強化学習を用いた金融市場における日次取引戦略の獲得

Acquiring a Daily Trading Stragety in a Financial Market Using Reinforcement Learning

松井 藤五郎[*1] 畔上 康一[*2] 後藤 卓[*3] 和泉 潔[*4] 大和田 勇人[*5]
Tohgoroh Matsui Koichi Azegami Takashi Goto Kiyoshi Izumi Hayato Ohwada

[*1]とうごろう機械学習研究所 Tohgoroh Machine Learning Research Institute
[*2]キヤノン株式会社 Canon Inc.
[*3]三菱東京 UFJ 銀行株式会社 The Bank of Tokyo Mitsubishi UFJ, Ltd.
[*4]産業技術総合研究所 National Institute of Advanced Industrial Science and Technology
[*5]東京理科大学 Tokyo University of Science

This paper descibed methods to acquire a daily trading strategy in a financial market using reinforcement learning. In general, reinforcement learning methods assumed that the length between each time step is fixed. However, the length is varied in daily trading, because financial markets are closed in Saturdays, Sundays, and holidays. In this paper, we propose two methods to solve this problem and show that an experimental result which indicates that one of the proposed method is efficient to acquire a daily trading strategy in the government bond market.

1. はじめに

1980 年代から，コンピューターを駆使して自動的に金融商品の売買を行うシステム・トレードが盛んになった．初期のシステム・トレードは単純な条件を持たせたプログラムによって人間の替わりに自動的に売買注文を出すというものであったが，1990 年代後半からニューラル・ネットワークや GA（遺伝的アルゴリズム）などの人工知能技術を用いたシステム・トレードが行われるようになった [田中 09]．その後も，人工知能の分野では新しい機械学習およびデータ・マイニングの理論や手法が次々と生み出されているが，金融市場が複雑であるため，新しく開発された理論や手法の応用は進んでいない．そこで，様々な分野で成果を上げている人工知能の理論や手法を金融市場へ応用し，安定した資産運用を自動的に行うことへのニーズが高まっている．

資産運用は，資産を預金や投資によって効率よく増やすことである．金融市場における取引戦略とは，一般には，金融市場において資産運用を行うための戦略と考えられている．本研究では，これを広義の取引戦略とし，資産運用の問題を

1. リスクが小さくなるように金融資産を預金，株式，債券などに配分して取引対象を決めるポートフォリオ構築,
2. 配分された資産をその対象に投じて殖やす注文決定,
3. 決定された注文を効率よく行う注文執行

に分けて考える．そして，2 の注文決定に用いられる戦略を狭義の取引戦略とする．

筆者らは，これまでに，株式と国債を対象として，強化学習を用いて狭義の取引戦略を獲得する手法を開発してきた [松井 07, 松井 08b, 松井 08a, 松井 09]．これまでの国債を対象とした研究では週次取引を対象としていたが，本論文では，残存期間[*i]10 年の日本国債の日次取引を対象とする．週次取引に比べると，日次取引は利益を上げる機会が増えるとともに，利率の変化によるリスクを減らすことが期待できる．

ところが，強化学習を用いて日次取引のための取引戦略を獲得しようとすると，時間の間隔が一定でなく不規則であるという問題に直面する．これは，通常の強化学習ではステップ間隔を一定と仮定していることが原因である．そこで，本論文では，強化学習を用いて日次取引のための取引戦略を獲得する際に直面する問題について述べ，その解決方法を提案する．また，日本国債を対象とした実験によって提案手法の有効性を示す．

連絡先: 松井藤五郎, TohgorohMatsui@tohgoroh.jp

∗i 債券の実際の価値は金利を受け取れる期間に応じて異なるため，債券市場では償還期日までの期間を同じにしたときの価値に換算して取引が行われる．この償還期日までの期間を残存期間という．

2. 強化学習を用いた取引戦略の獲得

2.1 日本国債の取引

国債は，株式の個別銘柄とは異なり，企業動向によるミクロの影響を受けにくく，個別企業のファンダメンタルズに左右されることもほとんどない．このため，株式の個別銘柄と比較して，テクニカル分析に基づいた取引戦略を策定しやすい．

