3L3-OS-26a-4

実取引環境における複利型強化学習を用いた取引戦略の獲得

Acquiring Trading Strategy Using Compound Reinforcement Learning in Online Trading Platform

松井 藤五郎[*1] 長瀬 舜[*1] 後藤 卓[*2] 和泉 潔[*3] 陳 ユ[*3] 鳥海 不二夫[*3]
Tohgoroh Matsui Shun Nagase Takashi Goto Kiyoshi Izumi Yu Chen Fujio Toriumi

[*1]中部大学 Chubu University
[*2]株式会社三菱東京 UFJ 銀行 Bank of Tokyo-Mitsubishi UFJ, Ltd.
[*3]東京大学 The University of Tokyo

This paper describes an application of compound reinforcement learning to an online trading platform. We use TradeStation that is the most popular online trading platform among individual investors. We propose a measure for commission fee in order to improve the winning rate.

1. はじめに

我々は，これまで，日本株の仮想取引環境であるカブロボにおいて複利型強化学習 [松井 11a, Matsui 12, 松井 11b, 松井 13b, 松井 13a] を用いて取引戦略を獲得する手法を開発してきた [松井 07, 松井 09, 後藤 13]．カブロボ[*1]は，Java によるマルチプラットフォーム環境をサポートしており，パフォーマンス分析やテクニカル指標など株取引に必要な API が提供されていることから，様々な環境で高度な取引戦略を実装することができる [鳥海 06]．しかしながら，注文を出す機会が前場が開く前と後場が開く前の 1 日 2 回だけに制限されており，カブロボではデイトレードを行うことができない．一般的には注文を出す機会が多いほど安定した運用を行うことができる可能性が高いことから，本研究では，デイトレードによる安定運用を目指す．

これに対し，TradeStation[*2]は，価格情報が更新されるごとにプログラムが実行されるため，更新間隔を短くすることによってデイトレードを行うことができる．また，TradeStation は，開発したプログラムを用いて実際の取引を行うことができる．

そこで，本論文では，これまで開発してきた複利型強化学習を用いて取引戦略を獲得する手法を TradeStation 上に実装し，デイトレードを行う方法を提案する．ただし，TradeStation は EasyLanguage という独自の言語を用いており，取引に必要な API も十分には提供されていない．そのため，複利型強化学習を用いて取引戦略を獲得するには実装上の工夫が必要となる．また，デイトレードにおいては，価格の変動が小さいため，手数料をありの場合に細かい取引を行うと勝率が極端に悪くなる．そこで，本論文では，これを防ぐための手数料対策の方法を提案する．

従来研究 [後藤 13] と本研究における環境の違いを表 1 に示す．従来研究との最大の違いは，時間足が日足から分足になってデイトレードを行う点である．この他にも，[中原 13] など，強化学習を用いて株取引を行う研究を行っている研究がいくつかあるが，実取引が可能な環境では強化学習を行っていない．

表 1: 従来研究 [後藤 13] との比較

	[後藤 13]	本研究
プラットフォーム	カブロボ	TradeStation
使用言語	Java	EasyLanguage
バックテスト	○	○
実取引	×	○
対象商品	日本株	米国株
時間足	日足	分足

2. 複利型強化学習を用いた株取引戦略の獲得

複利型強化学習は，割引複利利益率（割引複利リターン）

$$(1+R_{t+1}f)(1+R_{t+2}f)^{\gamma}(1+R_{t+3}f)^{\gamma^2}\ldots = \prod_{k=0}^{\infty}(1+R_{t+k+1}f)^{\gamma^k}$$

の期待値を最大化するような行動規則を学習する．ここで，R_t は時刻 t に観測された利益率（リターン），γ は割引率パラメーター，f は投資比率パラメーターを表す．割引複利利益率は，対数を取ることで，従来の強化学習と同じように再帰的な形で表すことができる．すなわち，行動規則 π の下での状態 s の価値 $V^{\pi}(s)$ と行動規則 π の下での状態 s における行動 a の価値 $Q^{\pi}(s,a)$ は次のように表される．

$$\begin{aligned} V^{\pi}(s) &= \mathrm{E}_{\pi}\left[\log\prod_{k=0}^{\infty}(1+R_{t+k+1}f)^{\gamma^k}\,\middle|\, s_t=s\right] \\ &= \sum_{a\in\mathcal{A}}\pi(s,a)\sum_{s'\in\mathcal{S}}\mathcal{P}^a_{ss'}\left(\mathsf{R}^a_{ss'}+\gamma V^{\pi}(s')\right) \qquad (1) \\ Q^{\pi}(s,a) &= \mathrm{E}_{\pi}\left[\log\prod_{k=0}^{\infty}(1+R_{t+k+1}f)^{\gamma^k}\,\middle|\, s_t=s, a_t=a\right] \\ &= \sum_{s'\in\mathcal{S}}\mathcal{P}^a_{ss'}\left(\mathsf{R}^a_{ss'}+\gamma V^{\pi}(s')\right) \qquad (2) \end{aligned}$$

