

NT 倍率取引における深層強化学習を用いた投資戦略の構築

常井祥太^{†1} 穴田一^{†1}

概要: 近年、人工知能を用いた投資戦略に関する研究が行われている。しかし、株価や為替には多くの変動要因があるため、人工知能による適切な投資戦略の構築は困難である。そこで本研究では、日経 225 先物と TOPIX 先物のような相関性の強い 2 つの金融商品に対して、売りと買いの両方を行う取引方法を考える。これらは概ね同じような値動きをし、価格差が拡大しても元に戻りやすい性質があり、この価格差から利益を狙うことができる。また、この取引では株価で考えられるような価格変動要因の大部分が相殺されるため、価格差のみに着目した取引が可能になる。このように本研究では、この取引によって状況を簡略化した上で、深層強化学習によって投資戦略を獲得するモデルを構築した。そして、現実の市場における過去の価格データを用いて本モデルの有用性を確認した。

キーワード: NT 倍率取引, 深層強化学習, ディープラーニング, 投資戦略, 人工知能

Trading System using Deep Reinforcement Learning

SHOTA TOKOI^{†1} HAJIME ANADA^{†1}

Abstract: In recent years, investment strategies using artificial intelligence have attracted a significant amount of research attention. However, it is difficult to construct an efficient investment strategy using artificial intelligence owing to the variable factors in market prices. Therefore, this study aims to focus on a trading method called the NT ratio transaction to reduce the number of price-variable factors. This transaction is an arbitrage transaction, which utilizes the difference in the price movements between Nikkei 225 futures and TOPIX futures. These futures generally exhibit similar price movements and even if the price differences expand, they tend to return to their original separation. Using this transaction, we can target profits from this price difference while offsetting a considerable number of price-variable factors. Therefore, in this study, we construct a model to acquire an investment strategy based on NT ratio transactions via deep reinforcement learning and confirm the effectiveness of this model.

Keywords: Deep learning, Artificial intelligence, Deep reinforcement learning

1. はじめに

近年、人工知能に関する研究が活発に行われている。そのような中で、金融分野でも人工知能を用いた投資戦略の研究が行われている。松井らは複利型強化学習という新たな強化学習の枠組みを提案した。複利型強化学習とは、試行錯誤を通じてエージェントが将来獲得する報酬ではなく、複利式のリターン（得た利益を掛け金に乗せして得るリターン）を最大化する行動規則を学習する枠組みである[1]。また、彼らは複利型強化学習における行動価値関数をニューラル・ネットワークで表した複利型深層強化学習を提案した。この手法で、日本国債の週次取引における行動規則を学習し、利益率が向上していく様子が確認できた[2]。しかし、最終的な利益率を見ると、学習が十分であるとは言い難い。これは国債や株価などには価格変動要因がかなり多く存在し、それらを十分に考慮できていないことが原因であると考えられる。しかし、これらは各国のニュースによる変動への影響など定量化が困難なものが多い。そこで、本研究では価格変動要因を減らすため、NT 倍率取引という取引手法に着目する。NT 倍率取引とは、日経 225 先物

と TOPIX 先物の値動きの違いを利用した裁定取引である。これらのような相関性の強い 2 つの金融商品に対して「買い」と「売り」をそれぞれ同時に行うことにより、価格の変動要因の大部分が相殺できるため、2 つの価格差のみに着目した取引が可能になる。また、松井らの手法では状態変数が 2 つと少なく、多数の状態変数を扱える深層強化学習の利点を活かし切れていない。そのため、状態変数を増やすことで、現在の状況を適切に捉えた上で、より良い投資行動を行えるようになるのではないかと考えた。以上のことを踏まえた、NT 倍率取引における投資戦略を、深層強化学習によって獲得する数理モデルを構築し、その有用性を確認した。

2. 提案手法

本研究はコンピュータシミュレーションによって行う。コンピュータ上につくられた仮想的な投資家が、1 日 1 回市場の状態を観測し、その状態におけるそれぞれの投資行動の価値 (Q 値) を推測する。その価値が高い行動を選択し、実行し、結果が良ければその行動に報酬を与えて、同じ状態においてその行動をとりやすくする。この Q 値の推測は

