深層強化学習を用いた取引戦略の学習

1　はじめに

近年，機械学習の急速の発展に伴い，深層強化学習による株取引が注目を集めている．特に，深層 Q 学習 (Deep Q-Network: DQN) の基本的な拡張手法を 6 種類組み合わせた手法である Rainbow を用いた株取引の研究 [1] もなされている．

そこで本実験では，DQN を用いて Open AI Gym の CartPole-v0 を実装しデータ解析を行う予備実験を行う．(→ DQN を用いた株取引) を行いデータ解析を行う．

2　要素技術

2.1　Open AI Gym

OpenAI とは人工知能を研究する非営利企業であり，Gym はその企業が作成した強化学習のシミュレーション用プラットフォームである．

2.2　Deep Q-Network

Deep Q-Network(DQN) は，代表的な強化学習手法である Q-Learning を用いた深層強化学習である．DQN では深層強化学習に基づく Q-Network と呼ばれる，強化学習における価値に相当する Q 値を多層ニューラルネットにより近似する．Q-Network の更新には replay memory と呼ばれる状態の繊維を経験として蓄積したものを利用する．

2.2.1　Q-Network

Q-Network とは，Q 値を求める多層ニューラルネットで，従来の状態と行動を入力として一つのスカラー値を返す方法とは異なり，状態のみを入力とする．出力層においては各行動ごとの Q 値を持つ．これにより，一度の入力に基づくニューラルネットの出力計算により，全種類の行動の Q 値が得られるため，行動数によって計算量が増えることがほとんどないという利点を持つ．

2.2.2　Experience Replay

多くの場合，深層強化学習アルゴリズムは i.i.d, つまり学習データに相関が無いことが想定されているが，強化学習では問題の性質上時間的に偏ったデータになりがちである．Experience Replay とは，過去の遷移情報を十分な数保存し，そこからランダムサンプリングすることで，擬似的にデータの方よりをなくす工夫である．この経験の蓄積を replay memory と呼ぶ．遷移情報は「状態 s_t で行動 a_t を選択したところ，報酬 r_t を獲得し，次の状態が s_{t+1} であった」場合，これらを含む 4 つの組から構成される $(s_t,\ a_t,\ r_t,\ s_{t+1})$ を経験した順に記憶し続ける．設定したメモリの上限を超える場合は最も古い経験から破棄する．

2.2.3　Q-Network の更新

十分に replay memory にデータを蓄えられたら，replay memory からランダムサンプリングし，以下の式に従って Q-Network を更新する．

$$Q_\theta(s_t,a_t) \leftarrow (1-\alpha)Q_\theta(s_t,a_t)+\alpha(r+\gamma \max_{a_{t+1}} Q_\pi(s_{t+1},a_{t+1})) \tag{1}$$

ここで $Q_\theta(s_t,a_t)$ はパラメータ θ を持つニューラルネットワークであり，$Q_\pi(s_t,a_t)$ は教師信号出力用のニューラルネットで，$Q_\theta(s_t,a_t)$ のコピーになっている．$\max_{a_{t+1}} Q_\pi(s_{t+1},a_{t+1})$ は遷移先の状態 s_{t+1} における最大の Q 値，α は学習率 $(0 \leq \alpha \leq 1)$, r は報酬，γ は割引率 $(0 \leq \gamma \leq 1)$, t は時刻である．

3　提案手法

4　数値実験

5　結果と考察

6　おわりに

7　参考文献

参考文献

[1] テスト