ここで，$\pi(s,a)$ は行動規則 π の下で状態 s において行動 a が選択される確率（行動選択確率），$\mathcal{P}^a_{ss'}$ は状態 s において行動 a を行ったときに次の状態が s' になる確率（状態遷移確率），$\mathsf{R}^a_{ss'}$ は状態 s において行動 a を行って次の状態が s' になった

連絡先: 松井藤五郎, 中部大学, 愛知県春日井市松本町 1200, TohgorohMatsui@tohgoroh.jp

∗1 http://www.kaburobo.jp
∗2 http://www.tradestation.com

Algorithm 1 複利型 OnPS アルゴリズム.

入力: 割引率 γ, 強化学習率 α, 初期優先度 p, 初期投資比率 f, 投資比率学習率 η
for all s, a **do**
  $P(s,a)$ を p に初期化
  $f(s,a)$ を f に初期化
end for
loop (各エピソードに対して繰り返し)
  $c(s,a) \leftarrow 0$ **for all** s, a
  状態 s を初期化
  repeat (エピソードの各ステップに対して繰り返し)
    P から導かれる行動規則に従って s での行動 a を選択
    $c(s,a) \leftarrow c(s,a) + 1$
    行動 a を実行し，利益率 R と次の状態 s' を観測
    for all s, a **do**
      $P(s,a) \leftarrow P(s,a) + \alpha \log(1 + Rf(s,a))c(s,a)$
      $c(s,a) \leftarrow \gamma c(s,a)$
    end for
    $f(s,a) \leftarrow f(s,a) + \eta \frac{R}{1+Rf(s,a)}$
    $s \leftarrow s'$
  until s が終端状態
end loop

ときに得られる利益率に投資比率を掛けて 1 を加えたものの対数の期待値

$$\mathrm{R}^{a}_{ss'} = \mathrm{E}_{\pi}\left[\log(1 + r_{t+1}f) \middle| s_t = s, a_t = a, s_{t+1} = s'\right] \quad (3)$$

を表す．複利型強化学習では，すべての s, a に対してこの $Q^{\pi}(s,a)$ を最大化するような行動規則 π を学習する．

本論文では，取引戦略の学習にオンライン勾配法を用いて投資比率最適化をする複利型 OnPS [後藤 13, 松井 13b] を用いる．複利型 OnPS のアルゴリズムを Algorithm 1 に示す．

複利型強化学習における状態は，終値と移動標準偏差に基づいた二次元空間で表現する．株価は大きく変動するため，直近のデータと比較した相対的な値として正規化することによって，株価が大きく異なる場合でも学習した行動規則を利用できるようにする．具体的には，移動平均および移動標準偏差の算出期間を k とし，以下のようにして相対化 [Matsui 09] する．

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

ここで，v_t は t における値，$\mu_{t,k}$ は時刻 t の直近 k 個のデータから求めた移動平均，$\sigma_{t,k}$ は同じく移動標準偏差を表す．終値を相対化した値を相対終値（RCP），移動標準偏差を相対化した値を相対移動標準偏差（RMSD）と呼ぶ．RCP が正のときは現在の株価が移動平均株価より大きい，すなわち，株価が上昇していることを表している．RMSD が正のときは現在の標準偏差が移動平均標準偏差より大きい，すなわち，株価の変動が大きくなっていることを表している．これらの値は共に連続値をとるので，15×15 の格子状に配置した動径基底関数を用いて線形関数近似を行う．

エージェントの行動は買いと売りの 2 種類である．株式を購入している状態をロング・ポジション，株式を信用売りしている状態をショート・ポジションという．エージェントは，複利型強化学習によって学習された取引戦略によって行動を選択し，オンライン勾配法によって学習された投資比率 f によってポジションの大きさを調整する．