^{†1} 東京都市大学
Tokyo City University

ニューラル・ネットワークを用いて行い、報酬に応じてその重みを変えることを繰り返して学習を進めていく。

2.1 既存手法からの変更点

本研究では、松井らの手法[2]をベースに総資産の最大化を目的として、以下の点を変更した。

(1) 取引手法

松井らの手法では、日本国債の週次取引に対する行動規則を学習した。しかし、国債には多くの価格変動要因が存在し、適切な行動選択を困難にしている。これらの変動要因をすべて取り入れて行動を選択することは不可能である上、多くの場合取り入れていない要因からも大きな影響を受けるため、安定した学習ができなくなってしまう。そこで、まず「考慮しなければならない価格変動要因を減らし、状況を簡略化すること」を考えた。具体的には、相関性が強く、価格差が拡大しても元に戻りやすいような2つの金融商品に対して、「買い」と「売り」をそれぞれ同時に行う裁定取引を考える。これにより価格変動要因の大部分を相殺可能である。このような相関が強い金融商品として、日経225先物とTOPIX先物がある。この2銘柄に対して「買い」と「売り」をそれぞれ同時に行う取引をNT倍率取引という。日経225先物とTOPIX先物の価格の推移を図1に示す。



図1 日経225先物とTOPIX先物の価格推移

図1の横軸は期間、縦軸は価格である。日経平均株価とTOPIXには約10倍の違いがあるため、この図ではTOPIXに10をかけたものをプロットしている。これを見ると、変動の仕方がかなり似通っていることが分かる。これは、日経平均株価とTOPIXがどちらも東証一部上場企業の株価や時価総額から計算される指標だからであり、変動の仕方がわずかに異なるのは計算に用いられている企業や、株価か時価総額かの違いによるものである。このように、定量化が困難な各国のニュースなどの影響の大部分はどちらも等しく受けており、2銘柄の価格の違いに着目した投資判断を行うことによって、価格変動要因の大部分が相殺された状態での取引が可能になる。そこで本研究では、NT倍率取引を取引手法として選択した。

(2) 学習方法

松井らの手法では、取引量を調節しながら利益率の複利効果を最大化するため、投資比率と複利リターン[2]を考慮した学習を行っている。しかし、本研究ではモデルを単純化するため、取引を1単位ずつの売買もしくはポジションの解消に制限した。

(3) 行動

本研究では行動として「1単位NT買い(日経225先物買い, TOPIX先物売り)」、「1単位NT売り(TOPIX先物買い, 日経225先物売り)」、「NT買いポジション解消」、「NT売りポジション解消」、「何もしない」の5つとする。ここで、日経225先物の最低取引単位(1単位)は日経平均株価の1,000倍、TOPIX先物の裁定取引単位(1単位)はTOPIXの10,000倍である。NT買い(売り)ポジションとは、日経225先物を1単位以上保持(空売り)、TOPIX先物を1単位以上空売り(保持)している状態を指し、それを解消することは保持している分を売り、空売りした分を買い戻すことを指す。

(4) 状態

松井らの手法では、状態変数として終値を相対化した値を用いている。これは、金融商品の価格などは大きく変動するため、そのまま状態として用いると、学習していない未知の状態に陥ってしまう可能性があるからである。時刻 t の状態変数 v_t を相対化した値 O_t を以下のように定義する。

$$O_t = \frac{v_t - \mu_{t,k}}{4\sigma_{t,k}} \quad (1)$$

ここで、 $\mu_{t,k}$ は時刻 t から過去 k 期間のデータから求めた移動平均、 $\sigma_{t,k}$ は同様に求めた移動標準偏差を表す。これにより、 $[\mu_{t,k} - 4\sigma_{t,k}, \mu_{t,k} + 4\sigma_{t,k}]$ の範囲を $[-1, 1]$ の範囲に正規化できる。松井らは終値とその移動標準偏差をそれぞれ相対化した2つの状態変数を用いている。

本研究では、深層強化学習の多数の状態変数を扱えるという利点を活かし、より状況を適切に捉えるため、状態変数の数を6に増やす。まず、TOPIX先物の終値に対する日経225先物の終値の割合であるNT倍率と、その移動標準偏差を相対化した値を状態変数とした。NT倍率は、松井らの終値と同様に現在の市場の動向を表す指標として採用する。次に利益確定を学習するために「含み損益」を加えた。 t 日目の含み損益 $prof_t$ は以下のように定義する。

$$prof_t = \frac{(P_t^N - P_{t-e}^N)S_t^N + (P_t^T - P_{t-e}^T)S_t^T}{A_0} \quad (2)$$

