時系列を考慮した複利型深層強化学習を用いた金融取引戦略獲得

Compound deep reinforcement learning to acquire trading strategies with time series

近藤 巧麻 [1*]　松井 藤五郎 [2,3]
Takuma Kondo[1]　Tohgoroh Matsui[2,3]

[1] 中部大学 大学院工学研究科 情報工学専攻
[1] Department of Computer Science, Graduate School of Engineering, Chubu University
[2] 中部大学 生命健康科学部 臨床工学科
[2] Department of Clinical Engineering, College of Life and Health Sciences, Chubu University
[3] 中部大学 工学部 情報工学科
[3] Department of Computer Science, College of Engineering, Chubu University

Abstract: 本論文では時系列を考慮したニューラルネットワークを用いることでより性能のよい金融取引戦略を獲得する複利型深層強化学習の方法を提案する。従来研究では全結合層のみから特徴を抽出しており、それより過去の情報は前日のテクニカル変数のみでしか反映されていない。そこで本論文では時系列データを学習することのできる LSTM や CNN を用いてより過去の情報を考慮した特徴から金融取引戦略を獲得する。また、TOPIX を対象とした取引に提案手法を適用し、その有効性を確認する。

1 はじめに

近年、機械学習は発達に伴って、様々な場面で使用されている。その中でも強化学習を用いて金融取引戦略を学習する研究が行われている [1, 2, 3]。

先行研究 [2] では、一般的に使われている深層強化学習を複利型に拡張した複利型深層強化学習を用いて取引戦略を獲得する手法が開発されてきた。この研究では、実際に深層強化学習を複利型に拡張することの有効性を確認することが出来ている。

しかしながら、実際の取引現場ではその日の状況だけで判断することはまずありえなく、チャートなどから取引当日までの過去データを考慮しながら判断をしていることが多い。先行研究では単純なニューラルネットワークで特徴量を獲得しているため、その日の状況のみから売買の判断をしている。そのため、長期的なトレンドなど過去のデータを考慮したうえでの状態が反映できていないと考えられる。

一方、時系列性を考慮して株価を予測する際に使われている深層学習の手法としては、LSTM [4] や CNN [5] などが挙げられる [6]。

そこで、本研究では、時系列性を考慮して深層複利型強化学習の特徴を抽出する手段として、LSTM、畳み込み、プーリングを用いることを提案する。また、これらの手法を従来の複利型深層強化学習と比較することで、提案手法の有効性を確認する。

*E-mail: eptkm410@gmail.com

2 従来手法

2.1 複利型 PPO

方策ベースの深層強化学習である PPO [7] は、時刻 t における状態 S_t を入力すると行動選択確率 $\pi(S_t, a)$ と状態の価値 $V(S_t)$ を出力するネットワークを学習する。PPO は、状態 S_t における行動 A_t のアドバンテージ $A(S_t, A_t)$ を次のように計算する。

$$\begin{aligned} A(S_t, A_t) &= Q(S_t, A_t) - V(S_t) \\ &= R_{t+1} + \gamma V(S_{t+1}) - V(S_t) \end{aligned}$$

ここで、$Q(s, a)$ は状態 s における行動 a の価値、$V(s)$ は状態 s の価値、R_{t+1} は S_t において A_t を実行した結果得られた報酬、γ は割引率を表す。

状態 S_t における行動 A_t の価値 $Q(S_t, A_t)$ を、それを実行した結果得られた報酬 R_{t+1} と次の状態の価値 $V(S_{t+1})$ に割引率 γ をかけたものの和によって推定し、

そこから状態 S_t の価値 $V(S_t)$ を引くことによってアドバンテージを求めている。$V(s)$ は、状態 s における行動の平均的な価値の見積と考えることができることから、アドバンテージ $A(s,a)$ は、状態 s における行動 a の価値 $Q(s,a)$ が状態 s における行動の平均的な価値をどのくらい上回っているかを表している。