株式においては価格が上昇すると運用利回りが上昇するが，日本国債を含む債券において価格が上昇すると運用利回りが減少する．したがって，金利が高い時に債券を買って，金利が低い時に債券を売るのが良い取引方法である．債券を買っている状態をロング・ポジションといい，債券を売っている（信用売りしている）状態をショート・ポジションという．

図 1 に，2004 年から 2008 年にかけての残存期間 10 年の日本国債の金利の推移を示す．本論文では，この期間のデータを対象として債券取引戦略の学習と評価を行う．

本研究では，各取引日において市場が閉まる直前に金利を観測でき，その値に応じてすぐに注文するとそのままの金利ですぐに取引が成立するものとする．そこで，市場が閉まる直前の金利を終値で近似することとする．また，取引手数料はかからないものとする．

これらは，銀行の債券取引部門など，実際に日本国債の取引を行っている現場でも同じように取引ができることから，このように簡単化しても大きな問題はない．

2.2 金融市場において強化学習が直面する問題

ここでは，強化学習を用いて金融市場における取引戦略を獲得する際に直面する問題について述べる．

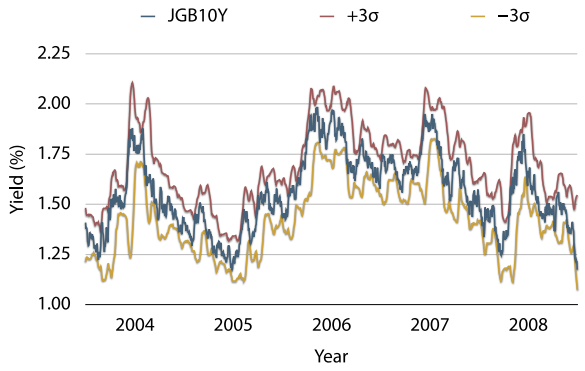

図 1　2004 年から 2008 年までの残存期間 10 年の金利（JGB10Y）と 20 営業日ボリンジャー・バンド（±3σ）の推移.

2.2.1　金融市場でなくても直面する問題

最初に直面する問題は，金融市場の状態を表すパラメーターが非常に多いという次元の呪いである．次に，基本的な強化学習は状態の集合と行動の集合をいずれも離散的で有限なものと仮定しているため，金融市場においては連続状態と連続行動が問題となる．さらに，金融市場の状態を完全に観測することはできないため，エージェントの観測は部分観測となる．

状態や行動そのものが複雑であるだけでなく，状態遷移も複雑なものとなる．エージェントが決定した注文は取引が成立しないと執行できないため，他の市場参加者の行動が状態遷移に大きな影響を与える．

2.2.2　金融市場で直面する観測値の範囲の問題

金融市場では，価格が大きく変化することは珍しくないため，観測値が大きく変わってしまう可能性がある．実際に，残存期間 10 年の日本国債の市場において，2006 年の金利水準は 2005 年に比べて高かったため，金利をそのまま観測値として用いた場合，2005 年のデータから獲得した取引戦略を 2006 年に用いて利益を上げることは難しいと考えられる．したがって，価格をそのまま観測値とするのではなく，その価格が高いのか低いのかを相対的に表した値を観測値とするべきである．

2.3　強化学習を用いた取引戦略の獲得

本研究では，profit sharing をオンライン型に改良した OnPS [Matsui 03] を用いている．OnPS は，等比減少信用割当関数を用いており，過去に選択した行動への信用割当をステップごとに割引率 γ ずつ減衰させる．

観測値は 2 つのパラメーターとする．これは，獲得した取引戦略を視覚化して分析することを容易にするためでもある．次に，連続状態については，RBF 特徴 [Sutton 98] を格子状に配置して関数近似を行うことによって対処している．また，連続行動となることを避けるために，学習時にエージェントが取り得るポジションを -1 または $+1$ のいずれかとしている．獲得した取引戦略を用いる際に行動選択確率から計算した期待ポジションを取ることによって，ポジションの強弱を表すことができる．観測値を 2 つに限定したことによって明らかに部分観測となっているが，OnPS は行動価値推定型のアルゴリズムに比べて非マルコフ環境に強いため，部分観測に対する特別な工夫はしていない．

