A-008

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

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複利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，状態変数が 2 つと少なく，深層強化学習の利点である多数の状態変数を扱えることを活かし切れていないからである．また，国債や株価などには多くの変動要因があるが，それらを全て状態変数に加えるには，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，相関性の強い 2 つの金融商品に対して「買い」と「売り」の両建てをする裁定取引を考える．これにより，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．このように状況を簡略化した上で，現在の状況を適切に捉え，投資行動を行えるように状態変数を追加する．そして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

松井らの複利型深層強化学習による学習手法[2]と同様に，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，結果が良ければその行動に報酬を与え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）取引対象

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」の両建てをする裁定取引を考える．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，日経平均株価先物と TOPIX 先物がある．これらの価格の推移を図 1 に示す．

図 1　日経平均株価と TOPIX の推移

図 1 の横軸は期間，縦軸は価格である．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，わずかなズレ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，それらの比を見て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，日経平均株価先物と TOPIX 先物の二つを取引対象として選択した．

（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，投資比率と複利リターンを考慮する必要がないように，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（日経平均株価先物買い，TOPIX 売り）」，「1 単位 NT 売り（TOPIX 買い，日経平均株価先物売り）」，「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．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，売った分を買い戻すことを指す．NT 買いと売りでポジション解消の行動を分けた理由としては，これらの行動をとるべき状況はそれぞれ違うと考えたからである．

† 東京都市大学, Tokyo City University

Copyright © 2018 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

（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，体験したことのない未知の状態に陥りやすくなってしまうからである．これは相対化することで防止できる．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．

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

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． しかし，多数の状態変数を扱えるという深層強化学習の利点を活かし切れていない上，適切な行動決定を行うためには不十分であると思われる．

そこで本研究では，状態変数の数を 11 に増やす．まず，TOPIX 先物の終値に対する日経平均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．また，松井らの手法では，1 つの期間 k に対してのみ相対化を行っていたが，本研究では短期 k_1，中期 k_2，長期 k_3 の 3 つの期間に対して相対化を行う．これによって，より多くの変動パターンが表現できると考えられる．これにより，「相対 NT 倍率」が 3 パターン，「相対移動標準偏差」が 3 パターンの計 6 つとなった．次に，「所有現金の初期資産に対する割合」を加えた．これは，現金所有量がマイナスにならないように学習を進めるために必要となる．次に利益確定を学習するために「含み損益」，「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」，「NT 売りポジションをとってからの現在の NT 倍率と最低 NT 倍率の差」，「現在のポジション」の 5 つを加えた．「含み損益」はポジションをとった時の価格と現在の価格の差に現在の金融商品のストック量をそれぞれかけ合わせたものを初期資産で割った値とする．これを状態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と考えた．「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「NT 売りポジションをとってからの現在の NT 倍率と最低 NT 倍率の差」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．

（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の最大化ではなく，総資産の最大化を目的としているため，報酬 r を以下のように定めた．

$$r = \frac{asset_t - asset_{t-1}}{asset_{t-1}} \tag{2}$$

ここで，$asset_t$ は t 回目の行動終了後の総資産を表す．これにより，総資産を増やすような行動に対する報酬が大きくなるようにした．さらに，保有現金が負になった時や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 の報酬を与えることで，このような行動をとらないようにした．

2.2 提案手法の流れ

実験は日経平均先物と TOPIX 先物の日次取引を対象として行う．訓練期間は 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．訓練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，100 エピソードを終えたら，テスト期間に移行する．まず，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，行動選択には，定数 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 $\varepsilon-$ 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．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．

$$q_t \leftarrow Q(\mathrm{X}, a) + \alpha\left(r + \gamma \max_{a'} Q(\mathrm{X}', a') - Q(\mathrm{X}, a)\right) \tag{3}$$

ここで，α は学習率を表し，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．これは，現在の Q 値から見込みの Q 値である $r + \gamma \max_{a'} Q(\mathrm{X}', a')$ へと α だけ近づけることを表している．

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. 結果と考察

発表時に詳細な結果と考察を述べる．

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).

Copyright © 2018 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.