ISSN 2186-5647

P-9

集中型マルチエージェント強化学習法における学習切り替え条件の最適化

日大生産工　○大野 拓哉　　日大生産工　山内　ゆかり

1. まえがき

機械学習の1種である強化学習には複数のエージェントが最適方策を学習しようとするマルチエージェント強化学習がある。マルチエージェント強化学習では各エージェントが他のエージェントの状態や行動を考慮し、最適方策を学習する。学習方法として、安定して最適方策を学習することが出来る集中型学習がある。

集中型学習とは、個々のエージェントが強調し合うことで得られる全体的な報酬の最大化を目指す学習方法である。しかし集中型学習は状態行動空間が大きくなるため学習に時間がかかるという欠点がある。そこで高速に学習を行えるように各エージェントが独立した環境で学習を行う分散型学習を集中型学習に組み合わせた学習法[1][2]が提案され、結果として通常の集中型学習より少ないエピソード数で最適方策を得られるという結果が得られている。

しかし従来研究では分散型学習をあるタイミングで停止する必要があり、停止基準となる閾値パラメータを予め準備する必要がある。閾値パラメータは最適方策を満たすような値で設定する必要があり、値が大きすぎると最適方策を学習する前に分散型学習を停止してしまい最適方策を得られない場合がある。そのため閾値を手動で設定する必要があった。

そこで提案手法として、閾値を設けず独立型学習器を停止させずに最適方策を得られる学習方法を提案する。

2. 従来研究

2.1 分散型学習と集中型学習

複数のエージェントが学習を行い最適方策を得るマルチエージェント強化学習において、学習方法にいくつかの種類がある。今実験では１つの学習器で全てのエージェントの方策を安定して学習することのできる集中型学習とエージェントごとに１つの学習器を用いて方策を学習する分散型学習の2種類の学習法を使用する。

集中型学習は１つの学習器で安定した最適方策を学習できる。その分行動状態空間が大きくなり、学習に時間がかかるという欠点がある。

分散型学習では、1エージェントにつき１つの学習器を与えられる。この学習器を独立型学習器と呼び、最適方策を得るのは集中型学習に比べ早いが他のエージェントの影響を受ける依存状態に対しては分散型学習では学習できない。以上のことからマルチエージェント学習では集中型学習を用いなければならない。

2.2 高速化手法

独立型学習器と集中型学習器を併用して学習させ序盤のみ独立型学習器のQ値を集中型学習器に転送することによって学習の高速化を計る。分散型学習では依存状態での最適方策を学習できない。そのため集中型学習器の最適方策の学習を乱してしまうため、序盤のみQ値を転送する。

独立型学習は上記の理由により、エージェントが共存しあう環境では前回の学習より良い方策が得られない場合があり、エピソードごとに集中型学習器に転送すると最適方策が乱れてしまう。そこで、Q値の転送は各学習器で良い方策を得られた場合のみ情報を転送する。良い方策を得られたかの判断はそれぞれの各学習器ごとに行われ、具体的な式は下記の(1)式で評価される。

$$E_i(t) = \sum_{s=1}^{L_i(t)} \gamma^{L_i(t)-s}\, r_{is}(t) \qquad (1)$$

$L_i(t)$はエピソードtにおける全行動回数,
$r_{it}(t)$はs回目の行動で得た報酬,γは割引率である。評価値$E_i(t)$は値が大きいほど評価が高いとされる。

転送された独立型学習器のQ値は集中型学習器のQ値となるが、集中型学習器は送られてきたQ値の平均をとって自身のQ値とする。しかし集中型と独立型の状態と行動の形式が異なるため、集中型学習器の形式に変換する必要がある。集中型学習器の各状態s,各行動aに対して変換後のQ値$Q_{IL}^{i}(s,a)$は独立型学習器の各エージェントiのQ値$Q_i(s,a)$から下記の(2)式で変換される。

$$Q_{CL}(s,a) = Q_i(IS_i(s), IS_i(a)) \qquad (2)$$

またQ値の転送は、上記で述べた通り良い方策が得られた場合のみ転送され、集中型学習器

Optimization of Learning Switching Conditions in Centralized Multi-Agent Reinforcement Learning Methods
Takuya OHNO and Yukari YAMAUCHI

では分散型学習器から転送されたQ値を用いてQ値Q_{CL}が下記の式(3)(4)で更新される。

$$Q_{CL}(s,a) = \frac{Q_{CL}(s,a) + \sum_{i=1}^{I} F_i(t) \cdot F_i(t) \cdot Q_{IL}^{i}(s,a)}{\sum_{i=1}^{I} F_i(t) + 1} \quad (3)$$

$$F_i(t) = \begin{cases} 1 & E_i \geq E_{max}^{i}(t) \\ 0 & otherwise \end{cases} \quad (4)$$