複利型強化学習 [1] は、報酬の代わりに利益率を観測し、利益率の複利効果を最大化する強化学習である。先行研究 [3] では、PPO を複利型に拡張し、金融取引戦略を学習している。複利型 PPO では、アドバンテージを次の式に置き換えることで、PPO を複利型に拡張している。

$$A(S_t, A_t) = \log(1 + R_{t+1}f) - \gamma V(S_{t+1}) - V(S_t)$$

ここで、R_{t+1} は状態 S_t において行動 A_t を実行した結果得られた利益率、f は投資比率を表す。利益率 R_{t+1} は、時刻 t における資産価格 p_t と時刻 $t+1$ における資産価格 p_{t+1} から次のように求められる。

$$\begin{aligned} R_{t+1} &= \frac{p_{t+1} - p_t}{p_t} \\ &= \frac{p_{t+1}}{p_t} - 1 \end{aligned}$$

2.2 LSTM

LSTM [4] は、時系列性を有するデータを学習できるニューラルネットワークであるリカレントニューラルネットワーク (RNN) の一種である。

RNN は隠れ層が隠れ層自身につながっており、ある時点での状態を次の状態の入力値として使うことが出来る。つまり、過去の情報を保存し、順に処理することが出来る。しかし、RNN は系列が長くなると勾配消失問題が発生してしまうという問題点がある。

これに対し、LSTM では全状態を次の状態へ入力するのではなく各種ゲートやメモリで制御しながら次の状態へ入力するため、系列が長くなっても問題なく学習できるようになっている。

2.3 畳み込みニューラルネットワーク (CNN)

CNN [5] は、画像のような 2 次元データにおいて空間的な情報を維持したまま処理を行うことができるネットワークである。通常の全結合のみでのネットワークとは違い、前層のニューロンと結合している領域を局所的にすることができる。その結果、データのずれなどに頑健になるという特徴がある。

本論文では、時系列データをたたみ込むために、1 次元畳み込みとプーリングを導入した。

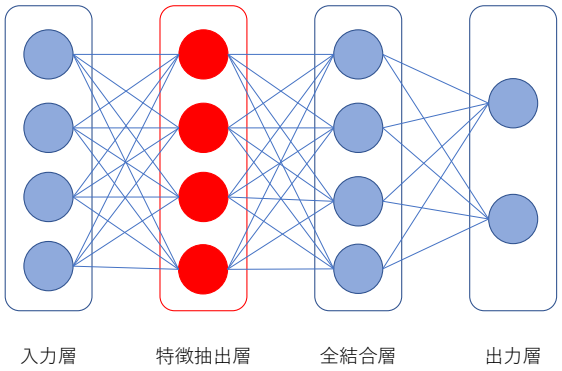

図 1: モデル構造

画像のような 2 次元畳み込みにおいては、対象となるピクセルの周囲のピクセルの特徴を畳み込んで対象となるピクセルの特徴とするが、将来を予測するの時系列データにおいては、対象となる時刻より後の時刻の情報を用いることができない。そこで、対象となる時刻よりも前の時刻の情報を畳み込んで対象となる時刻の特徴とする。

プーリングは、一定領域の特徴から最大値や平均値など一つの特徴を抽出するもので、画像のような 2 次元プーリングにおいては特徴の位置ずれを吸収することができる。時系列データにおいては、対象となる時刻以前の一定領域から一つの特徴を抽出することによって、時間のずれを吸収することができる。

3 提案方法

本論文では、従来手法 [3] では全結合層のみとなっていた特徴抽出層を、LSTM または CNN に置き換えたモデルを 3 種類提案し、その有効性を比較する。提案手法のモデル構造を図 1 に示す。

