人工知能学会研究会資料
SIG-FIN-011-04

強化学習を用いたブーム検知型株トレーディングシステムの構築

Trading System of Boom Detection using Reinforcement Learning

中原 孝信[1*]　宇野 毅明[2]　岡田 克彦[3]　羽室 行信[3]
Takanobu NAKAHARA[1]　Takeaki UNO[2]　Katsuhiko OKADA[3]　Yukinobu HAMURO[3]

[1] 関西大学 データマイニング応用研究センター
[1] Data Mining Applied Recerch Center, Kansai University
[2] 国立情報学研究所 情報学プリンシプル研究系
[2] Principles of Informatics Research Division, National Institute of Informatics
[3] 関西学院大学 経営戦略研究科
[3] Institute of Business and Accounting, Kwansei Gakuin University

Abstract: In this paper, we apply reinforcement learning algorithm to enhance return from technical trading in the Japanese stock market. Specifically, we employ MACD (Moving Average Conversion Diversion) signals to make buy / sell decision. MACD trading signal generates good return when the market is in a trending state but performs poorly when in a box-range state. Reinforcement learning endeavors to identify the state ex-ante and helps traders efficiently allocate capital. We demonstrate how we design our trading system and show the results of our simulations. Our results indicate the reinforcement learning of technical trading signals dramatically improves returns.

1　はじめに

強化学習は, 未知の環境や動的に変化する環境に置かれた学習対象 (エージェント) が, 試行錯誤を通じて得た自身の成功・失敗体験 (報酬) から価値を将来にわたって最大になるように行動方策を自律的に学習するエージェント学習の一手法である. 従来強化学習は、離散状態空間の課題に対して適用されてきたが、強化学習を実問題に適用する場合は、連続状態空間の問題として扱う必要があり、連続値を入力として価値関数と行動写像を関数近似器により近似する方法などが取られている [Sutton 98]。

本研究では株価動向を予測しながら株取引を行う投資モデルを構築するために、強化学習を応用する。近年ファイナンス領域における多くの研究の結果、株価動向には何らかの規則性が存在することがわかっており、「小型株効果」と「バリュー株効果」と呼ばれるものがある [Fama 92],[Fama 93]。これは長期的には時価総額が小さい小型株への投資収益が、時価総額が大きい大型株への投資収益よりも平均的に高いことを示す。

一方、小型株効果と独立して、「バリュー株効果」も存在する。バリュー株とは市場で評価される時価総額と会計上の株主価値である簿価総額の比率が低い株であり、帳簿上の価値よりも市場であまり大きく評価されていない銘柄を指す。これに対して「成長株」は、帳簿上の価値よりも市場評価が高い銘柄を指す。バリュー株への投資収益は成長株よりも平均的に高いことが知られており、これをバリュー株効果と呼ぶ。これらの効果は独立して存在するため、高い収益率を得るためには、小型株で且つバリュー株の銘柄群に投資することが薦められる。しかしながら、これらの効果は、長期では高い収益率が期待されるが、短期では不確実性が伴う。単年度の収支で評価されるプロの投資家は、毎年収益をあげることを顧客に要求されるため、短期的な収支管理が必要不可欠である。

本研究では、短期的な株価変動の規則性を時系列に捉えるために、ある指標によってグループ化した銘柄群に対して、日々の超過収益率の変動率を状態として扱うことで、 ブームを検知する。そして株取引における超過収益率が最大になるように 「売り」「買い」「様子見」「精算」という４つ行動から政策を学習する。

2　株取引エージェント

時々刻々と変化する株式市場において、株価はある一方向にトレンドを形成したり、方向感なく一定幅の

*連絡先：関西大学 データマイニング応用研究センター
吹田市山手町 3-3-35.
E-mail: nakapara@gmail.com

上下動（ボックスレンジ）を繰り返したりする。最近では、長期金利が下落したことを受けて、 不動産関連株にブームが発生し、継続的に値上がりする上昇トレンドを形成している。一方、不景気時は、収益率の高い製薬メーカーの株価は方向感のないボックスレンジの動きを示している。

