3H2-OS12b-2

人工データからの強化学習を用いた金融取引戦略の獲得

Learning A Financial Trading Strategy from Artificial Data using Reinforcement Learning

松井 藤五郎[*1] 後藤 卓[*2] 和泉 潔[*3]
Tohgoroh Matsui Takashi Goto Kiyoshi Izumi

[*1]とうごろう機械学習研究所 Tohgoroh Machine Learning Research Institute
[*2]三菱東京 UFJ 銀行株式会社 The Bank of Tokyo Mitsubishi UFJ, Ltd.
[*3]東京大学 University of Tokyo

This paper describes a new method to learn a trading strategy in financial market, from artificial data generated by random walk, using reinforcement learning. We use two probability distributions for random walk in this paper: Laplace distribution and nonparametric districution.

1. はじめに

筆者らは，これまでに，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方法を開発してきた [Matsui 09]．この手法は，観測パラメーターを相対化することによって，値が大きく変動する市場においても利用可能な取引戦略の獲得を可能とした．また，非常に少ないデータを有効に活用するために，エピソードの終端状態と次のエピソードの開始状態を重ね合わせる方法を提案した．

強化学習は試行錯誤を繰り返すことによって行動規則を学習する枠組みである．強化学習を用いて優れた行動規則を学習するには，非常に多くの試行錯誤，すなわち訓練データが必要となる．しかしながら，金融市場から観測できる実際のデータは非常に限られたものとなってしまう．たとえば，週次取引の場合，実際のデータは 1 年当たり 50 個ほどしかない．

[Matsui 09] の手法では多くの試行錯誤を行うために限られたデータを繰り返し用いて取引戦略を学習するために，過学習（オーバー・フィッティング）の問題が生じていた．つまり，獲得された取引戦略は，訓練データに対しては非常に良い性能を示すが，未知のデータに対してはあまり良い性能を示さないという問題があった．

そこで，本論文では，ランダム・ウォークを用いて時系列データを生成することで学習データが繰り返し用いられるという問題を解決する．実際のデータが少ないという欠点を補うことを試みる．ところが，ランダム・ウォークを単純に用いて強化学習に必要な時系列データをすべて生成すると，元の時系列とは乖離した時系列データとなってしまう．本論文では，この問題を解決するために，時系列データの生成と学習を繰り返し行う手法を提案する．そして，日本国債の週次取引を対象とした実験により，提案手法の有効性について議論する．

2. 強化学習を用いた取引戦略の学習

取引戦略の学習には，筆者らがこれまでに開発した手法 [Matsui 09] を用いる．観測値は，市場の値とその移動分散の 2 つとしている．今回は残存期間 10 年の日本国債の週次取引を対象としているため，その金利と 14 週移動分散（直近 14 週の金利の分散）が観測値となる．

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 $o_{t,n}$ を求める．観測値は連続な値なので，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．

強化学習の報酬はリターン（金利の前日比から 1 を引いたもの）である．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，エピソードの終端状態を次のエピソードの初期状態としている．

3. 人工時系列データの生成

本論文では，実データへのオーバー・フィッティングを避けるために，学習用のデータをリターンに基づくランダム・ウォークによって人工的に生成することを試みる．リターンの確率密度分布は，片側ラプラス分布と仮定し，あるいはノンパラメトリックな確率分布として実データから次のようにして推定する．

3.1 片側ラプラス分布

ラプラス分布（Laplace distribution）は，確率密度関数が

$$f(x) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right) \tag{1}$$

と表される連続的な確率分布である．ここで，平均は μ，分散は $2b^2$ である．

ラプラス分布は，両側指数分布とも呼ばれ，平均付近の確率密度がかなり高く，またガウス分布と異なり平均から離れても確率密度が極端に低くならない．この性質は，金融市場におけるリターンの分布の性質としては正規分布よりも好ましいと考えられる．ラプラス分布は平均を中心に左右対称の形をしているが，金融市場におけるリターンの分布は非対称であることが経験的に知られている．