ここで、 P_t^N は t 日目の N (日経225先物)の価格、 P_t^T は t 日目の T (TOPIX先物)の価格、 e はポジションをとってからの日数である。よって、 P_{t-e}^N はポジションをとつ

た時の価格になる。\$S_t^N\$ は \$t\$ 日目の \$N\$ (日経 225 先物) の
 ストック数であり、保有している分を正の値、空売りして
 いる分を負の値で表す。\$A_0\$ は初期資産である。これを状態
 変数として取り入れることで、今ポジションを解消したら
 どのくらい利益が得られるかを把握することができる。次
 に「“NT 買いポジションをとってからの最大 NT 倍率”と
 “現在の NT 倍率”の差」と「“現在の NT 倍率”と“NT
 売りポジションをとってからの最低 NT 倍率”の差」をそ
 れぞれ「機会損失幅 (NT 買いポジション)」と「機会損失
 幅 (NT 売りポジション)」として定義し、状態変数として
 導入する。これらは、最大利益を獲得できる時点から NT 倍
 率がどのくらい変わってしまったかを把握するための状態
 変数である。そして、現在のポジションを把握するための
 「現在のポジション」を加えた 6 つを状態変数として学習
 を行う。

(5) 報酬

松井らの手法では、複利リターンを最大化するため、利
 益率 \$R\$、投資比率 \$f\$ の時のグロス利益率 (利益率に 1 を
 加えたもの、つまりは資産の変化前に対する変化後の割合
 である) の対数 \$\log(1 + Rf)\$ を報酬としている。しかし、
 本研究では複利リターンを考慮しない。また、松井らの手
 法ではとった行動に対してすぐに報酬を決めて与えている
 が、金融取引において行動の良し悪しをすぐに決めるのは
 大変困難である。そのため、ポジションを取得してから解
 消するまでの全ての行動に対する報酬を、ポジションを解
 消した後一括で決定し付与する。このときすべての行動
 に対し、付与量は一定で「ポジションをとった時からポジ
 ションを解消した時までの総資産の変化率」とする。また、
 ポジションを保持していないときに「何もしない」を選択
 した時の報酬は 0 とする。

2.2 提案手法の流れ

提案手法での学習の流れを以下で述べる。

① 初期化

行動価値関数を表すニューラル・ネットワークを初期化
 する。

② 取引とデータ収集

行動価値関数から得られる行動規則に従って取引を行
 い、データ (状態変数ベクトル \$X\$、行動 \$a\$、報酬 \$r\$、次の
 状態を表す状態変数ベクトル \$X'\$) を収集する。収集したデ
 ータは *Replay Buffer* に保存するが、この時、ポジショ
 ンの状態に応じて異なる処理を行う。ポジションを保持し
 ていないときに「何もしない」を選択した場合は報酬 \$r = 0\$
 とし、得られたデータを即座に *Replay Buffer* に加える。
 ポジションを取得した時から解消する時までのデータ
 は即座には報酬を決めずに、一旦 *Temp List* に保存する。
 これらのデータは、ポジションを解消した時に報酬をまと
 めて決定され、*Replay Buffer* に保存される。その後、
Temp List 内のデータをすべて削除する。

この際の行動選択には、\$\epsilon\$ の確率でランダムに行動し、
 それ以外は \$Q\$ 値の一番高い行動を選択する \$\epsilon\$-greedy 法
 を用いる。これを \$M\$ 回繰り返す。

③ ニューラル・ネットワークの更新

Replay Buffer 内のデータからランダムサンプリング
 により、\$m\$ 個取り出してそれぞれ \$Q\$ 値を計算し、それらを
 教師データとして行動価値関数を表すニューラル・ネット
 ワークを更新する。ここで、\$t\$ 日目の状態 \$X\$ での行動 \$a\$
 に対する \$Q\$ 値、つまり、\$X\$ と \$a\$ を入力した時の望ましい
 出力 \$q_t\$ は以下のように求める。

$$q_t \leftarrow r + \gamma \max_{a'} Q(X', a') \quad (3)$$

ここで、\$r\$ は 2.1 で決めた報酬、\$\gamma\$ は将来の報酬に対する
 割引率である。これにより、今回の行動で得られた報酬と、
 次の状態での最大価値を持つ行動の \$Q\$ 値を割引いたもの
 の和を望ましい出力とする。

