The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

ゲーム内の学習エージェントに対する直接教示効率化のための支援手法の検討

A Preliminary Approach on Efficient Teaching Support to Non-player Learning Agents on Multiplayer Games

筒井壮太郎 *1　福田直樹 *2
Sotaro Tsutsui　Naoki Fukuta

*1静岡大学情報学部
Faculty of Informatics, Shizuoka University

*2静岡大学学術院情報学領域
College of Informatics, Academic Institute, Shizuoka University

Giving human knowledge to learning agents is a good way to speed up the process of reinforcement learning for learning agents. To give human knowledge to learning agent efficiently, it is important to estimate whether or not agents need more knowledge. In this paper, we present our preliminary approach on realizing efficient teaching on an application which can show the users the progress of learning in a video game.

1. はじめに

エージェントを用いることによって環境における解を得るための方法の中で有名なものの一つとして，強化学習があげられる．強化学習では，エージェントが試行錯誤を行い，報酬が最大になるようなふるまいを探索することで解を収束させる．

強化学習の手法を用いることによって生じる処理時間や計算量に関する課題を緩和するための方法の一つに，学習中のエージェントに対して人の知識を与えられるようにするものがあげられる．学習中のエージェントに対して人の知識を与えられるようにする方法に関する研究は過去に多く行われており，その方法の種類は複数提案されている．どのようにして人が学習中のエージェントに対して知識を与えるかという点についての議論は多くなされてる一方で，人が学習中のエージェントに対して知識を与えるために，どのようにして人が学習中のエージェントのふるまいを観測し，学習中のエージェントがどのような状態であった場合に人が知識を与えるべきかというような点についての議論は比較的少ない．

強化学習を用いて解を効率よく収束させるための方法として，人が学習中のエージェントに対して知識を与えられるようにする方法を用いる場合，エージェントに対して人が与える知識の量やその回数を最小限にしつつ効果的に利用させる必要がある．学習中のエージェントに対して人が効率よく知識を与えるためには，学習環境における各状態において，学習中のエージェント対して人が知識を与える必要があるかどうかを容易に判断できるようにする必要がある．本稿では，学習中のエージェントに対して人が知識を与える必要があるかどうかを容易に判断できるようにするために，学習の進行の度合いを数値化し，その数値に応じて表示が変化するような機構の試作を進めている．

2. 研究の背景

2.1 マルチプレイヤー戦闘ゲーム

本稿で述べるマルチプレイヤー戦闘ゲームとは，人が追跡者であるキャラクターのうちの一体を操作して追跡対象を倒すことを目標としたゲームである．本ゲームにおいてゲームを優位に進行させるためには，人が操作している追跡者キャラクターと，ゲーム側で行動を制御している他の追跡者キャラクターが協力して追跡対象を倒すことが必要である．本稿では，ゲームを追跡問題をヒントにして，その条件を拡張したものとして実装を行った．本稿で実装したゲームの外観を図 1 に示す．

図 1 で，追跡者は図 1①と②のキャラクターであり，追跡者には上下左右いずれか 1 マスへの移動と攻撃および上下左右 4 マスへ同時に捕獲行動が行えるようになっている．一方で，追跡対象は図 1 の③であり，上下左右 1 マスへの移動と停止を行動として用意した．本ゲームは，一方の追跡者，追跡対象，もう一方の追跡者，追跡対象の順に行動することで 1 ターンが経過し，N ターン経過で 1 ゲームが終了するとしている．一方の追跡者が隣接している追跡対象に捕獲行動をとった場合，追跡対象は次の行動が不可能となる．行動不能な追跡対象にもう一方の追跡者が攻撃を当てた場合はゲームに勝利し，ゲームに勝利しないまま N ターンが経過した場合はゲームに敗北する．本ゲームにおいて追跡対象を倒した場合の具体的な例を図 2 に示す．