3. TradeStation における Strategy 構築

TradeStation は，米国の TradeStation 社が提供している個人投資家向けの実取引環境である．TradeStation では，米国の株式，オプション，先物，FX の 4 種類の金融商品の取引を行うことができ，システムトレードとシステムトレードを行うためのバックテストができる．2011 年 4 月にマネックス・グループが同社の株を買収し，日本向けのプラットフォームも公開される予定である．本研究では，米国でのリリース版 TradeStation9.1 を使用している．

TradeStation における自動取引のルールは，Strategy と呼ばれる．Strategy は，TradeStation が提供している様々な機能のうち，Chart と呼ばれる機能に適用する．

Strategy は，EasyLanguage と呼ばれる専用の言語で記述される．TradeStation には TradeStation Development Environment と呼ばれる EasyLanguage 専用の開発環境が用意されており，これを利用して Strategy を作成・編集する．TradeStation において，銘柄，時間足，期間を指定すると，指令された銘柄，時間足，期間の Chart（時系列グラフ）が作成される．この Chart に Strategy を適用すると，Strategy に記述されているルールに従って，この期間のバックテストが行われる．

EasyLanguage は，TradeStation 専用のシステムトレードのルールを記述するための言語である．例えば，

```
Buy 100 shares next bar at market;
```

という命令は，次の足（`next bar`）に成行（`at market`）で 100 株（`100 shares`）買注文を出す（`Buy`）ということを表す．

しかし，単純なルールによるシステムトレード専用であり，かつ，実取引を前提とした言語であるため，複利型強化学習を用いた Strategy を構築する上で必要となるバックテスト中に取引ごとのリターン，保有資産評価額，総資産額等を計算するための API が用意されていない．そこで，そのため，強化学習をする上で必要となるリターン，保有資産評価額，総資産額は取引記録に基づいて算出した．

本論文では，状態変数である相対終値と相対移動標準偏差を直近 30 足のデータから求めている．相対化する前の移動標準偏差の計算に直近 30 足のデータを用いるため，状態変数を求めるには直近 60 足のデータが必要となる．証券取引所は，平日の昼間のみ取引が行われるため，市場が開いた直後には，直近 60 足のデータに前日のものが含まれてしまう．市場が開いた直後の株価は前日終値と大きく乖離していることがあるため，本研究では，市場が開いてから 60 足の間は取引を行わない．

実際に Strategy をバックテストした TradeStation の実行画面を図 1 に示す．Chart の横軸が時間，縦軸が株価を示し，ボリンジャーバンド（移動標準偏差）を表示している．足に対して下からの矢印が買注文，上からの矢印が売注文で，数値は取引した株数を示す．Chart 中の縦の破線は営業日の境目を示す．この Chart から，営業日の境目で株価が大きく変わっていることが確認できる．また，市場が開いてからしばらくの間，取引を行っていないことも確認できる．

我々は，[長瀬 13] において，上場投資信託である SPDR S&P 500 ETF Trust (SPY) を取引対象とし，時間足を 1 分足，手数料を 0 として実験を行った．学習期間を 1 週間，2 週間，1 か月，3 ヶ月，6 ヶ月，1 年，2 年，運用期間を 1 日とし，それぞれ無作為に 30 回バックテストを行い，利益率，最大ドローダウン，シャープレシオを評価した．参考のため，2012 年から 2013 年にかけての SPY の値動き（日足終値）を図 2 に示す．

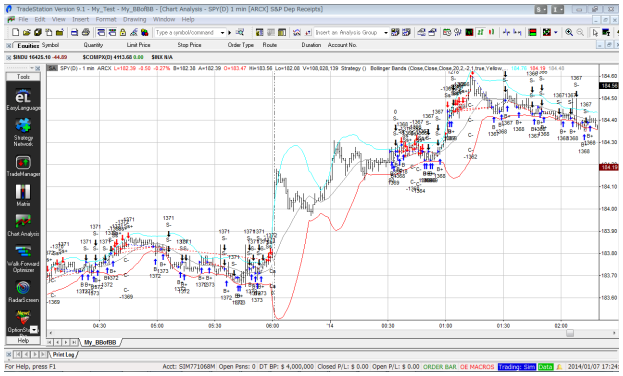

図 1: TradeStation の実行画面.

図 2: SPDR S&P 500 ETF Trust (SPY) の値動き.