観測値については，テクニカル分析のアイディアを用いて直近 n 個のデータから現在の値が相対的に大きいか小さいかを決め，実際の観測値としている．具体的には，観測する値 v_t に対して，直近 n 個の値 $v_t, v_{t-1}, \ldots, v_{t-n+1}$ から，平均 $\mu_{t,n}$ と標準偏差 $\sigma_{t,n}$ を計算し，これらの値に基づいて実際の観測値 $o_{t,n}$ を次のようにして求める．

$$o_{t,n} = \frac{v_t - \mu_{t,n}}{3\sigma_{t,n}}$$

これによって，この観測値が正規分布に従うと仮定すると，直近 n 個の値のうち 99.7% の値が $[-1, +1]$ の範囲に入る．

報酬は，評価損益の増分とする．また，一方のポジションを取ってから反対のポジションに変更する直前までを一つのエピソードとしている．

表 1　2004 年 1 月前半の残存期間 10 年の日本国債の金利（終値）.

日付	曜日	JGB10Y
2004/01/05	月	1.409
2004/01/06	火	1.378
2004/01/07	水	1.383
2004/01/08	木	1.382
2004/01/09	金	1.363
2004/01/13	火	1.314
2004/01/14	水	1.308
2004/01/15	木	1.294
2004/01/16	金	1.297
2004/01/19	月	1.309

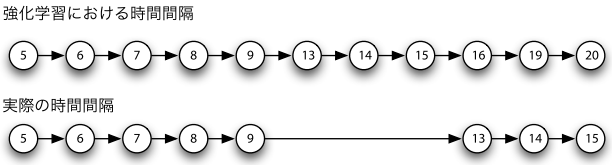

図 2　強化学習における時間間隔と実際の時間間隔．丸が状態，丸の中の数値が日を表す.

2.4　日次取引において直面する不規則間隔の問題

日次取引におけるデータは，表 1 に示すように，時間ステップの間隔が一定ではない．しかしながら，通常の強化学習では，時間ステップの間隔は一定と仮定している．したがって，時間間隔の違いを学習に反映させることができない．

強化学習における時間間隔と実際の時間間隔の違いを図 2 に示す．丸が状態，丸の中の数値が日を表している．たとえば，最初の状態は 2004 年 1 月 9 日を表している．

日次取引における基本時間間隔は 1 日であるため，1 月 5 日から 1 月 9 日の間は問題は生じない．問題となるのは，1 月 9 日と翌営業日である 1 月 13 日の間が 4 日あることである．土日は市場が開かれないため，金曜日と月曜日の間隔は常に問題となる．また，1 月 12 日のように，休日も市場が開かれないため休日の前後の間隔も問題となる．

通常の強化学習では，このように時間感覚が異なっていたとしても，同じ 1 ステップとして扱うため，これらの違いを考慮することができない．たとえば，1 月 14 日に報酬を獲得すると 1 月 13 日，1 月 9 日，1 月 8 日と時間間隔を考慮せずに信用が割り当てられてしまう．これを，本論文では不規則間隔の問題と呼ぶ．

これまでの週次終値を用いた取引でも金曜日が休日の場合には不規則間隔の問題が生じていたが，日次取引ほど時間間隔が大きく異ならないことから，この問題については考慮されていなかった．