一つ目のモデルは、特徴抽出層を LSTM に置き換えたものである。その時刻の価格やテクニカル指標だけでなく、その時刻以前の一定期間の価格とテクニカル指標の値を入力とし、LSTM を用いて特徴を抽出し、全結合層と出力層を通して行動 a の選択確率 $\pi(S_t, a)$ と状態価値 $V(S_t)$ を出力する。

二つ目のモデルは、特徴抽出層を 1 次元畳み込み層に置き換えたものである。LSTM に置き換えた一つ目の方法と同様に、その時刻以前の一定期間の価格とテクニカル指標の値を入力とし、1 次元畳み込み層を用いて特徴を抽出する。

三つ目のモデルは、1 次元畳み込み層にプーリング層を加えたものを用いて特徴を抽出する。プーリング

図 2: TOPIX 期間 1

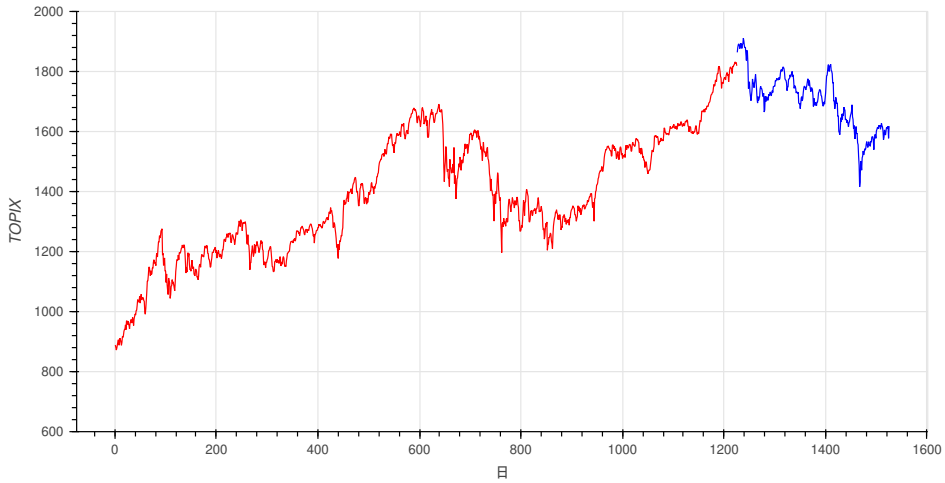

図 3: TOPIX 期間 2

層を加えることによって、時間のずれを吸収することが期待できる。本論文では、プーリングの方法として max pooling を用いた。

4 実験

4.1 実験方法

東証株価指数 (TOPIX) のデイリー取引を対象として提案手法の有効性を確認する実験を行った。

図 2 と図 3 に示す 2 つ期間を用いてデータセット作成した。期間 1 は、訓練期間を 2012 年 1 月 1 日から 2016 年 12 月 31 日までの 5 年間、テスト期間を 2017 年 1 月 1 日からの 300 日分とした。期間 2 は、訓練期間を 2013 年 1 月 1 日から 2017 年 12 月 31 日までの 5 年間、テスト期間を 2018 年 1 月 1 日からの 300 日分とした。

期間 1 は訓練期間とテスト期間の両方が全体的に上昇傾向となっているが、期間 2 は訓練期間は全体的に上昇傾向であるがテスト期間は下落傾向となっている。

強化学習の状態として、次の 10 個のテクニカル指標を状態変数とした。

1. 移動平均 (MA)
2. 移動標準偏差 (MSD)
3. 相対終値 (RCP)
4. 移動標準偏差の移動平均 (MSDMA)
5. 移動標準偏差の標準偏差 (MSDSD)
6. 相対移動標準偏差 (RMSD)

