3D9-5

強化学習を用いた株式取引エージェントにおける汎用政策の学習

Building A Generic Policy for Stock Trading Agents Using Reinforcement Learning

松井藤五郎　　大和田勇人
Tohgoroh Matsui　　Hayato Ohwada

東京理科大学 理工学部 経営工学科
Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science

This paper proposes a way to build a generic policy for stock trading agents using reinforcement learning. We have proposed a method to build a pair trading policy in Kaburobo. Pair trading is a basic strategy for stock trading, which is also used by investment funds. In this paper, we investigate the state space and the state distribution in the pair trading agent which we proposed, and shows the policy is useless for arbitrary stock pairs. We then propose that the states are represented by the percentage of difference from moving average.

1. はじめに

カブロボは，Java 言語を用いてプログラミングされたソフトウェア・ロボットによる仮想的な株式取引のプラットフォームである．カブロボでは，ロボット（ソフトウェア・エージェント）が，決められた額の仮想資金を元手に指定された銘柄を対象として実際の株式市場の値動きに連動した仮想的な取引を自動的に行い，その運用実績を競う．

カブロボは，2004 年に開催された第 1 回以来，毎回参加者を増やして発展し続けている．特に，昨年開催された第 1 回スーパー・カブロボ・コンテストは，優秀ロボット 10 体に 5,000 万円ずつ，計 5 億円を実際に投資した実運用を行って各方面の注目を集め，多くのメディアで取り上げられた [1, 2, 3, 4].

カブロボは，Java を用いてロボットを自由にプログラミングできることから，人工知能技術を株式取引の分野に応用するプラットフォームとして有用である．

筆者は，これまでに，強化学習アルゴリズムのひとつであるオンライン型 **profit sharing**（**OnPS**）[5, 7] をカブロボの行動学習に応用する方法を提案し [8]，このロボットを第 2 回カブロボ・コンテストに実際に参加させた [9].

本研究では，取引の対象を同業種の 2 銘柄に絞ったペア取引を行う．ペア取引は，ヘッジ・ファンドなども用いる基本的な取引手法の一つである．ペア取引の対象銘柄を固定し，一定量の成行注文だけに限定することによって，強化学習における行動を「買い」（主取引銘柄を買い，副取引銘柄を売る）と「売り」（主取引銘柄を売り，副取引銘柄を買う）の 2 つだけにできる．

しかしながら，これまでは，取引対象銘柄を固定していた．つまり，トヨタ自動車と本田技研工業の組を用いて学習し，同じ組に対して適用した．このようにして学習した政策は，他の銘柄の組に対して適用することができない，すなわち，汎用性が低いと考えられる．そこで，本論文では，従来手法で学習した政策の汎用性が低いことを示し，より汎用性が高い政策を学習する方法を提案する．

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

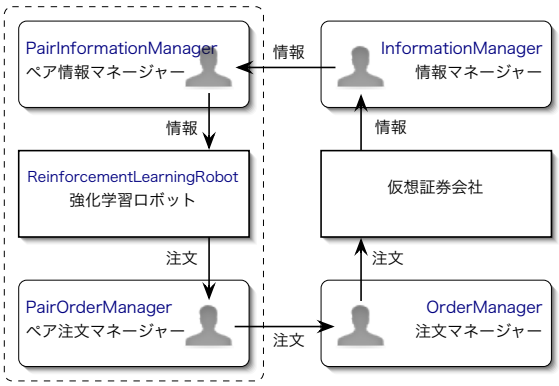

図 1　強化学習ロボットと仮想証券会社の関係．点線の枠内が構築した部分.

