7P-07

シュート判定器を用いた深層強化学習による Half Field Offence タスクの学習

島 健人† 相馬 隆郎†

東京都立大学†

1. はじめに

近年、人工知能技術への関心が高まっており、その中でも正解を人間が与えることが難しい問題を試行錯誤しながら学習する強化学習はロボット制御、自動運転など様々な分野で活躍が期待されている。強化学習は、設計者が適切な報酬を設定する必要があり、この報酬を得ることで学習が進むが、問題が複雑である場合や、困難である場合において報酬を得ることが難しく、学習が進まない場合がある。本研究では RoboCup Soccer 2D Simulation のサブタスクであり、ゴール以外での報酬設定が困難な Half Field Offence タスクにおいて効率的な学習を試みた。

2. RoboCup Soccer

RoboCup Soccer とは”2050 年までに、サッカーの世界チャンピオンに勝てる、ロボットを作る”という目標のプロジェクトであり、この目標に向けて、シミュレーション環境において強化学習を用いて、各種動作の獲得を目指した研究[1]や、複数の敵と味方エージェントが存在する keepaway タスクにおいて各エージェントの行動方策獲得を試みた研究[2]などがある。

3. シュート判定器を用いた HFO タスクの学習

川上ら[3]は深層強化学習を Half Field Offence タスクのに適用し、ゴール時にのみ報酬を与えオフェス 3 人対ディフェンス 2 人の環境におけるオフェス側の行動方策の獲得を試みた。Half Field Offence タスクでは、環境の簡略化のために図 1 のようなサッカーコートの半分にオフェンス及びディフェンスを配置し、オフェンスがゴールを Agent2D[4]をもとに用意されたドリブル、シュートなどの行動空間を用いて目指す。人が任意の状態の評価値を適切に構築することは困難であり、評価値を決定する必要のない強化学習は行動方策獲得に有効であると考えられるが、環境が難しくなるにつれて報酬を得ることができずに学習が進まなくなる場合がある。

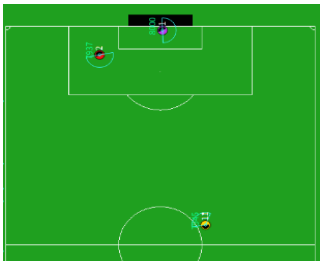

図 1 Half Field Offence タスク

本研究では任意の状態からのシュート成功率を予測する判定器を事前学習することで報酬を得る回数が少ない環境において効率的な学習を目指す。シュート判定器の結果が式(1)で表される基準 a より低い場合、シュートを行動選択肢から外し探索を行う。

$$a = 0.4 \times \left(1 - \frac{e}{E}\right)^2 \quad (1)$$

e は現在のエピソード数、E は総エピソード数である

4. 実験結果

4.1 シュート判定器の作成

4.1.1 作成手順

任意の状態からシュートが入る確率を予測するシュート判定器を作成するために下記のようにデータを作成する。

1) オフェンス 1 人をランダムに配置。
2) キーパーをゴール前に配置。
3) ディフェンス 1 人をランダムに配置。
4) この状態から 5 回シュートを行う。
5) 得られた結果からシュート成功率を算出。
6) 1～5 を繰り返す。

さらに上記で得られたデータを用いて状態から成功率を予測する 3 層(ノード数 128)ニューラルネットワークの学習を行う。

A Deep Reinforcement Learning Using Shoot Judge for Half Field Offense
†Shima Kento and Soma Takao, Tokyo Metropolitan University

4.1.2　シュート判定器学習結果

学習の訓練中とテスト結果を横軸にエポック数、縦軸に平均絶対誤差を表したグラフを図2に表す。

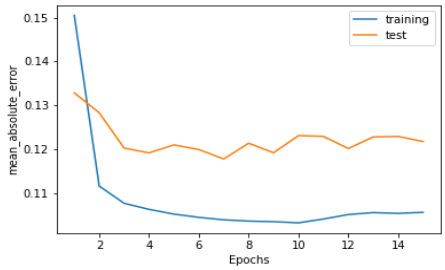

図2　シュート判定器の学習

訓練後の平均絶対誤差は 0.121 となっている。この結果からおおむねシュート成功率を予測できていることが分かる。

4.2 HFO タスクの学習

4.2.1　問題設定

本研究では、表1のようにオフェンスの人数を変えて2つのパターンで学習を行った。

表1　各エージェント数

	オフェンス	ディフェンス
パターン1	3人	2人
パターン2	1人	2人

学習には状態変数として各エージェントの x, y 座標を用いた。また行動選択肢には、周囲8方向へのドリブル、シュート、その場でボールを保持、味方へのパスとした。学習は3層(ノード数 128)ニューラルネットワークを用いた Deep Q-Network を使用し、5000 エピソード行い、その後 1000 エピソードのテストを行った。

4.2.2　パターン1（3対2）の学習

学習中にシュート判定器によって行動選択に制限をかけた場合とかけていない場合のそれぞれの結果を横軸にエピソード数、縦軸にゴール成功率を表したグラフを図3に表す。

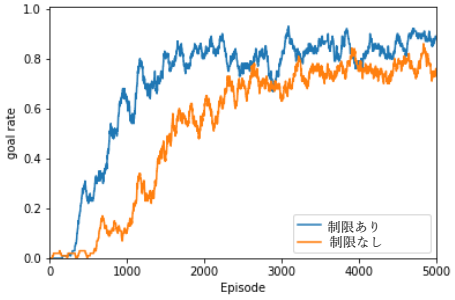

図3　パターン1学習結果

グラフから見て取れるように制限ありの方が学習の進みが早く、学習後のテスト結果でも制限ありで、89.7%、制限なしで、76.5%となり制限ありの方が高い結果となった。

4.2.3 パターン2（1対2）の学習

パターン 1 と同様に結果を表したグラフを図4に表す。

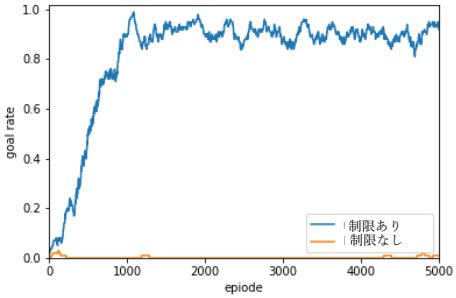

図4　パターン2学習結果

制限なしの場合、学習が進んでいないことが見て取れる。制限ありの場合では、学習後のテスト結果ではゴール率 95.5%となった。

5. おわりに

本研究では HFO タスクにおいて、シュート判定器を事前学習し、判定器の出力によって学習中の行動選択に制限をかけることで学習能力が向上することを確認した。また、ゴール率においても、パターン1と同様の条件下で行動選択肢のドリブルを Agent2D の判断で行った先行研究[3]のゴール率 80% を上回る 89.7% を達成した。

6. 参考文献

[1] Tomoharu Nakashima, Masayo Udo and Hisao Ishibuchi: A Fuzzy Reinforcement Learning for a Ball Interception Problem, RoboCup 2003:Robot Soccer World Cup VII, pp. 559-567, (2004).

[2] Peter Stone, Richard S. Sutton and Gregory Kuhlmann: Reinforcement Learning for RoboCup-Soccer Keepaway, Adaptive Behavior, 13(3), pp. 165-188, (2005)

[3] 田村 啓郎, 川上 翔平, 相馬 隆郎” ロボカップサッカーの keepaway および Half Field Offence タスクへの深層強化学習の適用 ”(2018)Akiyama Hidehisa. “Agent2d base code,” 2010.

[4] Akiyama Hidehisa. “Agent2d base code,” 2010.