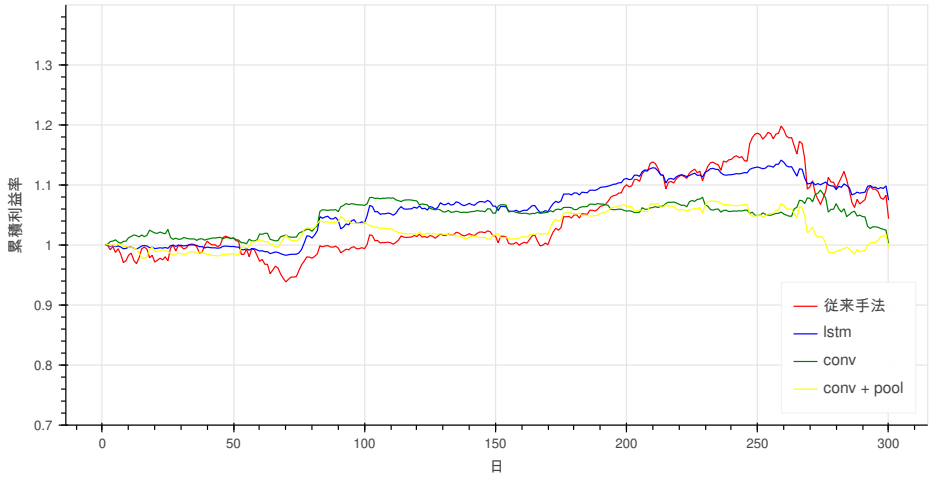

図 4: 期間 1 の平均累積利益率の推移

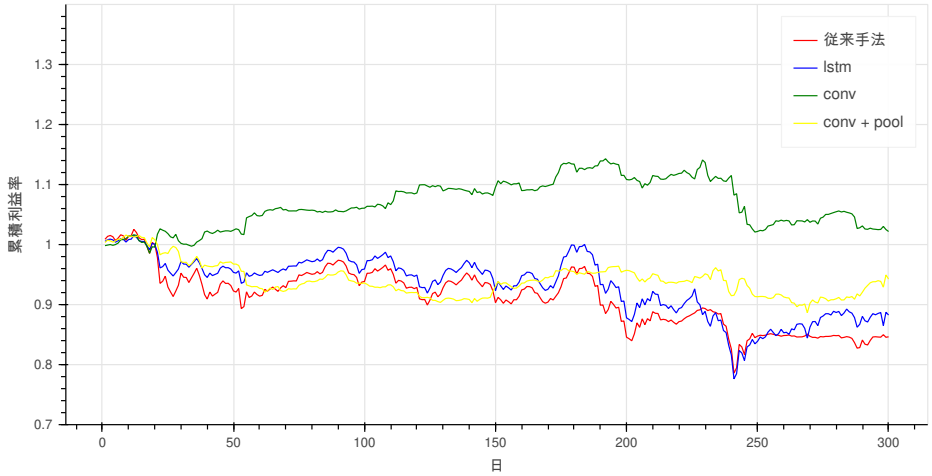

図 5: 期間 2 の平均累積利益率の推移

7. ボリンジャーバンドの上側
8. ボリンジャーバンドの下側
9. ROC (rate of change)
10. モメンタム

前処理として状態変数はそれぞれ 0 から 1 になるようにスケーリングしている。

強化学習の行動は、購入と売却の 2 種類とした。行動が購入のときは当日の始値で購入して終値で売却し、売却の時は当日の始値で信用売りして終値で買い戻す。

実験では複利型 PPO を使用し、それぞれ 500 エピソード分学習を行った。また、モデルを 3 か月ごとに更新し、これを 5 回行って一年間の利益率の平均を比較した。

4.2 実験結果

図 4 に期間 1、図 5 に期間 2 の結果を示す。横軸がテスト期間の経過日数、縦軸は 5 回の平均累積利率を表している。

図 4 では、全ての手法が 260 日あたりまでは上昇した。その中でも、LSTM (lstm) と畳み込みを用いた手法 (conv) は、図 2 の約 1,500 日目以後の急落に対しても他の二つの手法に比べて平均累積利益率の下落が小さかった。