2. 強化学習を用いた株式取引エージェント

本研究では，カブロボにおける注文決定問題を逐次意思決定問題としてとらえ，情報を参照できる最古の時刻を $t = 0$，時刻 t における状態 s_t のうちの観測可能な情報を I_t としたとき，ロボットが，各時刻 t において，情報 $\mathcal{I}_t = \{I_0, I_1, \ldots, I_t\}$ に基づいて，目標時刻 $T > t$ における効用 u_T が最大になるように，その時刻に出す注文の銘柄 $\mathfrak{s} \in \mathfrak{S}$，種類 $\mathfrak{t} \in \mathfrak{T}$，価格 $\mathfrak{p} \in \mathbb{N}$，株数 $\mathfrak{n} \in \mathbb{Z}$ の組 $\langle \mathfrak{s}, \mathfrak{t}, \mathfrak{p}, \mathfrak{n} \rangle$ の集合 $\boldsymbol{O}_t$ を決定する問題だと考えた．ここで，$\mathfrak{S}$ は注文可能銘柄の集合，$\mathfrak{T}$ は注文種別の集合，$\mathbb{N}$ は自然数の集合，$\mathbb{Z}$ は整数の集合を表す．本研究では，ペア取引を行うことによって，注文決定問題における銘柄 $\mathfrak{s}$，種別 $\mathfrak{t}$，価格 $\mathfrak{p}$ および株数 $\mathfrak{n}$ の絶対値を考慮する必要が無くなるため，ロボットは $\mathfrak{n}$ の正負のみを決定すればいいことになる．

本研究で構築したロボットの概要を図 1 に示す．このロボッ

1. すべての i（$i = 1, \ldots, n$）について：
   $\theta(i) \leftarrow \frac{1}{|\mathcal{A}||\mathcal{T}|}$
2. 状態 s を初期化する
3. 各エピソードに対して：
4. 　すべての i（$i = 1, \ldots, n$）について：
   $c(i) \leftarrow 0$
5. 　各ステップに対して繰り返し：
6. 　　すべての i（$i = 1, \ldots, n$）について：
   $\phi_{s,a}(i) \leftarrow \exp\left(-\frac{\|s - o_{a,i}\|^2}{2{\sigma_{a,i}}^2}\right)$
7. 　　$P(s,a) \leftarrow \sum_{i=1}^{n} \theta(i)\,\phi_{s,a}(i)$
8. 　　P から導かれる確率分布に従って s での行動 a を選択する
9. 　　すべての i（$i = 1, \ldots, n$）について：
   $c(i) \leftarrow c(i) + \phi_{s,a}(i)$
10. 　　行動 a を取り，報酬 r と次状態 s' を観測する
11. 　　$\vec{\theta} \leftarrow \vec{\theta} + \alpha r \vec{c}$
12. 　　$\vec{c} \leftarrow \gamma \vec{c}$
13. 　　$s \leftarrow s'$
14. 　s が終端状態ならば繰り返しを終了

図 2　株式取引のための強化学習アルゴリズム．n は特徴数，$\mathcal{A}$ は行動の集合，$\mathcal{T}$ は格子の集合，α はステップ・サイズ・パラメータ，γ は割引率パラメータ，$o_{a,i}$ と $\sigma_{a,i}$ は動径基底関数の中心と幅を表す．

トは，情報マネージャーを通して状態と報酬に関する情報を取得し，行動を決定し，注文マネージャーを通して仮想証券会社に注文する．

強化学習ロボットの学習アルゴリズムを図 2 に示す．ロボットの行動選択には，行動優先度 P を正規化した上で Gibbs 分布によるソフトマックス選択を用いた．

強化学習における報酬は，資産評価額が前日よりも増えた場合に +1 の報酬を与える．また，資産評価額が前日より減少した場合にエピソードを打ち切る．これにより，資産を減少させる原因となった行動は無視して，資産の増加につながった行動だけを強化できる．

強化学習における状態は，従来手法では，2 銘柄の価格比を正規化したレシオと割引ゴールデン・クロスを用いて表現した．これらを格子状に配置した動径基底関数（RBF）を用いた関数近似 [6] によって表現している．しかしながら，このように状態を定義すると，学習時に取引対象とした銘柄と異なる銘柄に対して適用できない，すなわち，汎用性が低い政策が学習されてしまう．