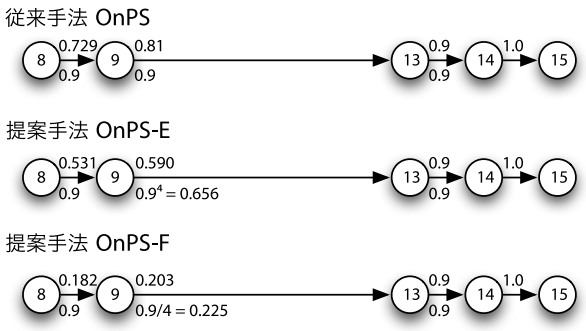

図 3　従来手法と提案手法．上段は基本割引率 $\gamma = 0.9$ のときの信用割当，下段は各ステップにおける割引率 γ_t を表す．

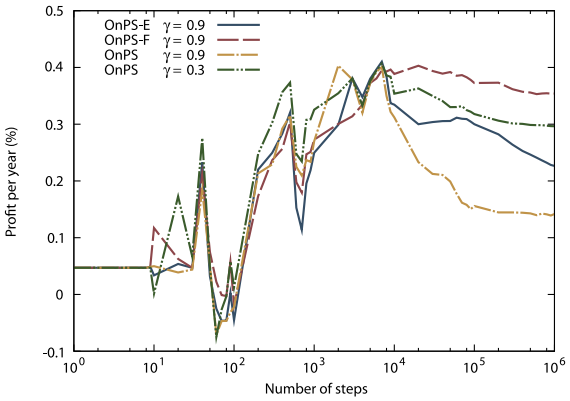

図 4　提案手法と従来手法による学習曲線の比較．

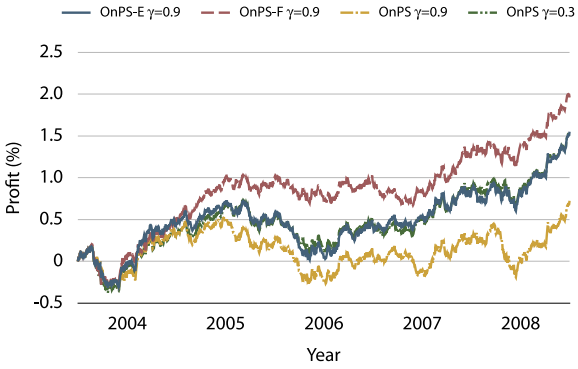

図 5　提案手法と従来手法による利益曲線．

3. 提案手法

OnPS が日次取引において不規則間隔の問題に直面するのは，常に割引率 γ が一定のためである．そこで，本論文では，割引率を時間ステップごとに変化させることによって，不規則間隔の問題を解決することを提案する．

提案手法では，次のようにして計算した割引率を用いる．まず，強化学習における基本ステップ間隔を Δ として定める．たとえば，日次取引であれば Δ は 1 日，週次取引であれば Δ は 7 日とする．そして，ステップ t とステップ $t+1$ のステップ間隔 δ_t とする．たとえば，1 月 9 日と 1 月 13 日の時間間隔は 4 日であるから，このときのステップ間隔 δ_t は 4 日である．この Δ と δ_t から，ステップ t における割引率 γ_t を求める．

本論文では，割引率をステップごとに変化させる手法を 2 種類提案する．ひとつは，基本ステップ間隔ごとにステップがあったものと仮定して，その分を信用割当から減衰させる手法である．もうひとつは，実際のステップ間隔と基本ステップ間隔の比に応じて割引率を決める手法である．

ひとつ目の手法では，次のようにしてステップ t における割引率 γ_t を求める．

$$\gamma_t = \gamma^{\frac{\delta_t}{\Delta}}$$

ここで，γ は割引率を変化させないときに用いられる基本割引率である．たとえば，基本割引率が $\gamma = 0.9$ で日次取引における 2004 年 1 月 9 日の場合，Δ は 1 日，δ_t は 4 日であるから，$\gamma_t = 0.9^4 = 0.656$ となる．この様子を図 3 に示す．本論文では，この手法を OnPS-E と呼ぶ[*ii]．

二つ目の手法では，次のようにして割引率 γ_t を求める．

$$\gamma_t = \frac{\Delta}{\delta_t}\gamma$$

たとえば，上の例と同じ基本割引率 $\gamma = 0.9$ で日次取引における 2004 年 1 月 9 日の場合，$\gamma = 0.25 \times 0.9 = 0.225$ となる．図 3 にはこの様子も示されている．本論文では，この手法を OnPS-F と呼ぶ[*iii]．

これらの手法により，実際の時間間隔が大きいステップでは割引率を小さくして信用割当を減らすことが可能となり，不規則間隔の問題を解決することが可能となる．

4. 実験結果

提案手法の有効性を確認するため，次のような実験を行った．実験データには，図 1 に示した 2004 年から 2008 年までの残存期間 10 年の日本国債（JGB10Y）の金利の日次終値を用いた．観測は，金利と 20 日営業日から求めたボリンジャー・バンドの幅とし，特徴数はそれぞれの次元に 5 つ，すなわち $5 \times 5 = 25$ とした．