本研究であつかうテクニカル指標（MACD) はトレンド形成時に収益率が高まるため、株価のブームを検知し、ブームに乗った銘柄群をポートフォリオに組み込むことで収益率の増加が期待できる。そこで、MACD を利用して上位・下位グループを作成し、超過収益率の変動を上下のグループ間で比較することでブームの検知を試みる。超過収益率の変動を表す指標として、過去 n 日の値を利用した時系列微分 Sharpe 比を提案する。時系列データ集合を $X = \{X_t | t = 1, 2, \cdots, n\}$、日 t における過去 n 日の時系列微分データ部分集合を $D_t(n) = \{X_t - X_{t-1}, X_{t-1} - X_{t-2}, \cdots, X_{t-n-1} - X_{t-n}\}$ とすると、時系列微分 Sharpe 比は式 (1) で表される。

$$SR_t(n) = \frac{mean(D_t(n))}{sd(D_t(n)) + 1.0} \tag{1}$$

ここで $mean$ は平均、sd は標準偏差である。分母の +1.0 はゼロ割り算を避けるためのものである。$SR_t(n)$ は、過去 n 日の増減値が等しい場合に、sd が 0 になるため、$mean$ と一致する。また、時系列データ集合 X の傾きによって $SR_t(n)$ の大きさは異なり、X が右上がりの場合は正に大きく、左下がりの場合は負に大きな値になる。

本研究では、時系列微分 Sharpe 比 ($SR_t(n)$) とセンチメントを利用して合計 3 つの値を状態として利用した。1 つ目は、上位下位グループの超過収益率の平均値を利用した $SR_t(n)$ をグループ毎に計算し、それを件数が均等になるように「高」「中」「低」 に離散化したものを利用した。2 つ目は、個別銘柄のセンチメント指数を計算し、その値を上位・下位グループで合計したものを連続値として利用した。センチメント指数は、ニュース記事テキストから作成される極性付き評価表現辞書に基づいた値であり、評価表現の出現頻度を日単位でカウントしたものである。評価表現辞書については筆者等が作成した辞書を利用した [Okada 12]。3 つ目は、日経 225 の時系列微分 Sharpe 比 ($SR_t(n)$) を計算し、それを正負で離散化した値を状態とした。

次に、強化学習における行動は、株取引で最低限必要な「購入」、「売却」、「様子見」、そして「精算」を加えた 4 種類の行動を扱う。

報酬は、取引した銘柄群の平均超過収益率を利用する。1 つの銘柄に対して「購入」または「売却」指示が出た日を s、「精算」指示の出た日を e とすると、超過収益率は式 (2) で定義される。

$$AR(s,e) = \begin{cases} (\frac{open_{e+1}}{open_{s+1}} - \frac{open225_{e+1}}{open225_{s+1}}) & \text{if } long \\ (\frac{open225_{e+1}}{open225_{s+1}} - \frac{open_{e+1}}{open_{s+1}}) & \text{if } short \end{cases} \tag{2}$$

ここで、$open$ は始値、$open225$ は日経 225 の始値を示す。$long$ は、購入を意味しており、日 $s+1$ で対象銘柄株を購入し日 $e+1$ で売却する取引を示している。$short$ は売却を意味しており、日 $s+1$ で空売りし日 $e+1$ で買い戻す取引である。そして、グループ内の取引対象となる銘柄群に対して超過収益率の平均を計算する。超過収益率の振れ幅は、上限下限が定まっていないため、収益率が異常に大きい (小さい) 場合の報酬に対する過度の影響を除外するために、標準シグモイド関数 $(1/1 + e^{-x})$ を $2 * (1/1 + e^{-x}) - 1$ とすることで、-1 から 1 に変換したものを報酬とした。