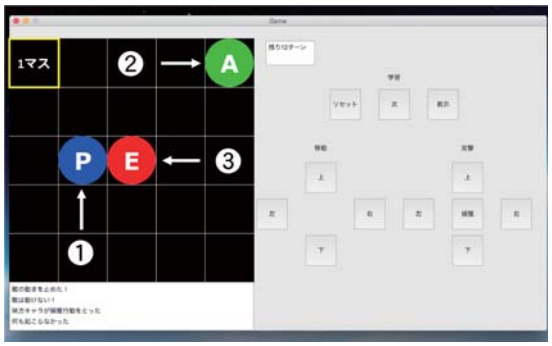

図 1: 本稿で実装したゲームの外観

2.2 学習エージェントに人が知識を与えるアプローチ

強化学習の手法を用いることによって生じる処理時間や計算量に関する課題を緩和するための方法の一つに，学習中のエージェントに対して人の知識を与えられるようにするものがあげられる．学習中のエージェントに対して人の知識を与えられるようにする方法に関する研究は過去に多く行われており，

連絡先: *1 筒井壮太郎，静岡大学情報学部情報社会学科，〒 432-8011 静岡県浜松市中区城北 3-5-1 ia14057@s.inf.shizuoka.ac.jp
*2 福田直樹，静岡県学術院情報学領域，〒 432-8011 静岡県浜松市中区城北 3-5-1，fukuta@inf.shizuoka.ac.jp

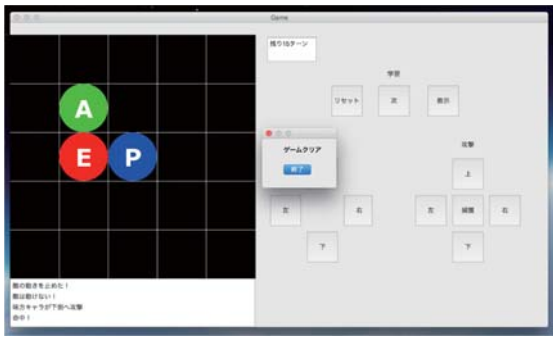

図 2: 本ゲームをクリアした場合の例

その方法の種類は複数提案されている．学習中のエージェントに対して人の知識を与えられるようにするための具体的な方法の例としては，学習中のエージェントのふるまいを人が監視し，エージェントが誤ったふるまいを行っていた時に人がエージェントに対して適切なふるまいを教示できるようにする方法 [Clouse 92] や，学習中のエージェントのふるまいを人が監視し，エージェントのふるまいに対して人が報酬を与えられるようにする方法 [Knox 08] や，環境から観測可能な状態の中で類似している状態を人が評価し，エージェントの学習速度をあげられるようにする方法 [Ariel Rosenfeld 17] などがあげられる．これらの研究は，どのようにして人が学習中のエージェントに対して知識を与えるかという点について議論をしているものである．

一方で，人が学習中のエージェントに対して知識を与えるために，どのようにして人が学習中のエージェントのふるまいを観測し，学習中のエージェントがどのような状態であった場合に人が知識を与えるべきかというような点についての議論は比較的少ない．人が学習中のエージェントに対して知識を与えられるようにする方法に関する研究における，人がエージェントに知識を与えるために学習中のエージェントのふるまいを観測するための方法としては，学習を行っている最中のエージェントのふるまいを人が観測するものや，学習を一定期間行ったのちに，エージェントのふるまいを人が観測するものなどがあげられる．

学習を行っている最中のエージェントのふるまいを人が観測する方法 [Knox 08][Thomaz 06] では，エージェントに学習を行わせつつ，同時並行で人がエージェントに対して知識を与えることを可能にしているため，エージェントに学習を行わせる回数を減少させられると考えられる．強化学習ではエージェントに繰り返し試行錯誤を行わせることで解を収束させるため，解を早く収束させるためにはエージェントに高速で繰り返し試行錯誤を行わせる必要がある．一方で，人が物体が移動している様子を映像として認識することが可能な速度には限界があるため，高速で試行錯誤を行っているエージェントのふるまいを人が観測することは困難であると考えられる．