図 3 は，30 回のバックテストの結果の幾何平均利益率から 1 年間を 250 営業日として年換算の利益率を求めた結果である．

学習期間が 3 ヶ月以下のときと 2 年のときは利益をあげることができなかったが，学習期間が 6 ヶ月のときは年換算で 10.2%，学習期間が 1 年のときは年換算で 23.2% の利益をあげることができた．このことから，学習期間を 1 年とすると，利益をあげる取引戦略を獲得できることが確認された．しかしながら，この結果は手数料をなしとしたときのものであり，手数料を一株当たり$0.1 とすると，極端に勝率が悪くなり，利益を出せなくなってしまうことが判明した．これは，購入価格からほとんど価格が動いていないにも関わらず，細かい取引を行うためである．そこで，本論文では，以下のような手数料対策を提案する．

4. 手数料対策

まず，株価の購入価格からの変動が手数料分よりも小さい間は，取引をしないことにした．これは，購入価格からの変動が手数料分よりも小さい間は，どのような取引を行っても必ず損失が発生してしまうからである．実際には，損失が発生してでも取引を行ったほうがよい場合もあり得るが，ここでは，そのような場合は考えないものとする．

次に，現在学習中の最適投資比率と実際の投資比率の乖離した際に生じる追加注文および部分決済によるポジション調整を取りやめる．ポジション調整を行わないことによって，細かい取引を大幅に削減することが期待できる．ポジション調整を行うと，平均購入価格が変動し，変動が手数料分よりも小さい間に取引を行わないよう対策の影響が大きくなってしまう．例えば，追加注文を行って，平均購入価格が上昇すると，平均購入価格に対して手数料を上乗せした分まで上昇しないと取引が行われない．

さらに，含み益を状態変数として追加した．したがって，エージェントの状態を相対終値，相対移動標準偏差，含み益の 3 次元で表現する．これによって，エージェントが，現在含み益がどのくらい出ているのかを知ることができ，含み益が出ているときだけ決済するなどの行動を学習することが期待できる．ただし，状態変数を加えると状態の特徴数が増えるため，動径基底関数は $15 \times 15 \times 15$ ではなく $9 \times 9 \times 9$ の格子状に配置した．

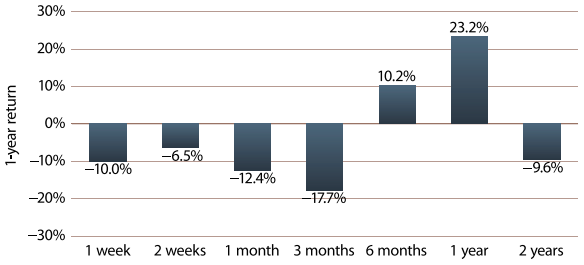

図 3: 年換算利益率.

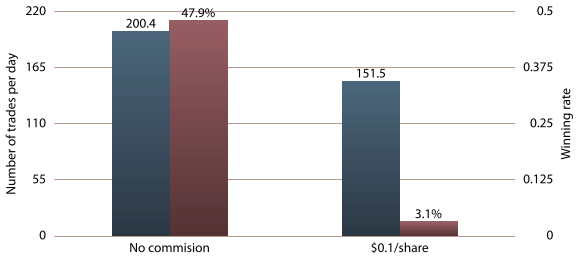

図 4: 定数量導入前後の学習期間中の一日あたりの取引回数と勝率の比較.

5. 実験結果

提案した手数料対策の有効性を確認するため，実験を行った．取引対象は手数料対策を導入する前と同じ SPY とし，時間足も導入前と同じ 1 分足とした．2013 年の各月の第 3 水曜日をテスト期間とし，その直前の 1 年間を学習期間とした．結果を図 5 に示す．

左側の縦棒は一年を 250 営業日として換算した一日あたりの取引回数を表し，右側の棒グラフは勝率を表している．手数料の導入によって勝率が 47.9%から 3.1%に激減したが，提案手法による手数料対策を行うことによって勝率が 37.1%にまで回復した．一日あたりの取引回数は，手数料なしのときは 200.4 回だったのに対し，手数料を導入した後は 151.5 回，手数料対策を行ったときは 24.5 回であった．

これは，取引を勝てる状況のみに絞り込むことによって，負け取引を大幅に削減していることを意味する．一日あたりの勝ち取引回数は，手数料対策を行ったことによって，4.71 回から 9.09 回に増加した．

提案した手数料対策によって勝率は大幅に改善されたが，手数料がある場合のトータルの運用成績は正にはならなかった．このことから，個人投資家向けの手数料が適用されている場合には，デイトレードを頻繁に行うと利益を得るのが難しいことがわかる．

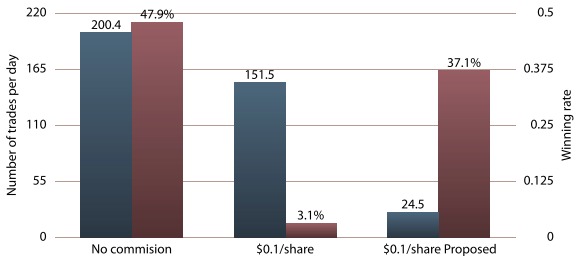

図 5: 学習期間中の一日あたりの取引回数と勝率.

