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

Trading System using Deep Reinforcement Learning

常井 祥太　　穴田 一

Shota Tokoi　　Hajime Anada

東京都市大学大学院工学研究科
Graduate School of Engineering, Tokyo City University

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.

1. はじめに

近年，人工知能に関する研究が画像認識やゲーム AI の分野を中心に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上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，これらを全て考慮するには，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されるため，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，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．よって，投資比率と複利リターンを考慮する必要がない．

（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，1 単位以上 NT 買い（売り）をしている状態を指し，それを解消することは買った分を売り，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 は以下のように求める．

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

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た．

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している．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（TOPIX 先物：T），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 はポジションをとった時の価格になる．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．その上でポジションの状態に応じた適切な報酬を与えられるように，以下の 3 つの場合に分けて報酬を与える．

・ポジションを取得した時

ポジションを取得した時には，順張りの取引を意識して報酬 r を以下のように定める．

$$r=\begin{cases} O_t \times \omega_1 & if\ pos = long\ and\ O_t > 0 \\ |O_t| \times \omega_1 & if\ pos = short\ and\ O_t < 0 \\ 0 & otherwise \end{cases} \quad (3)$$

ここで，O_t は相対 NT 倍率，ω_1 はその重み，pos は現在のポジションで $long$ は NT 買いポジション，$short$ は NT 売りポジションである．これにより，NT 買いポジションをとった時に NT 倍率が過去 k 期間の平均と比べて高ければ，その分だけ報酬を与え，NT 売りポジションをとった時に NT 倍率が過去 k 期間の平均と比べて低ければ，その分だけ報酬を与えることを表す．

・ポジション保持状態の時

ポジションを解消せず，保持し続けている時には，t 日目の含み損益 $prof_t$ を用いて報酬 r を以下のように定める．

$$r = (prof_t - prof_{t-1}) \times \omega_2 \quad (4)$$

ここで，ω_2 は含み損益の重みである．これにより，含み損益を増やすような行動に対して報酬を与えることができる．

・ポジションを解消した時

ポジションを解消した時には，その時に得られる損益である t 日目の実現損益 $Rprof_t$ を用いて報酬 r を以下のように定める．

$$r = Rprof_t \times \omega_3$$

$$Rprof_t = \frac{P_t^N \times S_t^N + P_t^T \times S_t^T}{A_0} \quad (5)$$

ここで，ω_3 は実現損益の重みである．これにより，実現損益が高くなるタイミングでの利益確定ができるようになると考えられる．

さらに，持っていないポジションを解消しようとした際にマイナス 1 の報酬を与える．例えば，NT 買いポジションをとっている時に NT 売りポジションを解消しようとした時などである．このような行動をとらないように負の報酬を設定した．

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い，M 回取引を行い，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．この際の行動選択には，パラメータ ε 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 ε − greedy 法を用いる．

③ ニューラル・ネットワークの更新

集めた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．

$$q_t \leftarrow r + \gamma \max_{a'} Q(\mathrm{X}', a') \quad (6)$$

ここで，r は 2.1 で決めた報酬，γ 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．それぞれの NT 倍率の推移を図 2 に示す．

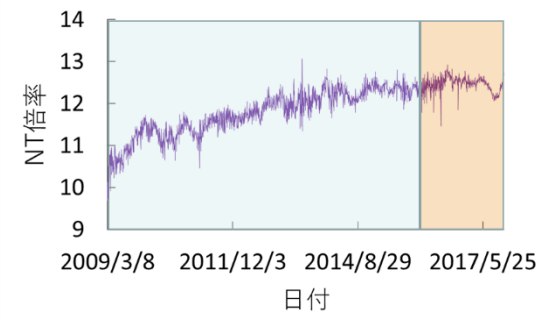

図 2：NT 倍率の推移．

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えるまでを 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数は $m = 20$ である．

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 0.05 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 $k = 5$，報酬の重みはそれぞれ $\omega_1 = 2.0, \omega_2 = 3.0, \omega_3 = 5.0$ とした．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．

