試行ごとに平均パス成功率を図 4 に示す．

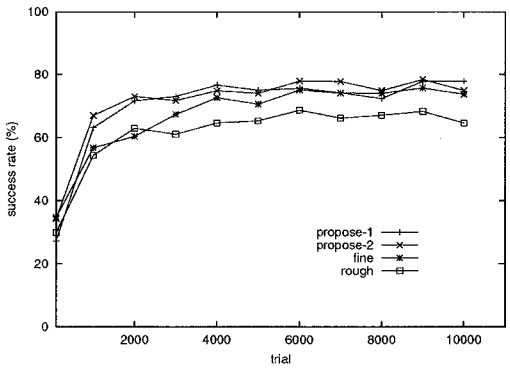

図 4　パス成功率の推移

5. 考察

タイリングを 4 枚重ね合せた[propose-2]の方がタイリングを 2 枚重ね合せた[propose-1]よりも行動の粒度が細かいので，より精度の良い解に収束するはずである．しかし，図 4 では [propose-1]と[propose-2]がほぼ同じ学習曲線を示している．このことから，実験を行った問題で必要とされる行動の粒度が[propose-1]の粒度で十分であったと考えられる．逆に言えば，行動数が余計に多い場合でも学習効率の低下は見られない．

また，計算コストを考えると，従来法の[fine],[rough]に比べて，提案手法の[propose-1]は 2 倍の計算コストを要する．図 4 の横軸を計算コストが一定になるようにスケール変換すると，[fine]と[propose-1]がほぼ同じ曲線になってしまう．したがって，実時間処理を行わない問題では提案手法の効果は小さくなる．しかし，RoboCup サッカーのように実時間処理を行わなければならない問題や行動の機会があまり与えられないような問題では汎化作用のある提案手法が有効である．

6. おわりに

連続値行動空間において粗い離散値行動の価値を重ね合せることで行動価値を細粒化する手法を提案した．

本手法は RoboCup サッカーシミュレーションのパスの学習において従来法に対して精度を落すことなく従来法よりも速い収束を実現した．

参考文献

[1] Watkins, C. J. C. H. and Dayan, P.: Technical Note: Q-Learning, Machine Learning 8, pp. 279-292 (1992).

[2] Richard S. Sutton and Andrew G. Barto: Reinforcement Learning: An Introduction, A Bradford Book, The MIT Press (1998).