図 5 では、畳み込みを用いた手法のみが平均累積利益率を伸ばし、それ以外の手法は 1 未満に落ち込んでしまった。また、LSTM と従来手法は、260 日までは平均累積利益率がほとんど同じ推移となった。

5 考察

訓練期間とテスト期間が共に全体的に上昇傾向であった期間 1 において、LSTM を使った手法が安定的に平均累積利益率を伸ばしていることから、LSTM の特徴である長期的なトレンドの特徴抽出がうまくいったのではないかと考える。

しかしながら、期間 2 においては、LSTM を使った手法はどのような状況でも同じ行動を選択するモデルを学習してしまうことがあった。訓練期間とテスト期間の全体的な傾向（長期トレンド）が似ている期間 1 では LSTM を使った手法がうまくいったことから、訓練データとテストデータの長期的トレンドが違うことが原因があると考える。

畳み込みを使った手法 (conv) に関しては、基本的に 5 回学習したモデル全てでバラバラの行動を選択することが多かった。このことから、畳み込みによって様々な特徴マップが作られ、他の手法に比べ毎回異なった多くの特徴が抽出でき、その結果多様なモデル作成に貢献していると考える。

畳み込みとプーリングの両方を使った手法 (conv + pool) に関しては、今回の環境においてはあまり効果的でないことがわかった。これは、畳み込みだけを使った手法 (conv) は効果的だったことと今回使ったプーリングが max pooling だったことから、プーリングに max pooling を使ったことが原因だと考えられる。max pooling では一定期間内の値が最大のものを抽出する。金融取引においては、値が小さいとことも重要な特徴であるため、最大のものだけを抽出することで重要な特徴を落としてしまっている可能性がある。深層学習による時系列予測についての実験的レビューを行った文献 [6] でも、畳み込みニューラルネットワークで精度を向上させるにはより多くのレイヤーを必要とするが、max pooling は必要ではないと結論付けている。

6 まとめ

本論文では時系列データの特徴を抽出できるニューラルネットワークを使用して学習を行う手法を提案した。今回は特徴抽出部分に LSTM、畳み込み、畳み込みとプーリングを組み合わせたものを用いた。

従来手法に比べてモデルの学習が安定している点、平均累積利益率が良い点を考えると、本論文の提案手法は従来手法に比べて優れていると言える。

ただし、本論文の提案手法では、LSTM, プーリングを使った手法で問題点があった。LSTM は長期的なトレンドが考慮された特徴を抽出できていることが分かったが学習がうまくいかない場合も多かったため、原因を調べる必要がある。プーリングに関しては、max pooling だけでなく min pooling を用いることやプーリングする前の特徴を同時に入力することなどを検証する必要がある。

参考文献

[1] 松井 藤五郎, 複利型強化学習—強化学習のファイナンスへの応用—. 計測と制御, Vol.52, No.11, pp.1022–1027, 2013

[2] 松井 藤五郎, 片桐 雅浩. 金融取引戦略獲得のための複利型深層強化学習. 第 16 回人工知能学会金融情報学研究会 (SIG-FIN), SIG-FIN-016-01, 2016

[3] 近藤 巧麻, 松井 藤五郎. 複雑な環境における複利型深層強化学習を用いた金融取引戦略獲得. 第 34 回人工知能学会全国大会 (JSAI 2020), 2J4-GS-2-03, 2020

[4] Sepp Hochreiter, Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, Vol. 9, No. 8, pp. 1375–1780, 1997.

[5] Hoo-Chang Shin, Holger R. Roth, Mingchen Gao, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, Vol. 35, No. 5, 1285–1298, 2016.

[6] Pedro Lara-Beítez, Mannuel Carranza-García, José C. Riquelme. An experimenta review on deep learning architectures for time series forecasting. *International Journal of Neural Systems*, Vol. 31, No. 3, 2130001, 2021.

[7] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, Oleg Klimov. Proximal policy optimization algorithms. *arXiv*:1707.06347, 2017