ISSN 2186-5647

2-45

アンサンブル学習型マルチエージェント強化学習

日大生産工(院)　○高萩 悠　日大生産工　山内 ゆかり

1. まえがき

強化学習 1)という学習手法がある。この手法は学習環境内のエージェントが行動選択を繰り返し、行動の結果最終的に獲得できる報酬を最大化する手法である。近年の研究ではマルチエージェントによる相互作用を用いた学習手法が主流になっている。マルチエージェント強化学習 2)には状態遷移の不確定性、不完全知覚問題など、学習に悪影響を及ぼす問題点がある。これらにより、学習収束が遅くなる、学習後の行動精度が高くならないなどの影響がある。

一方で複数の学習モデルを組み合わせ、一つの学習モデルを生成するアンサンブル学習 3)という手法がある。アンサンブル学習によって得られる制御や予測は、単一の学習器よりも高い精度を出すことが知られている。

本研究ではアンサンブル学習による学習能力の向上効果によってマルチエージェント強化学習の学習精度の向上を目指す。

2. 従来手法

2.1 強化学習

強化学習1)とはSuttonらが提唱した学習方法である。ある環境内におけるエージェントに行動選択をしてもらい、最終的に獲得する報酬を最大化する学習手法である。有名な手法にQ学習やTD学習がある。Q値の更新や状態価値の更新をする際には次の時間の行動価値や状態価値を使用する。

$$\Delta V(s) = r' + \gamma V(s') - V(s) \qquad (1)$$

式(1)はTD学習での状態価値の更新式である。これはTD誤差とも呼ばれる。$V(s)$が状態sでの状態価値であり、r'はその状態の報酬であり、γは割引率である。s'は次の時間の状態である。

TD学習の中で有名なものにActor-Criticといわれる学習手法がある。

2.1.1 Actor-Critic

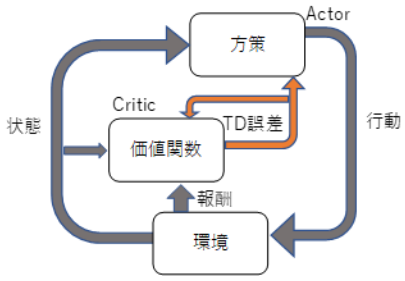

Figure 1　Actor-Critic のモデル

Figure1はActor-Critic4)のモデルを図にしたものである。Actorは方策によって行動を選択し、Criticは状態価値関数に応じて方策を修正する。一般的なActor-Criticは修正量としてTD誤差を利用する。TD誤差は先ほどの式(1)のものである。

$$V(s) \leftarrow V(s) + \alpha \Delta \mathrm{V(s)} \qquad (2)$$

式(2)はCriticによる評価値の更新式である。αは評価値の学習率であり、式(1)のTD誤差を用いて評価値を更新する。また、行動選択方策Actorも更新する必要がある。Actorは行動選択確率$p(s,a)$と行動選択方策によって行動を選択する。

$$p(s,a) \leftarrow p(s,a) + \beta \Delta V(s) \qquad (3)$$

式(3)はActorの更新式である。βはActorの行動選択確率の学習率である。Criticとは独立にそれぞれ学習するので別の学習率を用いる。

2.1.2 Q-learning

Q-learning5)とはQ値といわれる行動価値を更新していく学習手法である。

$$Q(s,a) \leftarrow Q(s,a) + \alpha(r' + \gamma \max_b Q(s',b) - Q(s,a)) \qquad (4)$$

式(4)はQ-learningにおけるQ値の更新式である。sは現在の状態を示しており、aは行動を

Multi-Agent Reinforcement Learning with Ensemble Learning

Yu TAKAHAGI and Yukari YAMAUCHI

示す。$Q(s,a)$は状態sにおける行動aを選択した際のQ値を示しており、Q-learningではこのQ値を更新する。s'は次の状態を示しており、r'は次の状態での報酬を示す。$\max_b Q(s',b)$は次の状態での最大のQ値を示す。γは学習率である。つまりQ-learningは現在の選んだ行動価値を状態遷移後の行動価値と報酬で学習する。

2.2 アンサンブル学習

アンサンブル学習3)とは複数の弱い学習器を組み合わせて、学習精度の向上を図る手法である。組み合わせ方にはいくつかの種類がある。その中でもスタッキングといわれる手法がある。

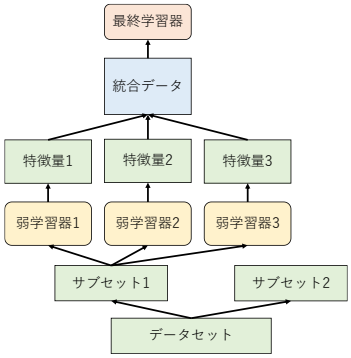

Figure 2　スタッキングのモデル

スタッキングとはその名の通り複数のモデルを積み上げる手法である。Figure2のように学習元となるデータセットをいくつかのサブセットに分け、その分けられたデータセットで学習を一旦行う。その後サブセットで得られた弱学習器を使い特徴抽出を行い新たなデータセットを再度作る。最後にその得られたデータセットを用いて再度学習を行う。スタッキングは一度学習器が出した予測値や特徴量などの出力結果を再度学習に利用するのが特徴的である。

3. 提案手法

本研究では、マルチエージェント強化学習の学習精度向上のためにアンサンブル学習の考え方を取り入れる。

マルチエージェント学習では他のエージェントの状態を全て把握すると状態空間が膨大になり計算コストが増大する。そこで提案手法では、一部のエージェントの状態のみ知覚することで、状態数を削減する。ただし不完全知覚が起きるので、学習は弱学習とみなされる。また、他のエージェントの出力情報である選択行動を知覚することで、スタッキング学習の効果が得られることを目指す。

4. 実験および検討

実験は7×7のトーラス状Grid World空間において追跡問題を行う。トーラス状とはドーナツのようにGrid Worldマップの上下左右の端が反対側につながっている状態を指す。

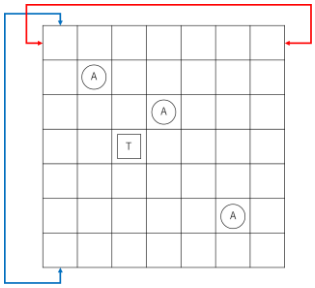

Figure 3　実験環境のイメージ

Figure3は実験環境のイメージである。Aがターゲットを追いかけるエージェントを示しており、Tが追跡されるターゲットである。本実験の環境は3エージェント1ターゲットである。

5. まとめ

本研究ではマルチエージェント強化学習にアンサンブル学習を取り入れることで学習精度が向上するかを試みる。しかし、現段階では提案手法が実装できていない。

参考文献

1) Richard S. Sutton and Andrew G. Barto, “Reinforcement Learning An Introduction” MIT Press (1998)
2) 荒井幸代, 宮崎和光, 小林重信, マルチエージェント強化学習の方法論 -Q-LearningとProfit Sharingによる接近- 人工知能学会誌 Vol.13 No.4 pp.609-618 (1997)
3) 上田修功, アンサンブル学習 コンピュータビジョンとイメージメディア, Vol 46, No SIG 15(CVIM 12), pp11-20 (2005)
4) Vijay R. Konda and John N. Tsitsiklis “Actor-Critic Algorithms” Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge (2000)
5) Watkins, C. J. C. H. “Learning from Delayed Rewards” King’s College Thesis Submitted for Ph.D. (1989)