6. まとめ

本論文では，複利型強化学習を用いて取引戦略を獲得する手法を TradeStation 上に実装し，デイトレードを行う方法について述べた．本手法では直近のデータから状態変数を求めているが，デイトレードにおいては市場が開いた直後の状態変数が前日のデータの影響を受けてしまうため，前日のデータの影響がある間は取引を行わない．本手法を用いた評価によって，学習期間を 1 年としたときに複利利益率が最大となることが確認されている．しかし，手数料をありにすると，勝率が極端に下がって利益を出すことができない．

そこで，本論文では，手数料対策として，価格の変動が手数料分よりも小さい間は取引を行わないこと，資産評価額の変動あるいは状態の遷移によって実際の投資比率が推定される最適投資比率と乖離してもポジション調整を行わないこと，状態変数に含み益を加えることの 3 点を提案した．実験によって，提案手法を用いることによって手数料ありの場合の勝率が大きく改善されることを確認した．

しかしながら，勝率が改善されても，運用成績を正にすることはできなかった．このことは個人投資家向けの手数料がかかる場合にはデイトレードによって資産を安定的に運用することが難しいことをしているが，今後は，強化学習パラメーターのチューニング等で運用成績を正にできるかどうか検討を行いたい．

留意事項

本論文は三菱東京 UFJ 銀行の公式見解を表すものではありません．

謝辞

本研究で使用している TradeStation のアカウントはマネックス証券株式会社より提供していただいています．ここに感謝の意を表します．

参考文献

[Matsui 09] Matsui, T., Goto, T., and Izumi, K.: Acquiring a government bond trading strategy using reinforcement learning, *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 13, No. 6, pp. 691–696 (2009)

[Matsui 12] Matsui, T., Goto, T., Izumi, K., and Chen, Y.: Compound Reinforcement Learning: Theory and An Application to Finance, in Sanner, S. and Hutter, M. eds., *Recent Advanced in Reinforcement Learning: Revised and Selected Papers of the European Workshop on Reinforcement Learning 9 (EWRL 2011)*, Vol. 7188 of *Lecture Notes in Computer Science*, pp. 321–332 (2012)

[後藤 13] 後藤 卓, 松井 藤五郎, 大澄 祥広：複利型強化学習の株式取引への応用, 第 27 回人工知能学会全国大会 (JSAI 2013), 4I1-OS-16-4 (2013)

[鳥海 06] 鳥海 不二夫：株式売買ソフトウェア スーパー株ロボを作ろう!, 秀和システム (2006)

[中原 13] 中原 孝信, 羽室 行信, 岡田 克彦, 宇野 毅明：強化学習を用いた相場のブーム検知と株取引への適用, 第 27 回人工知能学会全国大会 (JSAI 2013), 1E4-3 (2013)

[長瀬 13] 長瀬 舜, 松井 藤五郎, 後藤 卓, 和泉 潔, 陳 ユ, 鳥海 不二夫：TradeStation における複利型強化学習を用いた Strategy 構築, 第 12 回人工知能学会金融情報学研究会 (SIG-FIN), pp. 51–55 (2013)

[松井 07] 松井 藤五郎：カブロボへの招待–人工知能を用いた株式取引–, 人工知能学会誌, Vol. 22, No. 4, pp. 540–547 (2007)

[松井 09] 松井 藤五郎, 後藤 卓：強化学習を用いた金融市場取引戦略の獲得と分析, 人工知能学会誌, Vol. 24, No. 3, pp. 400–407 (2009)

[松井 11a] 松井 藤五郎：複利型強化学習, 人工知能学会論文誌, Vol. 26, No. 2, pp. 330–334 (2011)

[松井 11b] 松井 藤五郎, 後藤 卓, 和泉 潔, 陳 ユ：複利型強化学習の枠組みと応用, 情報処理学会論文誌, Vol. 52, No. 12, pp. 3300–3308 (2011)

[松井 13a] 松井 藤五郎：複利型強化学習—強化学習のファイナンスへの応用—, 計測と制御 (計測自動制御学会誌) , Vol. 52, No. 11, pp. 1022–1027 (2013)

[松井 13b] 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)