F_006

強化学習を用いた株式取引シミュレーション
Stock Trade Simulation with Reinforcement Learning

松井 藤五郎* 大和田 勇人*
Tohgoroh Matsui Hayato Ohwada

1 はじめに

筆者らは，これまでに，カブロボ・コンテスト [3] を対象として，強化学習を用いた株式取引エージェントを開発し，株式取引シミュレーションを行ってきた [4, 5]. 本論文では，これまでに行ったシミュレーションの結果から得られた知見に基づいて，強化学習を用いて株式取引を行う方法を示す.

カブロボ・プログラミング・コンテスト（略称：カブロボ）は，株式取引を対象としたソフトウェア・プログラミング・コンテストである．これまでに 2 回のコンテストが開催され，多数のチームが参加した．今年からスーパー・カブロボと名を改め，手数料や機関投資家に適用される規制などが考慮されたより大規模で実際的なシミュレーションが可能となった.

スーパー・カブロボの参加者は，株式取引を行うソフトウェア・ロボット（エージェント）を作成して提出する．提出されたロボットは，東証の 500 銘柄を対象として前場と後場の前にそれぞれ注文を決定し，所持金 5,000 万円からの運用成績を競う.

2 強化学習問題の設定

本研究では，取引の対象を同業種の二銘柄に絞り，二銘柄の価格比に着目したレシオ取引を行う．同業種の二銘柄に着目するペア取引は，ヘッジ・ファンドなども用いる基本的な取引戦略の一つである.

2.1 状態

本論文では，エージェントが観測したパラメータ（株価などの指標）を状態と呼んでいる．エージェントは一部のパラメータだけを観測しているので，正確には部分観測環境である.

本論文では，状態を二銘柄の価格比を正規化した値

$$\rho(t) = \begin{cases} p_s(t)/p_m(t) - 1 & p_s(t)/p_m(t) \leq 1 \text{ のとき} \\ 1 - p_m(t)/p_s(t) & \text{そうでないとき} \end{cases}$$

とその MACD2 $\mu(t)$ を用いて状態を表す．ここで，$p_m(t)$ は時刻 t の主取引銘柄の株価，$p_s(t)$ は同じく副取引銘柄の株価を表す．正規化することによって，主取引銘柄と副取引銘柄を入れ替えたときの絶対値を等しくするとともに，値域を $[-1, 1]$ とできる．MACD2 は，移動平均線による分析を発展させたテクニカル分析手法の一つであり，トレンドの転換を素早く察知することができる．短期の指数平滑移動平均（EMA）と長期の EMA の差を MACD と呼び，MACD の EMA をシグナルと呼ぶ．MACD2 は，MACD とシグナルの差である.

最終的には，$(\rho(t), \mu(t))$ を格子状に配置した動径基底関数（RBF）を用いた関数近似 [2] によって近似し，状態集合 S を表現する．本研究では，翌日の状態 $(\rho(t+1), \mu(t+1))$ が前の状態 $(\rho(t), \mu(t))$（と行動 $a(t)$）だけに依存する MDP であると仮定している.

2.2 行動

ペア取引を対象としているため，行動は「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）の二つだけである．すなわち，行動集合 $A = \{\text{買い}, \text{売り}\}$ となる.

注文量は，取引可能金額 $M_A(t)$ の一定割合を上限金額として，それぞれ $\lfloor \beta M_A(t)/2p_m(t)u_m \rfloor$, $\lfloor \beta M_A(t)/2p_s(t)u_s \rfloor$ とする．ここで，β は注文割合，u_m，u_s はそれぞれの単元株数である.

前回の行動と今回の行動が異なる場合は，まず保有している株の反対取引を行い，その後に今回選択した行動に基づく取引を行う．同じ場合は株を保有し続け，取引は行わない.

2.3 報酬

カブロボにおけるロボットの評価は総資産額を基にして行われるため，総資産額を基にして報酬を計算するのが自然である．総資産額の比を，シグモイド関数を用いて -1 から 1 の値に変換し，これを報酬とする.

$$r(t+1) = \frac{1 - \exp\left(-\kappa \frac{M_T(t+1)}{M_T(t)}\right)}{1 + \exp\left(-\kappa \frac{M_T(t+1)}{M_T(t)}\right)}$$

* 東京理科大学 理工学部 経営工学科

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1　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．

ここで，κ はシグモイド関数の傾きを表す定数を表す．

正の報酬または負の報酬が続く間のみ報酬を伝播させることとし，報酬が正から負または負から正に変わった時点でエピソードを終了させる．

3　強化学習アルゴリズム

本研究では，強化学習アルゴリズムにオンライン型 profit sharing（OnPS）[1] を用いる．OnPS は，行動優先度学習型の強化学習アルゴリズムであり，Q 学習や Sarsa(λ) など行動価値推定型のアルゴリズムに比べて少ない試行錯誤から学習できるという特徴を持っている．また，従来の（オフライン型）profit sharing は，目標状態が定義可能なエピソード型タスクにしか適用できないため，そのままカブロボのタスクに適用することはできない．OnPS は，少ない試行錯誤から学習でき，かつ，非エピソード型タスクにも適用可能なことから，株式取引に適した強化学習アルゴリズムである．

OnPS を基にした株式取引のための強化学習アルゴリズムを図 1 に示す．行動 $a(t)$ の選択には，Gibbs 分布によるソフト・マックス選択 [2] を用いる．学習効果を大きく反映させるため温度パラメータを $\tau = 0.1$ とし，わずかな優先度の違いでも行動選択確率が大きく変わるようにしている．また，ステップ・サイズは重ねた格子の枚数から $\alpha = 0.1$ とし，割引率は早く学習するよう $\gamma = 0.9$ と比較的大きい値にしている．

4　考察

これまでのシミュレーション [4, 5] では，保有株を決済せずに注文を重ねていたため，取引可能金額が不足して取引ができないことがあった．これは，強化学習から見ると，選択した行動が結果に反映されないことを意味しており，学習に支障を来していると考えられる．そこで，本論文では，前回と行動が変わった際に保有株の反対取引を行って決済することで，取引可能金額が不足しないようにした．

また，これまでは，過去の状態・行動対の学習のために得られた報酬をすべて伝播させていた．このため，直前までの長期間にわたって資産総額が増え続けていても，資産総額が減少したときには負の報酬が伝播していた．そこで，本論文では，正の報酬または負の報酬が続く間のみ伝播させるようにした．

さらに，これまでは，注文量を一定の単元としており，主取引銘柄と副取引銘柄の株価が大きく異なる場合には，株価の大きい銘柄に強く依存してしまっていた．そこで，本論文では，出来る限り同じ金額分の取引を行うよう注文量を変更した．

前者の二つは強化学習にとっては重大な問題点である．これらの変更により，より適切な強化学習が行われると考えられる．

参考文献

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

[2] R. S. Sutton and A. G. Barto. *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.

[3] カブロボ, 2004–2006. http://kaburobo.jp/.

[4] 松井藤五郎, 大和田勇人. 株式取引エージェントの強化学習への応用. 2005 年度人工知能学会全国大会（第 19 回）論文集, 1D4-1, 2005.

[5] 松井藤五郎, 大和田勇人. 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（第 20 回）論文集, 3C1-6, 2005.