ISSN 2186-5647

P-21

温度パラメータを調整することによる狭路問題の学習高速化

日大生産工（学部）　○村松 匡史　日大生産工　山内 ゆかり

1 まえがき

マルチエージェントシステムとは、多数の自律エージェントが近傍のエージェントとの相互作用を通して自律的にルールを作るシステムのことである。システム内外の環境変化に対して柔軟なシステムを構築することができると期待されている。しかし、多数のエージェントが限られた空間で競合するため、設計者があらかじめ個々のエージェントに適切な行動を組み込んでおくことが困難であるとされている。そこでこの問題を解決する方法として強化学習が注目されている。

強化学習は報酬を頼りに試行錯誤を通じて、環境に適応する学習制御のことである。教師あり学習とは異なり、状態入力に対する正しい出力を明示的に示す教師が存在しない。代わりに報酬という情報を頼りに学習するが、学習に時間を要する。そのため学習初期の行動直後の報酬を見るだけではその行動の正誤性の判断は難しい。したがって幾度の試行錯誤を通して学習を進めていく。

マルチエージェントがすれ違う狭路問題の研究として、Moriyamaらは、道を譲ったエージェントが周囲のエージェントから報酬を得ることで著しく不利益になることを軽減する近傍報酬を導入している[1]。しかし、各エージェントは、事前に相手の行動を知ることができないため、正誤性の判断ができなくなってしまう可能性がある。

この問題に対して、山田らはMoriyamaらの研究で発生した不完全知覚状態を低減し、従来よりも高い確率で競合回避をするため、信頼度を用いて強化学習の割引率を自律的に調整する手法を提案した[2]。その結果不完全知覚状態が発生する状況でも競合回避ができ、従来よりも高い確率で競合回避行動を獲得することができた。しかし、学習速度の面については普通のQ-Learningより向上していなかった。

そこで、本研究では行動選択の際に用いるボルツマン選択の温度パラメータTを調整することで、学習速度の高速化を目指す。

2 従来研究

2.1 山田らは信頼度を用いて自律的に割引率を調整する強化学習を使って、マルチエージェント狭路問題での競合回避行動を以前よりも高い確率で獲得することに成功した。

次に提案手法のアルゴリズムを示す。

観測した状態における行動価値の Q 値を使い、式(1)により行動を選択する。実行した行動に対し報酬が与えられ、式(2)により割引率γを調整する。Q 値を新しい割引率γに基づいて更新し、信頼度 R(s,a)を式(3)、式(4)により更新する。

終了条件を満たすまで 1 から 3 を繰り返す。

2.2 行動選択および自律的な割引率の調整

次式は、状態 s における行動 a の選択確率 p(a|s)をボルツマン方策を用いて示したものである。T は温度パラメータ、b は行動集合 A の要素を表す。

$$p(a|s) = \frac{\exp(Q(s,b)/T)}{\sum_{b\in A}\exp(Q(s,b)/T)} \tag{1}$$

割引率γは信頼度 R(s,a)に基づき次式により調整する。

$$\gamma_t = \min\left(1, \frac{1}{R(s,a)}\right) \tag{2}$$

また、信頼度 R(s,a)は次式により更新される。

$$\varsigma = \delta^2 + \gamma_R R^2(s',b) - R^2(s,a) \tag{3}$$

$$R^2(s,a) \leftarrow R^2(s,a) + \alpha_R \varsigma \tag{4}$$

信頼度 R(s,a)は、エージェントが行動した際に得た報酬と Q 値との二乗誤差δとの累計を表される。したがって Q 値と報酬の誤差が大きい場合、信頼度 R(s,a)の値は大きくなり、γの値は小さくなる。逆に Q 値と報酬の誤差が小さい場合、信頼度 R(s,a)の値も小さくなりγの値は大きくなる。Q 学習の性質として

An Efficient Learning of Narrow Road Problem by Adjusting Temperature Parameter

Masashi MURAMATSU and Yukari YAMAUCHI

割引率γの値が小さいとき、エージェントは即時報酬を重要視して学習し、割引率γの値が大きいとき、長期報酬を重要視して学習する性質がある。

3　提案手法

本研究では、学習速度の面については普通の Q-Learning と大きく変わっていないという問題を解決するべく、従来研究に対して行動選択部分の温度パラメータ(T)を調整することにより学習の高速化を図る。

