1K-05

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† 東京都市大学知識工学部‡

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．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，相関性の強い 2 つの金融商品に対して「売り」と「買い」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．そこで本研究では，裁定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

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[2]をベースとし，実現益の最大化を目的として，以下の点を変更した．

Construction of Investment Strategy using Deep Reinforcement Learning

† Shota Tokoi, School of Engineering, Tokyo City University Graduate Division

‡ Hajime Anada, Faculty of Knowledge Engineering, Tokyo City University

(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，国債には多くの変動要因が存在するため，価格変動を予測し，適切な行動規則を学習することは困難である．そこで，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，「売り」と「買い」の両建てをする裁定取引を考える．このような金融商品として，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と計算方法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，それらの比を見ることによって，価格変動要因の大部分が相殺され，価格差のみに着目した取引が可能になる．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量を 1 単位で固定した．

（3）状態

本研究では松井らの手法と同様に，状態変数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．

松井らの手法では，相対化は 1 つの期間 k に対してのみ行っていたが，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．

具体的には，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それぞれの期間で相対化する．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 パターンとなる．これに利益確定を学習するための「含み損益」を加えた 7 つの状態変数を用いて学習を行う．含み損益は相対化せず，投資額で除算した値を状態変数として用いる．

（4）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，本手法では複利リターンの最大化ではなく，実現益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る．

$$r = \alpha \times PL_{ur} + \beta \times PL_r \quad (2)$$

ここで，PL_{ur} は含み損益，PL_r は実現損益を表し，α と β はこれらを調節するパラメータである．含み損益について報酬を与え，利益確定を行った際に生じる実現損益についても報酬を与えることで，含み益の高い金融商品を保有するだけでなく，より稼げるタイミングでの利益確定を学習できるようになると考えられる．

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 日分のデータになる．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．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 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．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．

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

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)