3.　状態空間と汎用性

図 3 と図 4 に，2006 年 4 月から 2007 年 3 月までの，トヨタ自動車（トヨタ），本田技研工業（ホンダ），日立製作所（日立），東芝，三菱電機（三菱）の株価（終値）を示す．この株価は，Yahoo!ファイナンスの株価時系列データ [10] から株価データを取得し，株式分割による影響を取り除いた調整後株価に変換したものである．この図から，同業種の株価は良く似た値動きをしていることがわかる．

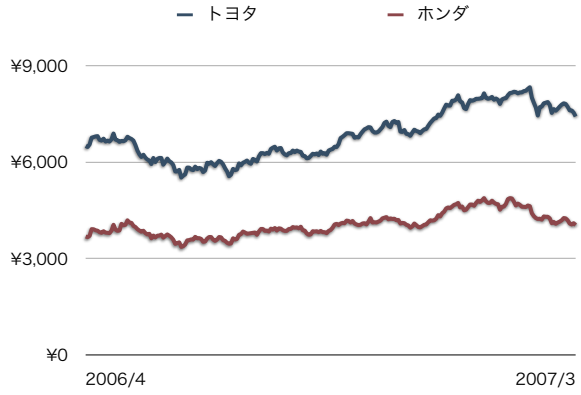

図 3　2006 年 4 月から 2007 年 3 月までのトヨタ自動車と本田技研工業の株価．

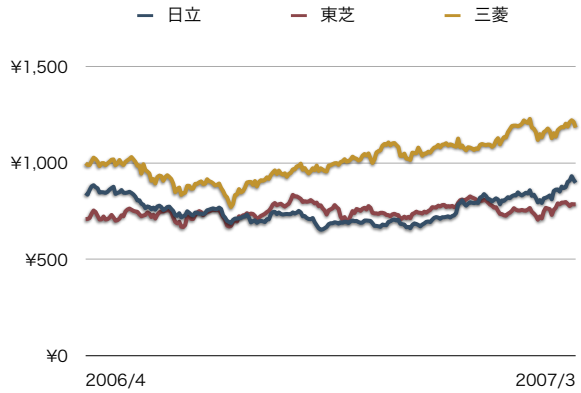

図 4　2006 年 4 月から 2007 年 3 月までの総合電機メーカー 3 社（日立製作所，東芝，三菱電機）の株価．

従来 [8] は，取引対象銘柄の株価の比に着目していた．トヨタとホンダ，日立と東芝，東芝と三菱，三菱と日立について，株価比と移動平均比を −1 から +1 に正規化した値を用いて状態を表現したパラメータとした時の状態分布を図 5 に示す．この図より，それぞれの銘柄の組が取り得る状態の分布に重なりが少ないことがわかる．つまり，このように状態を表現したときには，ある銘柄の組を取引して学習した政策は，他の銘柄の組の取引には適用できない汎用性の低いものとなってしまう．たとえば，東芝と三菱の組を取引して政策を学習し，その政策をトヨタとホンダの組の取引に適用しても，同じ状態を訪れることがないために学習の成果はほとんど反映されない．

そこで，本論文では，学習対象銘柄以外の組を取引する際にも適用可能な汎用政策を学習するために，銘柄ごとに大きく異なる値を取らない指標を用いて状態を表現することを提案する．例として，移動平均乖離率を用いて状態を表現した時の状態分布を図 6 に示す．移動平均乖離率とは，現在の株価が移動平均からどれだけ離れているかを表し，通常は，次のように計算さ