そこで，本論文では，実データのリターンを非負と非正に分け，それぞれ同じ絶対値で符号が反転した値を生成し，それぞれのラプラス分布のパラメーターを最尤推定によって推定す

連絡先: 松井藤五郎, TohgorohMatsui@tohgoroh.jp,
http://とうごろう.jp, @TohgorohMatsui

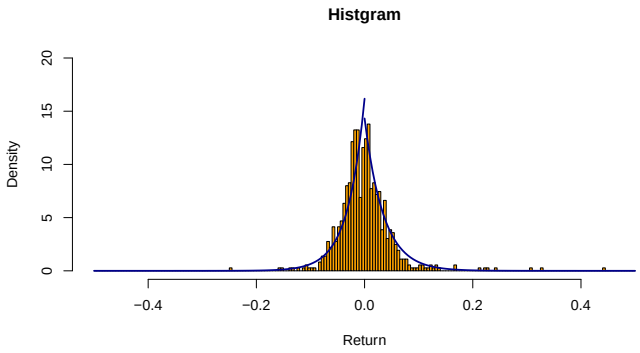

図 1: 1996 年から 2009 年までの残存期間 10 年の日本国債の週次の金利の確率分布とそこから推定した片側ラプラス分布の確率密度関数.

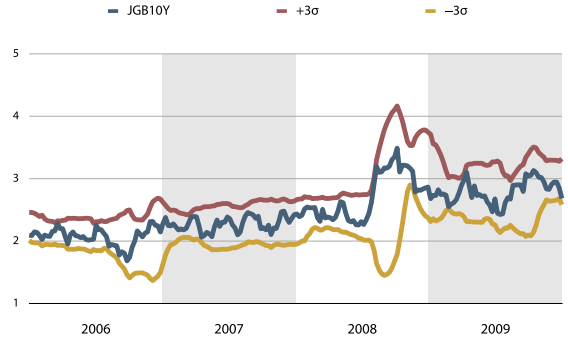

図 2: 2006 年から 2009 年までの残存期間 10 年の日本国債の週次の金利と $\pm 3\sigma$ のボリンジャー・バンド.

る．そして，非正，非負のラプラス分布を半分ずつ組み合わせて確率密度関数を合成する．図 1 に，1996 年から 2009 年までの実データのリターンのヒストグラムと，そこから推定した片側ラプラス分布の確率密度関数を示す．

3.2 ノンパラメトリックな確率分布

ノンパラメトリックな確率分布とは，リターンの実データに基づいて確率分布の形を推定する方法である．N 個の実時系列データ $x_0, \ldots, x_{N-1}$ から N 個の人工時系列データ $\dot{x}_0, \ldots, \dot{x}_{N-1}$ を生成するとき，0 以上 $N-1$ 未満の整数 u を一様なランダムに生成し，実データ x_u と x_{u+1} から実際のリターン $R_t = x_{u+1}/x_u - 1$ を求め，人工データのリターンとして用いる．つまり，この手法では，実データにおけるリターンのヒストグラムをそのまま確率密度関数として用いる．

3.3 人工時系列データの生成

強化学習には多数の学習データが必要であり，本論文ではこれをランダム・ウォークによって生成する．具体的には，リターンの実データから推定した確率密度分布に基づいてランダムにリターン R_t を生成し，一つ前の人工データ $\dot{x}_t$ にグロス・リターン $1 + R_t$ をかけることによって次の人工データ $\dot{x}_{t+1}$ を生成する．ところが，100 万ステップもの時系列データを生成すると，初期値からの乖離が大きくなってしまうという問題が生じる．

そこで，本論文では，乖離が大きくならない程度の一定期間のデータを生成してそこから学習を行うことを繰り返す方法を提案する．これによって，人工時系列データが実データから大きく乖離することを防ぐことができる．

