4P-04

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† 穴田 一†
東京都市大学 大学院工学研究科†

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．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．松井らは深層強化学習によって，日本国債の週次取引における行動規則を学習した[1]．しかし，この手法による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．そこで，本研究では考慮しなくてはいけない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．この取引は，相関性が強い 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ため，2 つの金融商品に共通する価格変動要因を相殺できる．その上で，状況を適切に捉え，投資行動を行えるように松井らの手法[1]の状態変数や報酬の与え方などに変更を加えた，深層強化学習によって投資戦略を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

2.1 既存研究からの変更点

本研究では，松井らの複利型深層強化学習による学習手法[1]をベースとし，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，国債には多くの変動要因が存在し，適切な行動選択を困難にしている．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日経 225 先物と TOPIX 先物に対して，「買い」と「売り」をそれぞれ同時に行う NT 倍率取引を考える．ここで，日経平均株価と TOPIX は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であり，違いは計算に組み入れられている企業や，株価か時価総額かのみである．そのため，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で，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，取引手法として NT 倍率取引を選択した．

（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[1]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「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 つを行動とする．ここで，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所有する金融商品をすべて現金化することを指す．

（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．時刻 t の状態変数 v_t を相対化した値 o_t は以下のように定義する．

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

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て，状態変数として用いている．

本研究では，深層強化学習の多数の状態変数

Trading System using Deep Reinforcement Learning
Shota Tokoi†, Hajime Anada†
†Graduate School of Engineering, Tokyo City University

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 10 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する．この時，移動平均を求める期間 k は短期，中期，長期の 3 パターン設定し，それぞれに対して相対化を行う．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．含み損益とは，取得した時の価格と時価を比較した未決済の損益のことである．これを初期資産で割ったもの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」を加えた 10 個の状態変数を用いて学習を行う．

（4）報酬

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即座に決めるのは大変困難である．そこで本研究では，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に一括で決定し，付与する．このとき付与量はポジションの状態によって異なるように設定した．買い（売り）ポジションの取得時と保持時には，「“最大（最低）NT 倍率”と“現在の NT 倍率”の差の絶対値」を報酬とする．このとき，最大（最低）NT 倍率の時点より前の行動に関してはそのままで正の報酬，後の行動に関しては -1 をかけて負の報酬とする．これにより，前者は「現在の NT 倍率からこのポジション中に NT 倍率がどれだけ上がる（下がる）か」，後者は「最大で稼げる NT 倍率からどのくらい下がって（上がって）しまったか」を考慮した報酬であることを表す．買い（売り）ポジションの解消時には「“現在の NT 倍率”と“ポジション取得時の NT 倍率”の差（“ポジション取得時の NT 倍率”と“現在の NT 倍率”の差）」を報酬とする．これは「ポジションを取得した時の NT 倍率からどれだけ上がった（下がった）か」，つまり利益をどれだけ出せたかを考慮した報酬であることを表す．

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

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．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 日分のデータである．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．提案手法の学習の流れは以下の通りである．

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は $r = 0$ 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し，$Replay\ Buffer$ に保存する．その後，$Temp\ List$ 内のデータをすべて削除する．ここまでを 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に対して，その状態におけるその行動の価値（Q 値）を計算し，これらを訓練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．

④ 終了判定

②～③を任意の回数繰り返す．

また，行動選択法は，訓練期間では 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，テスト期間には，常に Q 値の一番高い行動を選択する greedy 法を用いる．

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，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).