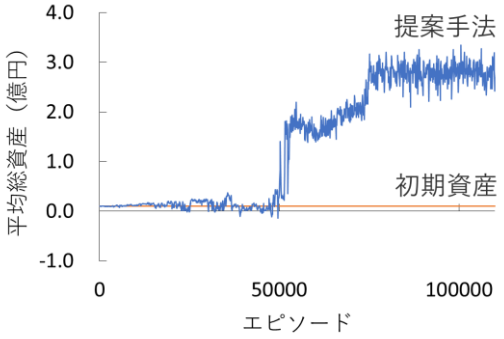

図 3：学習期間の最終総資産の推移．

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，50,000 エピソードまでは初期資産の周りを振動するだけであるが，それ以降に大きく総資産を伸ばし，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．

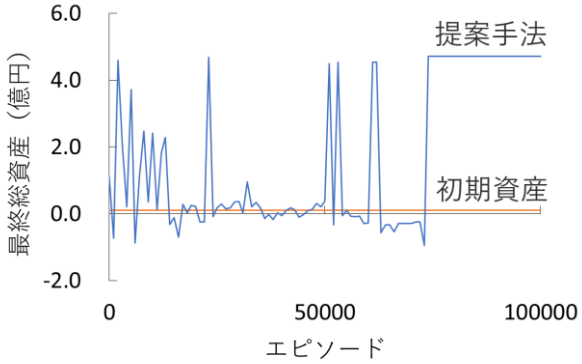

図 4：テスト期間の最終総資産の推移．

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．テスト期間は 1000 エピソード毎に行うため，序盤はかなり激しく振動しているが，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．

最後に，状態変数の「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）」がある場合とない場合の学習期間での比較を図 5 に示す．

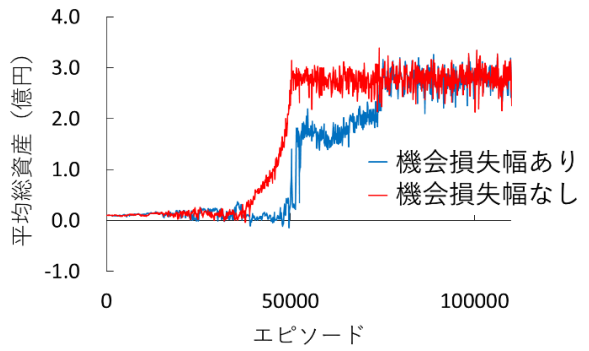

図 5：機会損失幅の有無の比較

図 5 の横軸はエピソード，縦軸は総資産である．こちらは図 3 と同様に学習期間の平均最終総資産の推移であり，青の折れ線グラフが機会損失幅あり，赤の折れ線グラフが機会損失幅なしである．これを見ると，序盤と終盤は同じような動きをしていることが分かる．しかし，中盤において機会損失幅がある方が，総資産の上昇が遅い．これは機会損失幅という新しい情報を受けて，新しい方策を創出しようとしているが，うまくいかずに同じ値で収束してしまっている，つまりは新しい情報を活かしきれていないと考えられる．

また，どちらも最終的に同じ値の周辺で収束しているが，この時に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．これは，学習期間の NT 倍率が長期的にみると上昇トレンドであるため，NT 買いをし続けることである程度稼ぐことができてしまい，「NT 買いをし続ける」という局所解に陥ってしまっていると考えられる．

5. 今後の課題

実験結果より，上昇トレンドに特化した「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，機会損失幅がある場合でもその情報を上手く活かし切れていないことが分かった．そのため，機会損失幅を状態変数に加えるだけでなく，報酬にもある程度の影響を与えるように変更することを検討している．また，学習期間の NT 倍率が長期的に見て上昇トレンドであることも原因の 1 つであると考えられるため，そこも改善が必要である．具体的には学習期間を短くし，上昇トレンドの比率を相対的に小さくすることなどを検討中である．

さらに，現在は Deep Q-Network という学習方法を用いているが，最新の手法として 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，非同期的に複数のエージェントから集めたデータを用いてニューラル・ネットワークの重みの変更を行うため，データをランダムサンプリングする必要がない．それによって 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，Vol.28, No.3, pp. 267-272 (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] 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), pp. 1928–1937 (2016)

[４] Sepp Hochreiter, Jürgen Schmidhuber : Long Short-Term Memory, Neural computation, 9(8), pp. 1735–1780 (1997)