3 強化学習アルゴリズム

強化学習では、政策改善と価値推定が重要となる。政策改善は、よりよい政策を積極的に探索 (Exploration) すべきか、現在の報酬で満足し搾取 (Exploitation) すべきかを判断しなければならない。そこには、探索と搾取のトレードオフが存在している。政策改善は、一般的にはグリーディな政策ではなく、確率的な改善法が利用される。本稿では、政策改善に ϵ グリーディ政策を用いており、確率 1-ϵ でグリーディ政策により行動を選び、確率 ϵ で一様ランダムに行動を選ぶ。

価値推定は、強化学習の分野では価値関数を近似するために様々な手法が提案されている。もっとも知られている SARSA と Q 学習に適学度トレースを考慮した SARSA(λ) と Q(λ) の 2 種類を利用して株取引システムの実験を行った。まずは SARSA(λ) について説明する。$t+1$ 期に状態 s で行動 a が選択されたとき、得られる行動価値 $Q_{t+1}(s,a)$ は、式 (3) により更新される。

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha\delta_t e_t(s,a) \tag{3}$$

ここで、α は学習率であり、過去に得られた推定値と現在得られた結果をどの程度優先するかを表したパラメータである。一般的には α=0.1 の値を設定することが多い [Takadama 03]。また、式 (3) の δ_t は、

$$\delta_t = r_{t+1} + \gamma Q_t(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t)$$

である。r は報酬を表し、γ は割引率で報酬が将来になればなるほどその影響を割り引くパラメータを表す。一般的には γ=0.9〜0.99 を設定することが多い。また、すべての状態 s、行動 a に対して $e_t(s,a)$ は、

$$e_t(s,a) = \begin{cases} \gamma\lambda e_{t-1}(s,a) + 1 & (s = s_t \text{かつ } a = a_t \text{のとき}) \\ \gamma\lambda e_{t-1}(s,a) & (\text{それ以外のとき}) \end{cases}$$

である。適格度トレース $\lambda(0 \leq \lambda \leq 1)$ によって、k ステップ先までの情報を用いて行動価値 Q を更新する。λ が 1 に近づくにつれて k の値が大きくなり、多くの情報を用いて Q 値は更新されることになる。

次に Q(λ) は、SARSA(λ) と類似しているが政策オフ型のアルゴリズムであり、δ_t の更新は以下で行われる。

$$\delta_t = r_{t+1} + \gamma \max_a Q_t(s_{t+1}, a) - Q(s_t, a_t)$$

この式は、更新時に現在の状態と行動価値 $Q(s_t, a_t)$ を状態 s_{t+1} で最大の価値 $max_a Q(s_{t+1}, a)$ に近づけ、価値を伝播させることを意味する。Q 学習では価値関数の更新が政策に依存していないことから政策オフ型と呼ばれている。一方 SARSA は、$t+1$ に政策によって選択された行動が次の期の行動になるため、政策オン型のアルゴリズムと言われている。

ここまでは、状態として離散空間を対象にしてきたが、連続値空間を扱う場合には、 距離的に近い状態や行動では Q 値も近い値を持つと仮定すると、関数近似を用いることができる。関数近似によりこれまで経験したことのない状態に対しても、似た状態の経験を用いることで適切な行動の選択が可能となる。本研究では、連続値を扱う場合には、線形ガウス関数を利用した関数近似によって式 (5) のように特徴ベクトルとパラメータベクトルを掛け合わせることで行動価値を求める。ここで、特徴ベクトルは、$\overrightarrow{\phi}_{s,a} = (\phi_{s,a}(1), \phi_{s,a}(2), ..., \phi_{s,a}(n))^T$ であり、パラメータベクトルは、$\overrightarrow{\theta_t} = (\theta_t(1), \theta_t(2), ..., \theta_t(n))^T$ である。特徴ベクトルは、基底関数としてよく利用されるガウス関数 (式 (4)) を用いて求める。

$$\phi_{s,a} = \sum_{b=1}^{B} \exp(-\frac{\|\boldsymbol{x} - \boldsymbol{c}_b\|^2}{2\sigma^2}) \tag{4}$$