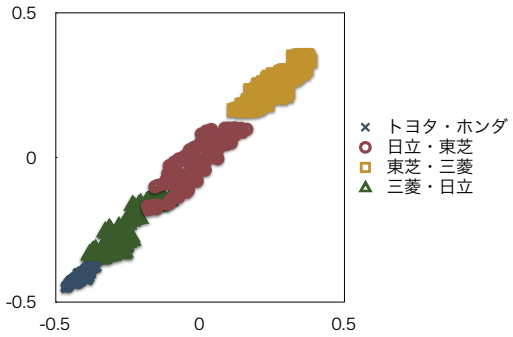

図 5　株価比と移動平均比を正規化したもので状態を表したときの状態分布．x 軸は株価比，y 軸は移動平均比をそれぞれ正規化した値を表す．

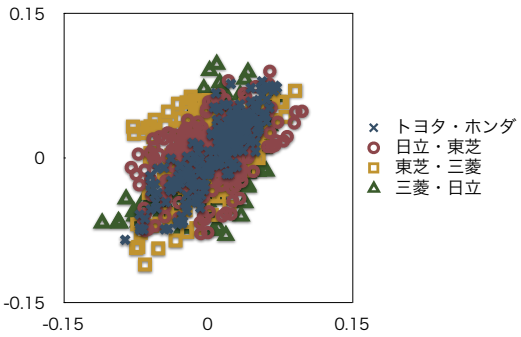

図 6　移動平均乖離率を用いたときの状態分布．x 軸は主取引銘柄の移動平均乖離率，y 軸は副取引銘柄の移動平均乖離率をそれぞれ正規化した値を表す．

れる．

$$\text{移動平均乖離率} = \text{株価}/\text{移動平均} - 1$$

本研究では，移動平均乖離率を値域が -1 から 1 までとなるように正規化し，

$$\text{移動平均乖離率} = \begin{cases} \text{株価}/\text{移動平均} - 1 & \text{if 株価}/\text{移動平均} \leq 1 \\ 1 - \text{移動平均}/\text{株価} & \text{otherwise.} \end{cases}$$

とする．

図 6 の状態分布は，図 5 と比べ，異なる組み合せを用いた場合でも状態分布が重なっている．したがって，ある銘柄の組を取引して学習した政策は，他の銘柄の組の取引にも適用できる．

4. 考察

本論文では，強化学習を用いて株式取引の政策を獲得するエージェントに対し，学習した銘柄とは異なる銘柄に対しても適用可能な汎用政策を学習するために，状態空間について検討した．その結果，株価などのように，銘柄ごとに大きく異なるパラメータを用いて状態を表現した場合には，異なる銘柄を対象としたときに状態分布が重ならないことを確認した．すなわち，エージェントは，学習した銘柄と異なる銘柄を対象として株式取引を行うと，学習時に訪れた状態を訪れることがないという問題が生じる．

これを解決する方法として，本論文では，移動平均乖離率のような，どのような銘柄に対しても同じような値を取るパラメータを用いて状態を表現することを提案した．提案手法を用いることによって，政策学習時と政策適用時で同じような状態を訪れることができるようになる．ただし，どの状態も良く似たパラメータで表現されることになるため，状態の区別がつきにくくなってしまう．

今後は，この状態空間において有効な政策を学習する方法について，特に，エピソードの区切り方と報酬の与え方について検討する必要がある．

参考文献

[1] 株自動売買ソフトで学ぶ. *AERA*, 2006 年 8 月 7 日号, p. 69, 2006.

[2] Is your fund manager plugged in? *Business Week*, 2006 年 8 月 8 日号, p. 14, 2006.

[3] チャートは使い方次第で天国と地獄!? 日経マネー, 2006 年 10 月号, p. 55, 2006.

[4] カブロボ，飛翔—2007 年夏，ロボットが市場を席巻する．ITmedia エンタープライズ．http://www.itmedia.co.jp/enterprise/articles/0612/19/news010.html, 2006.

[5] T. Matsui, N. Inuzuka, and H. Seki. On-line profit sharing works efficiently. In *Proc. of the KES-2003, LNAI-2773*, pp. 317–324, Springer, 2003.

[6] R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 1998. 三上, 皆川 共訳. 強化学習. 森北出版, 2000.

[7] 松井, 犬塚, 世木. 線形関数近似を用いた profit sharing 強化学習法. 2002 年度人工知能学会全国大会, 2D3-03, 2002.

[8] 松井, 大和田. 株式取引エージェントへの強化学習の応用. 2005 年度人工知能学会全国大会, 1D4-1, 2005.

[9] 松井, 大和田. 強化学習を用いた株式取引エージェントの評価. 2006 年度人工知能学会全国大会, 3C1-6, 2006.

[10] Yahoo!ファイナンス．時系列データ, 2006. http://quote.yahoo.co.jp/.