2.3 学習エージェントの観測と知識付与

人がエージェントのふるまいを観測しやすいようにエージェントの試行錯誤の速度を遅くすると，解が収束するまでにかかる時間が増加してしまうと考えられる．学習を一定期間行ったのちに，エージェントのふるまいを人が観測する方法 [Clouse 92] では，一定期間エージェントに学習を行わせたのちに，その学習データを用いてエージェントに実環境で行動をさせ，エージェントがどのようにふるまうかを観測するものである．この方法では，人が観測することが可能な試行錯誤を行っているエージェントのふるまいの速度には限界があるという点を解決するための一つの方法であると考えられる．一方で，学習を一定期間行ったのちにエージェントのふるまいを人が観測する場合，人がエージェントに対して知識を与えるかどうかを判断するためには，エージェントがどのような状態においてどのようにふるまうかを網羅的に観測する必要がある．単純な学習環境においては，エージェントが観測することが可能な状態は比較的少なく，人がエージェントのふるまいを網羅的に観測することは比較的容易であることが予想される．複雑な学習環境においては，エージェントが観測することが可能な状態は増加するため，人がエージェントに対して知識を与える必要があるような状態を発見するためには，非常に多くのエージェントのふるまいを観測する必要がある．人が観測する必要のあるエージェントのふるまいの数を減らすために，エージェントに対して知識を与える必要があると考えらえるような状態を人が予想し，その予想した状態におけるエージェントのふるまいを局所的に観測するという方法が考えらえる．一方で，学習の環境が複雑な場合，人がエージェントに対して知識を与える必要があると考えられる状態を人が予測することが困難であると考えられる．

3. 本稿におけるアプローチ

3.1 機構および機構に必要な指標の検討についての概要

強化学習において，人がエージェントに対して知識を与えられるようにする方法を用いて効率よく解を収束させるためには，エージェントに対して人が与える知識の量やその回数を最小限にしつつ効果的に利用させる必要がある．学習中のエージェントに対して人が効率よく知識を与えるためには，環境の各状態において，学習中のエージェント対して人が知識を与える必要があるかどうかを容易に判断できるようにする必要がある．本稿では，学習中のエージェント対して人が知識を与える必要があるかどうかを容易に判断できるようにするために，エージェントの学習の進行度を数値化し，その数値に応じて表示が変化するような機構を検討を進めている．

機構の試作を行うにあたって，強化学習においてエージェントの学習の進行度合いを数値化するために用いることが可能であると考えられる指標について，検討する必要がある．機構に用いるために必要な指標として，本稿では，各状況における学習の進行度を示すための指標を検討し，その指標を用いた機構の試作を検討する．

3.2 各状況における学習の進行度合いを示すための指標の検討

特定の学習エピソードまでに，学習環境に対する各状況における学習の進行度合いを示すための指標を検討するにあたって，本稿では，状態 s を観測した回数，状態-行動の対 (s, a) における value 関数 $Q(s, a)$，およびそれらを用いて新たに定義するヒューリスティックな指標を考え，本稿で試作を検討する機構への適用をそれぞれ試みる．

一つ目のヒューリスティックな指標として，状態 s を観測した回数を学習環境に対する各状況における学習の進行度合いを示すための指標として用いることを検討する．状態 s を観測した回数と学習の進行度合いの関係は学習に用いるエージェントに従わせる探索の方法に依存する点や，学習を十分行うために必要な状態 s の具体的な観測回数は環境における状態やその状態における選択可能な行動の数に依存する点から，学習環境に対する各状況における学習の進行度合いを示すため