④ 終了判定

②~③を任意の回数繰り返す。

テスト時には、行動価値関数から得られる行動規則に従い、
 テスト期間の取引を行う。この際、行動選択には、常に \$Q\$
 値の一番高い行動を選択する greedy法を用いる。

3. 実験

実験は日経 225 先物と TOPIX 先物の日次取引を対象と
 して行う。学習期間は 2009/3/4~2015/12/31 で 1682 日分、
 テスト期間は 2016/1/4~2017/12/29 で 506 日分のデータ
 を用いた。また、過去の実験より、NT 倍率が長期的に上昇ト
 レンドであることから、単に「1 単位 NT 買い (日経 225 先
 物買い、TOPIX 先物売り) をし続けてしまう」という局所
 解に陥ってしまったという問題があった。そこで本実験で
 は、学習期間の NT 倍率の時系列データからトレンドを除
 去する。原系列とトレンド除去後の NT 倍率の推移を図 2
 に示す。

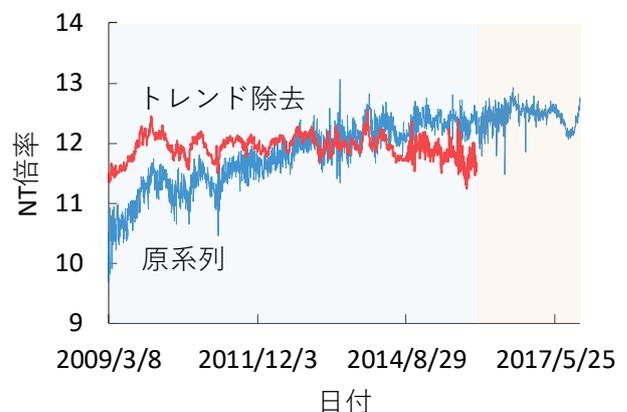


図 2 NT 倍率の推移

図2の横軸は日付、縦軸はNT倍率である。青い折れ線グラフはNT倍率の原系列、赤い折れ線グラフはトレンド除去後のNT倍率を表している。青背景の部分が学習期間、橙背景の部分がテスト期間である。取引は1日1回、前日の終値を観測し、当日の始値で行う。学習期間での取引をすべて終わるまでを1エピソードと定義し、1000エピソードを終える度にテスト期間の取引を行い、それを終えたらまた学習期間の取引を行う。

本研究で用いる深層強化学習のモデルはDeep Q-Networkである。ここで用いられるニューラル・ネットワークの中間層は2つで、そのユニット数は入力側から36, 25である。重みはXavierの初期値を用い、活性化関数は、中間層から出力層の間が線形結合、それ以外はランプ関数(ReLU)とした。最適化手法はAdam、学習時のニューラル・ネットワークの更新間隔は $M = 100$ 、ランダムサンプリングは優先順位付き経験再生(Prioritized Experience Replay)でサンプリング数は $m = 20$ で行う。

学習期間の行動選択方法は ϵ -greedy選択、テスト期間はgreedy選択とした。ランダムな行動を選ぶ確率 ϵ は0エピソード時には1.00とし、50,000エピソードかけて0.1まで線形に低下していくように設定した。Q値更新時の将来報酬の割引率は $\gamma = 0.95$ とした。

状態変数の相対化に用いる期間は $k = 5$ とし、初期資産は $A_0 = 10,000,000$ で実験を行った。

4. 結果と考察

まず、学習期間の最終総資産の推移を図3に示す。

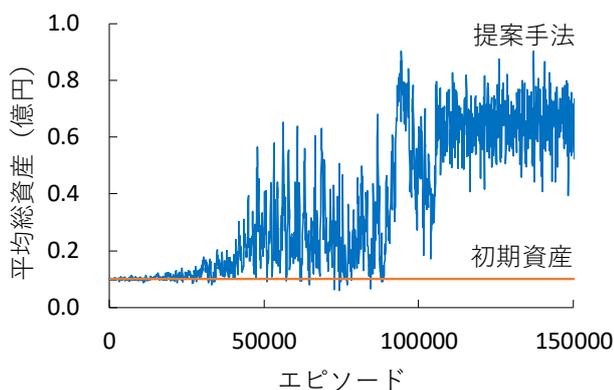


図3 学習期間の最終総資産の推移

図3の横軸はエピソード、縦軸は総資産である。青い折れ線グラフは、1エピソードの終わり時点での総資産を100エピソード毎に平均し、プロットしたものである。また、橙色の直線は初期資産である。これを見ると提案手法は、ゆるやかに総資産を伸ばし、最終的にはかなり高い値で収束していることが分かる。