基本割引率を $\gamma = 0.9$ として，二つの提案手法 OnPS-E, OnPS-F と従来手法 OnPS を比較した．また，提案手法は時間間隔が大きいときに割引率を小さくする手法であるため，従来手法において基本割引率が小さいときと比較した．このときの基本割引率は $\gamma = 0.3$ とした．

ランダム・シードを変えて 10 回の実験を行い，その平均を求めた．結果を図 4 に示す．横軸は学習したステップ数，縦軸は 1 年あたりの平均利益を表す．

実験の結果，OnPS-F が最も優れていることがわかった．OnPS-E は同じ基本割引率の従来手法よりは優れていたものの，基本割引率を小さくした従来手法よりは劣る結果となった．

ランダム・シードを変えて行った 10 回の実験のうち，最も最終利益が大きかった取引戦略の利益曲線を図 5 に示す．このグラフから，OnPS-F が獲得した取引戦略は，運用開始直後の金利が急上昇する局面で利益がマイナスとなるものの，その後は比較的安定して利益を出していることがわかる．また，このときに OnPS-F が獲得した取引戦略と同じ割引率（$\gamma = 0.9$）の OnPS が獲得した取引戦略を図 6 に示す．縦軸が金利，横軸がボリンジャー・バンドの幅を表している．赤い丸がロング・ポ

*ii　Exponentiation の頭文字をとった．
*iii　Fraction の頭文字をとった．

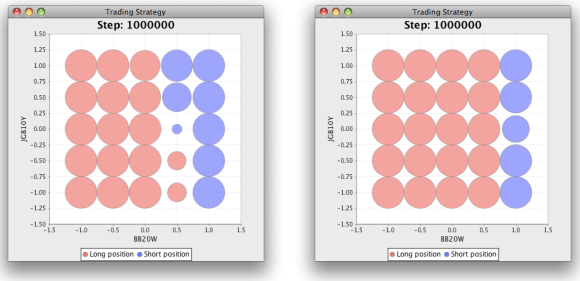

図 6　OnPS-F が獲得した取引戦略（左）と $\gamma = 0.9$ の OnPS が獲得した取引戦略（右）．

ジション，青い丸がショート・ポジションを表し，半径がポジションの絶対値の大きさを表している．

この結果は，ステップ間の時間間隔に応じて割引率を変化させるというアイディアが有効であることを示している．基本割引率が大きいとき，OnPS-E では時間間隔が大きくなっても割引率がそれほど小さくならない．図 3 に示したように，基本割引率が $\gamma = 0.9$ のとき，時間間隔が 4 倍になったとしても OnPS-E の割引率は $\gamma_t = 0.656$ とまだ大きい．これに対して，OnPS-F の割引率は $\gamma_t = 0.225$ とかなり小さくなる．このことが，OnPS-E の結果がそれほど良くなかった理由であると考えられる．

5. 関連研究

これまでに，強化学習を金融市場における取引戦略の獲得に応用する研究がいくつか行われている．Sherstov と Stone [Sherstov 04] は，強化学習を用いて Penn Exchange Simulator (PXS) の上で動く三種類の自律型株式取引エージェントを構築し，比較を行った．Bertoluzzo と Corazza [Bertoluzzo 07] は，回帰型強化学習を用いて株式市場のインデックスを対象とした取引戦略を獲得する研究を行った．Lee ら [Lee 07] は，日次の株式取引にマルチエージェント強化学習のアプローチを応用した．Bates ら [Bates 03] は，進化計算と強化学習を組み合わせた手法で為替取引（FX）の注文の流れを解析した．また，Gold [Gold 2003] は高頻度の為替取引のための取引戦略を回帰型強化学習を用いて獲得することを試みた．このように金融市場における取引戦略の獲得を対象とした研究がいくつか行われているものの，債券市場における取引戦略を対象とした研究は本研究が最初である．