の指標としてそのまま用いることは困難であると考えられる．本稿では，状態 s を観測した回数を学習環境に対する各状況における学習の進行度合いを示すための指標として用いることを可能にするための方法として，状態 s における学習の進行度合い $Progress(s)$ を

$$Progress(s) = 1 - (1+p)^{-\Upsilon(s)} (0 \leqq Progress(s) < 1)$$

と定義し，本稿で試作を検討する機構への適用を試みる．$\Upsilon(s)$ とは，状態 s を観測した回数 $n_{visits(s)}$ を用いた関数であり，

$$\Upsilon(s) = \sqrt{n_{visits(s)}}$$

と定義する [Silva 17]．状態 s を観測した回数が増加するにつれて $\Upsilon(s)$ の値は増加し，$Progress(s)$ の値も増加する．$Progress(s)$ の値の増減は，$p(0 < p)$ の値によって制限され，p を大きな値に設定するほど $Progress(s)$ の値は急激に増加する．

二つ目のヒューリスティックな指標として，状態-行動の対 (s,a) における value 関数 $Q(s,a)$ を学習環境に対する各状況における学習の進行度合いを示すための指標として用いることを検討する．一般に強化学習においては真の Q 値は不明であるために，エージェントによって更新させた Q 値と真の Q 値を比較することが困難であるという点から，状態-行動の対 (s,a) における value 関数 $Q(s,a)$ を学習環境に対する各状況における学習の進行度合いを示すための指標としてそのまま用いることは困難であると考えられる．本稿では，状態-行動の対 (s,a) における value 関数 $Q(s,a)$ を学習環境に対する各状況における学習の進行度合いを示すための指標として用いることを可能にするための方法として，状態 s における学習の進行度合い $Progress(s)$ を

$$Progress(s) = 1 - (1+m)^{-\psi(s)} (0 \leqq Progress(s) < 1)$$

と定義し，本稿で試作を検討する機構への適用を試みる．$\psi(s)$ とは，状態 s およびその状態において選択可能な行動 a の対におけるそれぞれの value 関数 $Q(s,a)$ の収束の度合いを示す関数であり，

$$\psi(s) = |max_a Q(s,a) - min_a Q(s,a)|$$

と定義する．強化学習では，学習が進行するにつれて Q 値は更新され，真の Q 値へと近づく．Q 値が更新されていくにつれて，状態 s における行動 a の価値はそれぞれ更新されていき，大きな報酬を得られるような行動に対する Q 値は大きな値に更新され，小さな報酬を得られるような行動に対する Q 値は小さな値に更新される．$Progress(s)$ の値の増減は，$m(0 < m)$ の値によって制限され，m を大きな値に設定するほど $Progress(s)$ の値は急激に増加する．

三つ目のヒューリスティックな指標として，状態 s を観測した回数と状態-行動の対 (s,a) における value 関数 $Q(s,a)$ を併用した関数を学習環境に対する各状況における学習の進行度合いを示すための指標として用いることを検討する．状態 s を観測した回数のみを用いて学習環境に対する各状況における学習の進行度合いを示す場合，状態 s において選択可能なそれぞれの行動 a についてどれほど学習が進行しているのかは確認が困難となることが予想される．一方で，比較的な正確な Q 値に更新されるまでには時間がかかるため，状態 s において選択可能なそれぞれの行動 a についてどれほど学習が進行しているかを確認できるような関数のみを用いて学習環境に対する各状況における学習の進行度合いを示す場合，十分に状態 s を観測していない場合でも，学習の進行度合いを示す関数は大きな値を取ってしまう可能性があると予想される．本稿では，状態 s を観測した回数を考慮に入れつつ，状態 s におけるそれぞれの選択可能な行動 a に対する value 関数を学習環境に対する各状況における学習の進行度合いを示すための指標として用いるために，状態 s における学習の進行度合い $Progress(s)$ を

$$Progress(s) = 1 - (1+n)^{-\Psi(s)} (0 \leqq Progress(s) < 1)$$

と定義し，本稿で試作を検討する機構への適用を試みる．$\Psi(s)$ とは，状態 s を観測した回数を考慮しつつ，状態 s およびその状態において選択可能な行動 a の対におけるそれぞれの value 関数 $Q(s,a)$ の収束の度合いを示す関数であり，