4. 実験結果

提案手法の有効性を確認するために，残存期間 10 年の日本国債の週次取引を対象として，次のような実験を行った．人工時系列データのリターンの確率分布を推定するための実データの期間は 2006 年から 2008 年とした．図 2 にこの期間の実データの金利の推移と $\pm 3\sigma$ のボリンジャー・バンドを，図 3 に，この期間の実データのリターンのヒストグラムとそこから推定した片側ラプラス分布の確率密度関数を示す．

強化学習アルゴリズムには複利型の OnPS と Q 学習 [松井 10] を用い，100 万ステップの間に学習された取引戦略を 2008 年の実データと 2009 年の実データに対して適用してその性能を

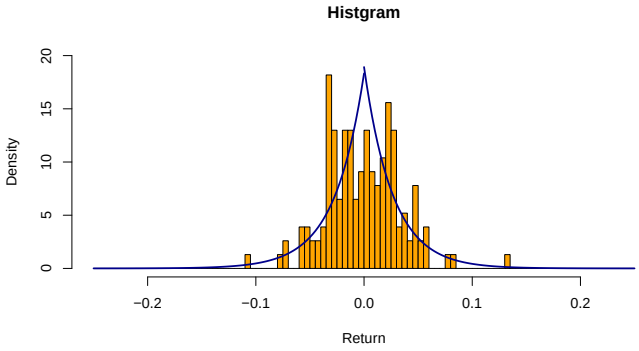

図 3: 2006 年から 2008 年までの残存期間 10 年の日本国債の週次の金利の確率分布とそこから推定した片側ラプラス分布の確率密度関数.

調べた．学習時の行動選択には温度 $\tau = 0.2$ の Gibbs ソフトマックス選択を用いた．複利型 OnPS の評価時は温度 $\tau = 0.01$ の Gibbs ソフトマックス選択から導かれる行動選択確率に基づく期待ポジションを取った．複利型 Q 学習に対しては温度 $\tau = 0.1$ の Gibbs ソフトマックス選択を用いて同様に評価を行った．

実験はランダム・シードを変えて 5 回行い，その中央値を求めた．その結果を図 4, 5 に示す．左のグラフは 2008 年の実データに対する評価，右のグラフは実データに対する評価であり，青い線はラプラス分布に基づいて生成した人工データから学習したときの学習曲線，赤い線はノンパラメトリックな確率分布に基づいて生成した人工データから学習したときの学習曲線，黄色い線は実データから学習したときの学習曲線を表している．

実データから学習したとき（黄色い線）は，複利型 OnPS と複利型 Q 学習の両方で学習データの一部である 2008 年に対しては高い性能を示すが学習データに含まれていない 2009 年には性能が極端に低くなるという過学習を引き起こした．ランダム・ウォークによる人工データから学習した場合は，いずれも，過学習を生じることはなかった．しかしながら，人工データから学習された取引戦略は，過学習された取引戦略よりは未知のデータに対する性能が良いものの，年間の利益は 0 に近く，総合的には良くも悪くもないというものであった．