次に、テスト期間の最終総資産の推移を図4に示す。

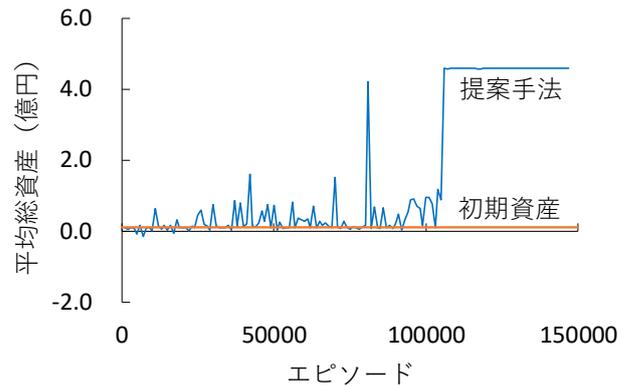


図4 テスト期間の最終総資産の推移

図4の横軸はエピソード、縦軸は総資産である。青い折れ線グラフは、テスト期間の取引結果の最終総資産をプロットしたものである。また、橙色の直線は初期資産である。序盤は大きく振動しているが、最終的には初期資産よりもかなり高い値で収束していることが分かる。

この収束している部分でどんな行動をとっているのかを調べたところ、基本的に「1単位NT買い(日経225先物買い、TOPIX先物売り)」をし続けていることが分かった。それ以外の行動をとることもあるが、それは ϵ -greedy選択によるランダム行動のときのみである。

今回、学習期間のNT倍率のデータからトレンドを除去することによって、「NT買いをし続ける」という局所解に陥ってしまうことを回避しようと試みたが、今回も同様の局所解に陥ってしまった。これは報酬の与え方に原因があると考えられる。ポジションを取得した時、ポジションを保持している時、ポジションを解消した時の報酬が一定であるため、ポジションを保持している時の報酬の方が回数の違いにより相対的に多くなってしまい、保持しやすくなってしまわないかと考えられる。

5. 今後の課題

実験結果より、「NT買いをし続ける」方策を学習していることが分かった。しかし、NT倍率が長期的に見て上昇トレンドであるとは言え、そこだけで稼ぐのではなく、短期的な下降トレンドでも利益を出せる方がより大きな利益を出せるはずである。今回の結果から、報酬の与え方により、行動の選択肢に偏りが生まれてしまっている可能性があることが示唆された。そのため、報酬の付与量を一定にするのではなく、ポジションの状態に応じて変更することを検討している。

さらに、現在はDeep Q-Networkという学習方法を用いているが、最新的手法としてA3C[3]というものが開発されている。A3C(Asynchronous Advantage Actor-Critic)は、Deep Q-Networkを発展させたモデルである。このモデルには、Asynchronous(複数のエージェントを同時に動かし、個々

の経験を集めて学習), Advantage (1 ステップ先ではなく, 数ステップ先の報酬を考慮) などの特徴がある. これを用いることで学習時に 1 つ先の報酬だけでなく, もう少し先の報酬も考慮できるようになる他, LSTM[4]などの時系列データの扱いに長けたニューラル・ネットワークの使用が可能になる. このような理由から, A3C の導入を検討している.

参考文献

- [1] 松井藤五郎, 後藤卓, 和泉潔, 陳ユ. 複利型強化学習における投資比率の最適化. 人工知能学会論文誌. 2013, vol. 28, no. 3, p. 267-272.
- [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), 2016, SIG-FIN-016-01.
- [3] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, Koray Kavukcuoglu. Asynchronous Methods for Deep Reinforcement Learning. In Proceedings of the 33rd International Conference on Machine Learning (ICML), 2016, p. 1928–1937.
- [4] Sepp Hochreiter, Jürgen Schmidhuber. Long Short-Term Memory. Neural computation, 1997, 9(8), p. 1735–1780.