$$\Psi(s) = \Upsilon(s)|max_a Q(s,a) - min_a Q(s,a)|$$

と定義する．$\Psi(s)$ の値は，状態 s の観測回数が増加し，それぞれの行動 a に対する Q 値が更新されていくにつれて増加する．$Progress(s)$ の値の増減は，$n(0 < n)$ の値によって制限され，n を大きな値に設定するほど $Progress(s)$ の値は急激に増加する．$\Psi(s)$ の値が大きいほど，状態 s における学習は進行していると考えることが可能である．

4. シミュレーション

ここで検討しているそれぞれの指標が，エージェントの学習過程でどのように更新されるかを，予備実験としてのシミュレーションにより確認する．シミュレーションには，追跡問題の条件を拡張し，ゲームとして試作をしたものを用いて，エージェントに学習を行わせる．

4.1 シミュレーション条件

本稿で，2 章で述べたように，追跡問題 [BENDA 86] をヒントに，その条件を拡張したゲームを用いてシミュレーションを行う．本ゲームは 5×5 の格子状のフィールドを用い，追跡者を 2 体，追跡対象を 1 体用意した．それぞれの追跡者にエージェントを導入し，本ゲーム内の行動を学習させた．本稿で試作する機構に用いる指標について調べるために，一方の追跡者が観測した状態を計測した．計測した具体的な状態は図 3 に示した通りで，追跡対象は行動可能であるものとする．エージェントの学習にはそれぞれ Q 学習を用い，学習率 0.2，割引率 0.9 とし，探索には ϵ-greedy 法を用い，$\epsilon = 0.3$ とした．報酬には，終了条件を満たした時に追跡対象に攻撃をしていた追跡者に+10，捕獲行動をしていた追跡者に+5，行動をとるたびに-1 とした．20 ターンで 1 エピソードとし，合計 50000 エピソードを学習および計測をした．

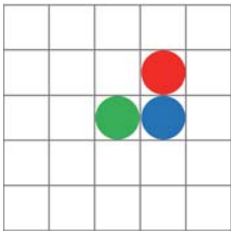

図 3: シミュレーションにより計測する状態例

4.2 シミュレーション結果

シミュレーションの結果を図 4 に示す．縦軸に学習の進行度をとり，横軸にエピソードをとっている．進行度の値が大きいほど，学習が進んでいることが期待される．項目「観測回数のみ」とは，シミュレーションで計測をする対象である状態において，その状態を観測した回数に基づいたその状態における学習の進行度を表現したものである．項目「Q 値のみ」とは，シミュレーションで計測をする対象である状態において，選択可能な行動それぞれに対する Q 値に基づいたその状態における学習の進行度を表現したものである．項目「ハイブリッド」とは，シミュレーションで計測をする対象である状態において，その状態を観測した回数と，その状態において選択可能な行動それぞれに対する Q 値の両方に基づいたその状態における学習の進行度を表現したものである．