独立型学習器では依存状態に対する最適方策を作成できない。したがって独立型学習器を動かし続けると集中型学習器における最適方策の学習が妨害されてしまう。よって障害が生じる前に独立型学習器を停止させる必要がある。

2.3 独立型学習機の停止

独立型学習器の停止判断にはTD誤差が用いられ、TD誤差は学習の進捗具合を確認できる。Q学習のTD誤差は下記の(5)式で求められる

$$TD_i(s_i, a_i) = r_i + \gamma max Q_i(s'_i, a) - Q_i(s_i, a_i) \quad (5)$$

TD誤差は学習が進むにつれて0に近づく。この性質を利用し、TD誤差の絶対値$|TD_i|$がある閾値TD_{lim}以下に収まるときエージェントiの独立型学習器が停止する。しかしTD誤差は状態、行動の数だけ存在しまだ経験していない状態、行動のTD誤差は算出されない。そこでエピソード中に経験した状態、行動のTD誤差を算出し、その絶対値の最小値TD_{lim}が閾値以下になった場合独立型学習器が停止される。

3. 提案手法

3.1 問題点の改善

従来研究では閾値パラメータを予め手動で設定する必要があり、且つ閾値パラメータの値が大きすぎると独立型学習器が最適方策を得る前に停止してしまうという欠点があった。そこで本研究では閾値パラメータを用いず、目標座標に辿り着いたステップ数に応じて、Q転送するQ値を調節する手法を提案する。

3.2 Q値の転送方法

従来研究では序盤の学習では独立型学習器の全てのエージェントが目標座標に達した場合、正の報酬与え、エージェントが一つでも目標座標に達していない場合は全てのエージェントに負の報酬与えていた。しかしそれでは目標座標に達したエージェントの行動も負の報酬を受け取ることになってしまう。そこで本実験ではゴールしたエージェントには正の報酬を与え、全てのエージェントが正の報酬を受け取った時、分散型学習から集中型学習にQ値を転送する。

また従来ではTD誤差を利用して独立型学習器からのQ値の更新を停止していたが本実験では最後まで独立型学習器からのQ値の転送は停止せず、目標座標に達するまでにかかったステップ数により従来の式(4)の式を下記の式(6)で更新する。

$$F_i(t) = \begin{cases} 1 & E_i > E_{max}^{i}(t) \\ 0.5 & E_i = E_{max}^{i}(t) \\ 1 - \frac{t}{T} & E_i < E_{max}^{i}(t) \\ 0 & otherwise \end{cases} \quad (6)$$

上記の式にすることにより、分散型学習の欠点であったエージェントが互いに影響し合う依存状態に対して最適方策を得られなかった場合は、転送する値を小さくすることにより集中型学習器の学習を大幅に乱すことなく、Q値の転送ができる。且つ、学習の最後まで独立型学習器を停止させずに済むため、問題点であった閾値を事前に適当な値で設けなければいけないという点は改善できると考えられる。

4. 実験

本実験では、各エージェントは上下左右の1マス移動及び停止の中から一つの行動a_iを選択し、状態遷移は確定的である。エージェントは番号順に行動し、他のエージェントがいる座標や壁を越えての行動はできず、そのような行動をした場合はその場に留まるものとする。また、一部のエージェントのみが目標座標に達成したとしても、全エージェントが目標座標に達するまで目標座標に留まるものとする。

各エージェントは$\varepsilon - greedy$法で行動を選択し、確率εでランダムな行動を選択する。またQ学習でQ値を更新し、各学習における割引率は$\gamma = 0.95$,エピソード総数は$T = 20000$,と1エピソードの最大行動回数は1000回とする。

5. まとめ

本研究では、独立型学習器からの転送を最後まで停止せずに、より良い学習が出来たかの判断を行い、転送するQ値の値を調節する提案をした。それにより閾値を事前に設けなければいけないという問題点の改善を試みた。しかし、本実験ではエージェント数を2つで行ったが、エージェント数が増えれば増えるほど状態の空間が大きくなり学習の時間が増えてしまう。そこでDeep Q-Networkなどを導入してQTable状態から改善する必要がある。

参考文献

[1] 赤羽根拓真, 飯間等, 集中型マルチエージェント強化学習の高速化, 電気学会論文誌, Vol.140, No.2, pp.242-248

[2] 佐々木薫, 飯間等, 報酬設定を自動化した集中型高速マルチエージェント強化学習法, システム制御情報学会論文誌, Vol.35, No.3, pp.39-47,2022