ISS-P-041

マルチエージェントシミュレーションへの深層強化学習の適用検討

豊味 諒磨　　尾崎 敦夫
大阪工業大学　情報科学部　コンピュータ科学科

1. はじめに

マルチエージェントシミュレーションでは，各エージェントは，基本的に定められた規則に従い行動する．本研究では，同一の環境に対して，複数のエージェントの行動規則を学習させるための有効な手法を検討する．さらにシミュレーション終了後，シミュレーション自体を評価し，シミュレーション時間(論理ステップ数)が最短になるようなエージェントの行動を，学習により調整する手法の検討も行う．

2. 提案手法

2.1 対象とする問題

本研究では以下のような状況を対象とする．

(1) エージェントは，複数存在し，共通の目標を持つ．
(2) エージェント同士は，競争・競合する関係にある．
(3) エージェント同士は，お互いの位置や状態を認識できる範囲内で行動する．

これらの状況でシミュレーション時間が最小となるようなエージェントの動作を考えるものとする．

2，2 アルゴリズム

提案手法での処理内容は下記(1)～(3)により構成され，これらの処理の流れを示したものが図1である．

(1) 1 シミュレーション当たりの動作

各エージェントは異なる Q 関数を持ち，ε -greedy 法[1]により行動を選択する．行動実行後の各エージェントの報酬と，変化した環境から Q 関数を更新する．全エージェントが目標地点に到達したらシミュレーションを終了し，その時の全エージェントの Q 関数と各ステップにおける報酬値から評価値表を作成し，メモリに保存する．

(2) 複数回のシミュレーションでの評価

複数回シミュレーションを行い，メモリに保存したデータをもとに，各ステップでの全エージェントの最適行動を求める．ここでの評価は最終的にステップ数が少なかったシミュレーションの方が高い報酬を得るものとする．ここでは Deep-Q-Neteork[1]を用いて学習を行う．具体的には Qa 関数を用いて，全エージェントの状態(環境情報)s を与え，最も報酬 r が得られると期待される，行動(Qa 関数が最大となる 1 ステップ動作)a，を選択するように学習させる．

(3) Q 関数を更新

(1)(2)で得られた，各ステップでの動作を反映させて再度シテップでの動作を反映させて再度シミュレーションを行い各エージェントの Q 関数を更新する．更新した Q 関数を用いて再び学習させる．

3. 期待する効果

本研究では，多数のエージェントが存在する場合において，シミュレーションのステップ回数が最小となる結果を導出できることが期待できる．

また ε -greedy 法によりランダムな行動選択を交ぜ学習させた時の Q 関数がデータとして存在する．そのため，各エージェントがそれぞれの Q 関数をもとに最良の評価値となる行動を選択し続けることで，全体の結果(シミュレーション全体)の劣化を防げることが期待できる[2].

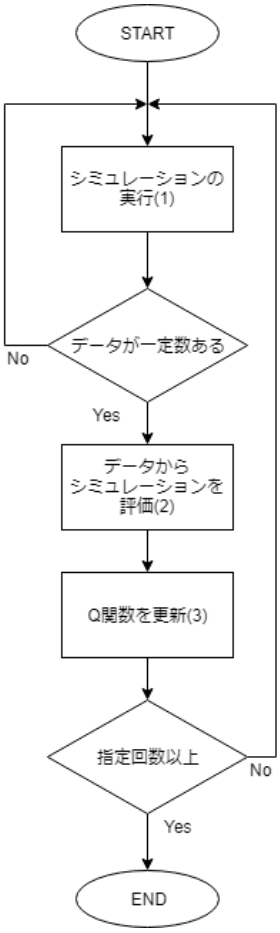

図 1. 提案手法での処理の流れ

4. 今後の課題

本研究では，Q 関数を持ったエージェントが学習し，その結果から Deep-Q-Network を用いてさらに学習する．即ち，1 度学習した結果でシミュレーションを行い，その結果から再度学習させる．そのため，評価する計算量が多くなると考えられ，その対策が必要となる．

また，現状では共通の目標を持ち，競争・競合関係にある問題を対象としているが，異なる目標を持ち，協力関係にある問題等にも対応できる手法の検討も今後の課題である．

参考文献

[1] 佐藤季久恵「Deep Q-Network を用いた交通信号制御システムの提案」， 2017 年度人工知能学会全国大会(第 31 回) 3I2-OS-13b-4, 2017

[2] 山分翔太「マルチエージェントタスクに対する群強化学習法」，計測自動制御学会論文集，2013 年 49 巻 3 p.370-377,2013