参考として，複利型 OnPS および複利型 Q 学習が獲得した

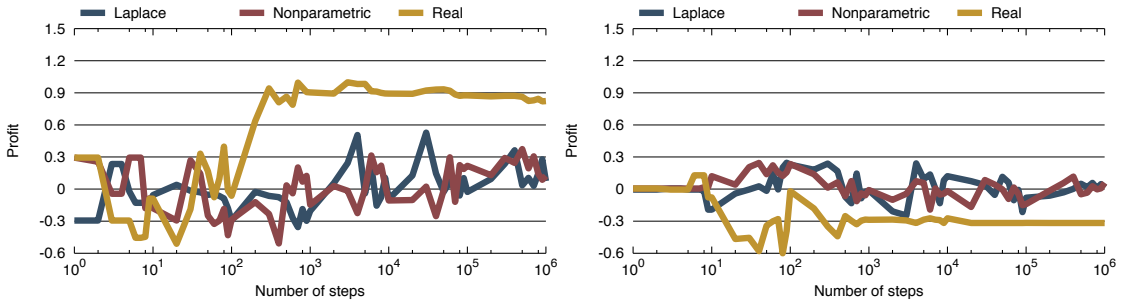

図 4: 複利型 OnPS による学習曲線．左は 2008 年，右は 2009 年の実データによる評価．

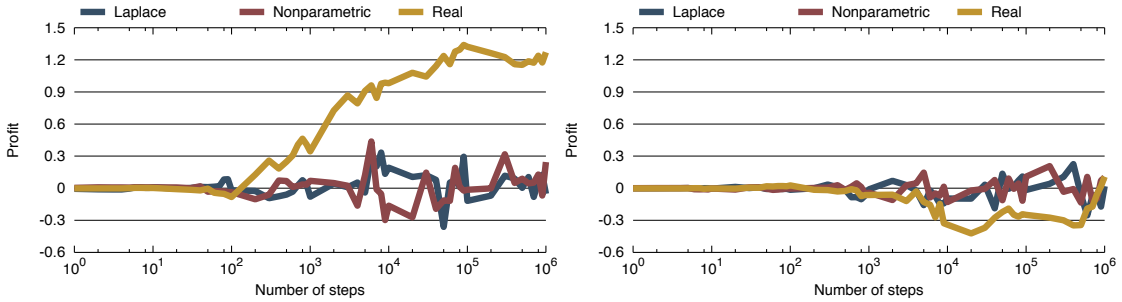

図 5: 複利型 Q 学習による学習曲線．左は 2008 年，右は 2009 年の実データによる評価．

取引戦略の例を図 6 に示す．縦軸が金利の水準，横軸が移動分散の水準を表す．また，赤い円は買いポジションを，青い円は売りポジションを表し，円の大きさがポジションの大きさを表す．

5. 考察

本論文では，ランダム・ウォークによって生成した人工時系列データから強化学習を用いて金融市場における取引戦略を学習する試みについて述べた．

上でも述べたように，人工データから学習された取引戦略は，絶対的な性能は良いとも悪いとも言えないものの，実データに対して過学習された取引戦略よりは未知のデータに対する性能が良いものであった．これは，ランダム・ウォークによって生成した人工時系列データから学習することによって，過学習を抑えることには成功したと見ることもできる．

ラプラス分布とノンパラメトリックな確率分布の差が大きく見られないことから，ランダム・ウォークの生成に用いる確率密度関数の違いはそれほど大きく影響しないと考えられる．しかしながら，分布の違いによる影響とデータ期間の違いによる影響はさらに調査する必要がある．

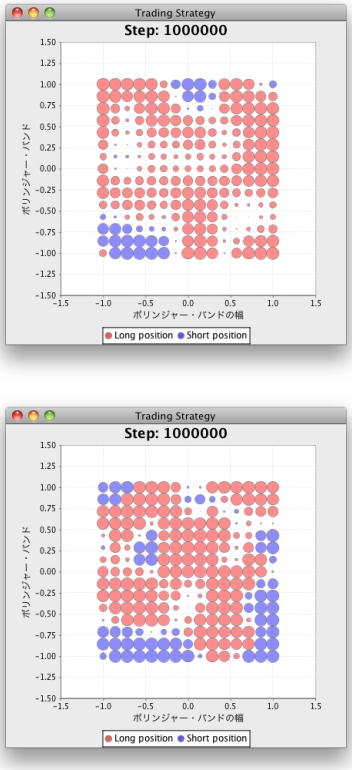

図 6: 獲得した取引戦略の例．上が複利型 OnPS，下が複利型 Q 学習．

参考文献

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

[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)

[松井 10] 松井 藤五郎：複利型強化学習, 2010 年度人工知能学会全国大会, 1A3–2 (2010)