基本は従来手法同様のアルゴリズムで学習させるが、式(1)の行動選択確率 p(a|s)の温度パラメータ(T)を従来では0.1に固定して実験を行っていた。それを環境によって変化させることにより 0 に漸近させる。

4　実験環境

本実験では、以下の2次元平面環境の狭路問題を用いて実験を行う。

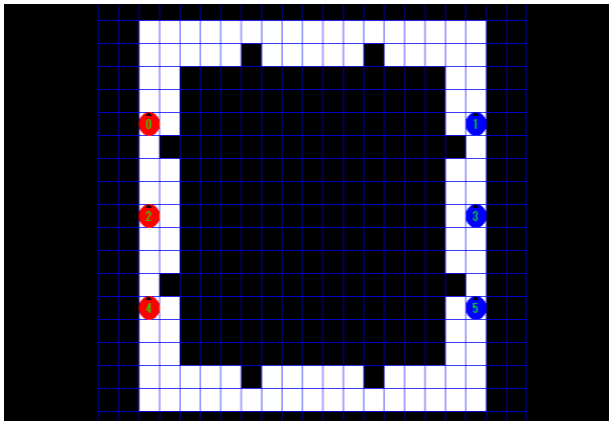

図1.実験環境

スタート位置は図1のエージェントの位置を固定とする。番号が偶数のエージェントが時計回りに、奇数のエージェントが反時計回りに周回する。また、エージェントには進行方向に対し、11近傍の状態と通路の内側外側の判別できる。入力はマスが空白の時0、壁の時1、周回方向が同じエージェント時2、周回方向が異なるエージェントの時3とする。エージェントの視界の様子を図2に示す。

エージェントが500ステップ行動したとき1エピソードとし、エピソードを更新するときエージェントは初期位置に戻って再行動し直す。3000エピソードを1トライアルとし、これを25トライアル行う。また、報酬は通路に対して前に進めたときのみ報酬1を与え、他のエージェントや壁に衝突した際には罰則2を、周回方向に対して前に進めなかったときは罰則1を与える。

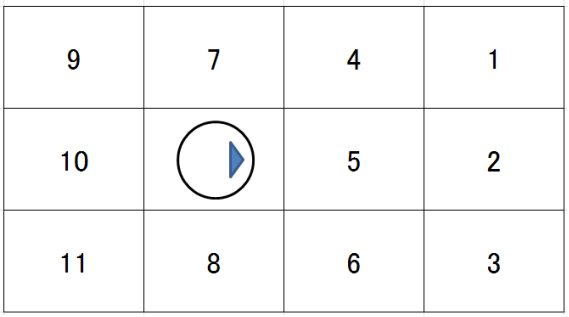

図2.エージェントの視野

そして本研究では、正の報酬回数獲得の早期収束を目標とする

5　まとめ

本研究ではマルチエージェント環境下での狭路問題に対して、従来研究の問題点の一つであった学習速度を温度パラメータ(T)調整することにより改善するものを提案した。

今回の手法が実装することができれば、複雑な環境下でも真の解を選ぶ確率が高くなり、より効率的な学習を可能にできると考えている。

今後の課題は、エージェントの視野を進行方向によって場合分けをした上で提案手法を実装することである。

「参考文献」

1) K. Moriyama and M. Numao: Self - Evaluated Learning Agent in Multiple State Games, Proc. 14th European Conference on Machine Learning, ECML-2003, 289/300, (2003)
2) 山田和明, 高野慧,「マルチエージェントのための信頼度を用いた強化学習-相補ゲームにおける競合回避行動の獲得-」, 計測自動制御学会論文集, Vol.49, No.1, 39/47.(2013)
3) 野田五十樹, Kim Hyun-Tae , "マルチエージェント学習下における温度パラメータの調節手法", 人工知能学会全国大会予稿集, pp. 1G1-1, 人工知能学会, 2011