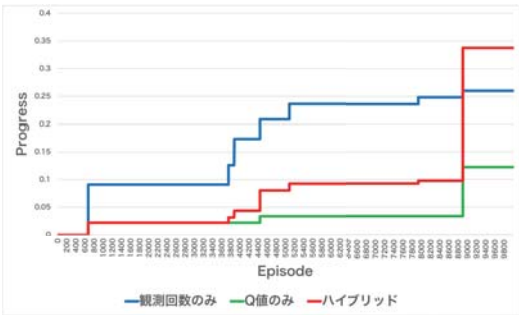

図 4: エピソード経過による三つの指標の値の遷移

計測した状態における各行動の Q 値の遷移を図 5 に示す．縦軸に Q 値をとり，横軸にエピソードをとっている．項目名に「移動」と表記されているものは，対応した移動方向に対しての Q 値を示しており，項目名に「攻撃」と表記されているものは，対応した攻撃方向に対しての Q 値を示している．項目「捕獲」は捕獲行動の Q 値を示している．表記されていない行動および攻撃は，計測する対象である状態においては選択不可能なため省略している．9000 エピソード以降，項目「下移動」の Q 値が大きく増加していることがわかる．一方で，図 4 の項目「観測回数のみ」は，9000 エピソードよりも前から増加し続けている．項目「ハイブリッド」は，9000 エピソード以降に大きく増加し，最も大きな値となっている．このことから，本稿で行ったシミュレーションの範囲においては，観測回数にのみに基づいて学習の進行度を表現した場合，その状態において十分に探索を行っていなくても大きな値をとってしまう恐れがあるということがわかる．一方で，観測回数と Q 値の両方に基づいて学習の進行度を表現した場合には，その状態における探索をどれほど行ったかを考慮した表現を実現することが可能であったのではないかと考えられる．

5. まとめと今後の研究課題

本研究では，学習中のエージェント対して人が知識を与える必要があるかどうかを容易に判断できるようにするために，学習の進行の度合いを数値化し，その数値に応じて表示が変化するような機構の試作を試みている．

本稿では，本研究で扱うマルチプレイヤーゲームを題材にした際に，強化学習においてエージェントの学習の進行度合いを数値化するために用いることが可能であると考えられる指標について検討をし，それを用いたエージェントへの教示支援

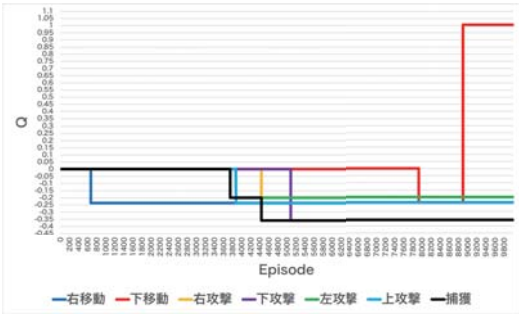

図 5: エピソード経過による Q 値の遷移

機構の試作を検討した．本機構に用いる指標として，各状況における学習の進行度合いを示すための三つの指標を検討した．検討したそれぞれの指標が，エージェントの学習を通してどのように更新されるかを，予備実験としてのシミュレーションにより確認した．

今後の課題としては，本稿で検討を進めているそれぞれの指標を用いることで，どれほど学習を効率的に行わせることが可能であるかを，本研究で対象とするゲーム上で調査することがあげられる．

参考文献

[Ariel Rosenfeld 17] Ariel Rosenfeld, S. K., Matthew E. Taylor: Leveraging Human Knowledge in Tabular Reinforcement Learning: A Study of Human Subjects, in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 3823–3830 (2017)

[BENDA 86] BENDA, M.: On Optimal Cooperation of Knowledge Sources : An Empirical Investigation, *Technical Report, Boeing Advanced Technology Center* (1986)

[Clouse 92] Clouse, J. A. and Utgoff, P. E.: A Teaching Method for Reinforcement Learning, in *Proceedings of the Ninth International Workshop on Machine Learning*, ML '92, pp. 92–110, San Francisco, CA, USA (1992), Morgan Kaufmann Publishers Inc.

[Knox 08] Knox, W. B. and Stone, P.: TAMER: Training an Agent Manually via Evaluative Reinforcement, in *IEEE 7th International Conference on Development and Learning* (2008)

[Silva 17] Silva, F. L. d. and Ruben Glatt, A. H. R. C.: Simultaneously Learning and Advising in Multiagent Reinforcement Learning, in *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 1100–1108 (2017)

[Thomaz 06] Thomaz, A. L. and Breazeal, C.: Reinforcement Learning with Human Teachers: Evidence of Feedback and Guidance with Implications for Learning Performance, in *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, AAAI'06, pp. 1000–1005, AAAI Press (2006)