強化学習の分野では，不規則間隔の問題はセミ・マルコフ決定過程（SMDP）環境としてこれまでに研究が行われている．Sutton ら [Sutton 99] は，不規則間隔で発生するイベントごとに行動選択を行う場合について調査し，SMDP 環境において行動価値を推定する方法について述べている．Stone と Sutton [Stone 01] は，RoboCup サッカーにおけるエージェントの行動規則の学習にエピソード型の SMDP を対象とした強化学習を用いた．しかしながら，profit sharing を SMDP に適用する研究はこれまで行われていない．

6. まとめ

本論文では，強化学習を用いて日次の金融市場取引戦略を獲得する際に直面する不規則間隔の問題を取り上げ，時間間隔に応じて OnPS の割引率を変化させる手法を提案した．日本国債を対象とした実験の結果，基本時間間隔との比に応じて割引率を変化させると，従来手法よりも優れた取引戦略が得られることがわかった．

様々な学習パラメーターを用いた実験を行って提案手法の特性を明らかにすること，株式市場や為替市場など他の金融市場を対象とした実験を行って提案手法の一般的な有効性を確認することなどが今後の課題である．

参考文献

[Bates 03] Bates, R.G., Dempster, M.A.H., and Romahi, Y.S.: “Evolutionary reinforcement learning in FX order book and order flow analysis”, In Proc. of the 2003 IEEE Int’l Conf. on Comput. Intell. for Finan. Eng. (CIFEr 2003), pp. 355–362 (2003)

[Bertoluzzo 07] Bertoluzzo, F. and Corazza, M.: “Making financial trading by recurrent reinforcement learning”, In Proc. of the 11th Int’l Conf. on Knowl.-Based Intell. Inf. and Eng. Syst. (KES 2007), pp. 2:619–626 (2007)

[Gold 2003] Gold, C.: “FX trading via recurrent reinforcement learning”, In Proc. of the 2003 IEEE Int’l Workshop on Comput. Intell. for Finan. Eng. (CIFEr 2003), pp. 363–370 (2003)

[Lee 07] Lee, J.W., Park, J., O, J., Lee, J., and Hong, E.: “A multiagent approach to Q-learning for daily stock trading”, IEEE Trans. on Syst., Man and Cybern., Part A, 37(6):864–877 (2007)

[Matsui 03] Matsui, T., Inuzuka, N., and Seki, H.: On-line profit sharing works efficiently, in *Proc. of the 7th Int’l Conf. on Knowl.-Based Intell. Inf. & Eng. Syst. (KES-2003)*, LNAI 2773, pp. 317–324 (2003)

[Sherstov 04] Sherstov, A.A. and Stone, P.:“Three automated stock-trading agents: A comparative study”, In Proc. of the AAMAS 2004 Workshop on Agent-Mediated Electron. Commerce (AMEC 2004), pp. 173–187 (2004)

[Stone 01] Stone, P. and Sutton, R.S.: “Scaling reinforcement learning toward RoboCup Soccer”, In Proc. of 18th Int’l Conf. on Mach. Learning (ICML 2001), pp. 537–544 (2001)

[Sutton 98] Sutton, R.S. and Barto, A.G.: *Reinforcement Learning: An Introduction*, The MIT Press (1998), 三上貞芳, 皆川雅章 共訳: 強化学習, 森北出版 (2000)

[Sutton 99] Sutton, R.S., Precup, D., and Singh, S.: “Between MDPs and semi-MDPs”, *Artif. Intell.*, 112:181–211 (1999)

[田中 09] 田中 雅：トレーディング，Art and Logic, トレイダーズ証券, http://www.traderssec.com/learn/tanaka/ (2009)

[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 08a] 松井 藤五郎, 後藤 卓, 和泉 潔, 大和田 勇人：強化学習を用いた金融市場取引戦略分析システムの試作, 人工知能学会 ファイナンスにおける人工知能応用研究会 (第 1 回) 研究会資料, pp. 12–17 (2008)

[松井 08b] 松井 藤五郎, 後藤 卓, 和泉 潔, 大和田 勇人：強化学習を用いた債券取引戦略の獲得, 2008 年度人工知能学会（第 22 回）全国大会講演論文集, 2C3–1 (2008)

[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 (2009)