ここで、$\boldsymbol{x}$ は、状態 s と行動 a を合わせたベクトル $\boldsymbol{x} = (s, a)^T$ であり、B は基底数、$\boldsymbol{c}_b$ は、ガウス関数の中心ベクトル、σ はガウス関数の幅を決める標準偏差である。そして、パラメータベクトルを式 (6) を利用して更新することで、Q 値が更新される。

$$Q_t(s, a) = \overrightarrow{\theta_t^T} \overrightarrow{\phi}_{s,a} = \sum_{i=1}^{n} \theta_t(i) \phi_{s,a}(i) \tag{5}$$

$$\overrightarrow{\theta}_{t+1} = \overrightarrow{\theta_t} + a\delta_t \overrightarrow{e}_t \tag{6}$$

ここで、

$$\delta_t = r_{t+1} + \gamma Q_t(s_{t+1}, a_{t+1}) - Q(s_t, a_t)$$

そして、

$$\overrightarrow{e}_t = \gamma\lambda \overrightarrow{e}_{t-1} + \nabla_{\overrightarrow{\theta_t}} Q_t(s_t, a_t)$$

である。この方法は最急降下型 Sarsa(λ) と呼ばれ、収束が保証された方法である [Sutton 98]。

4 強化学習による株の取引実験

本研究では、2002 年 2 月〜2012 年 12 月までに取引された銘柄の中からブームを捉えるために MACD の高い上位・下位 3%の銘柄を上下グループとして利用し、上記の方法によって状態を作成した。MACD は、銘柄毎に短期 (12 日) の指数移動平均と長期 (26 日) の指数移動平均の差を計算した値であり、MACD が正の値の場合には上げトレンド、負の場合には下げトレンドを示している。

表 1: シミュレーションで利用した状態一覧

状態	値	方法
上位 Grp 過去 10 日 SRt	H,M,L	件数均等 3 分割
下位 Grp 過去 10 日 SRt	H,M,L	件数均等 3 分割
日経過去 10 日移動平均	P,N	0 より小 or 0 以上
上位 Grp センチメント	基底関数 7	範囲均等 8 分割
下位 Grp センチメント	基底関数 7	範囲均等 8 分割
ポジション	0,1	保有有り・無し

表 1 は利用した状態の一覧を示している。「上位 Grp 過去 10 日 SRt」、「下位 Grp 過去 10 日 SRt」、そして「日経過去 10 日移動平均」は、既に第 2 節で説明した通りである。「上位・下位 Grp センチメント」は連続値で、ガウス関数の中心を定めるために、センチメント指数のレンジを 8 分割し、9 つの境界の中で、最小と最大を除く 7 つの値を中心値として利用した。ポジションは、銘柄の保有状況を示すための値であり、保有中であれば 1、それでなければ 0 の値をとる。センチメント以外は、離散値であり、それらを組み合わせると合計 36 種類の離散状態を扱う。

次に行動については、株が未保有の状態（ポジション=0）では、「購入」「売却」「様子見」の 3 つの行動から選択され、保有状態では、「様子見」「精算」の 2 つの行動から選択される。強化学習を実施するに際して、1 つのエピソードは取引開始日 (様子見が最初の場合もある) から精算日までの期間を表し、これを 1 つのサンプルとして学習する。実験にはエピソードを日々生成する方法を用いたので、1 年間に約 250 エピソードのサンプルが生成され、10 年間で約 2500 エピソードの学習を行う。実際に取引する銘柄は、上位グループに含まれる銘柄の中で、MACD の高い上位 3〜6 %に位置づけられる銘柄である。この方法で選択される銘柄は、平均 15 銘柄ほどである。

図 1 はシミュレーションの結果を示している。シミュレーションで利用したパラメータは、ϵ=0.05、α=0.1、γ=0.9 である。図の横軸は経過日数、縦軸は累積報酬で、各線は λ の値、状態、そしてアルゴリズムを表している。λ は 4 種類 (0.1,0.5,0.7,0.9) の値、状態は「d」が離散で「dn」は離散と連続の両方を状態に利用して

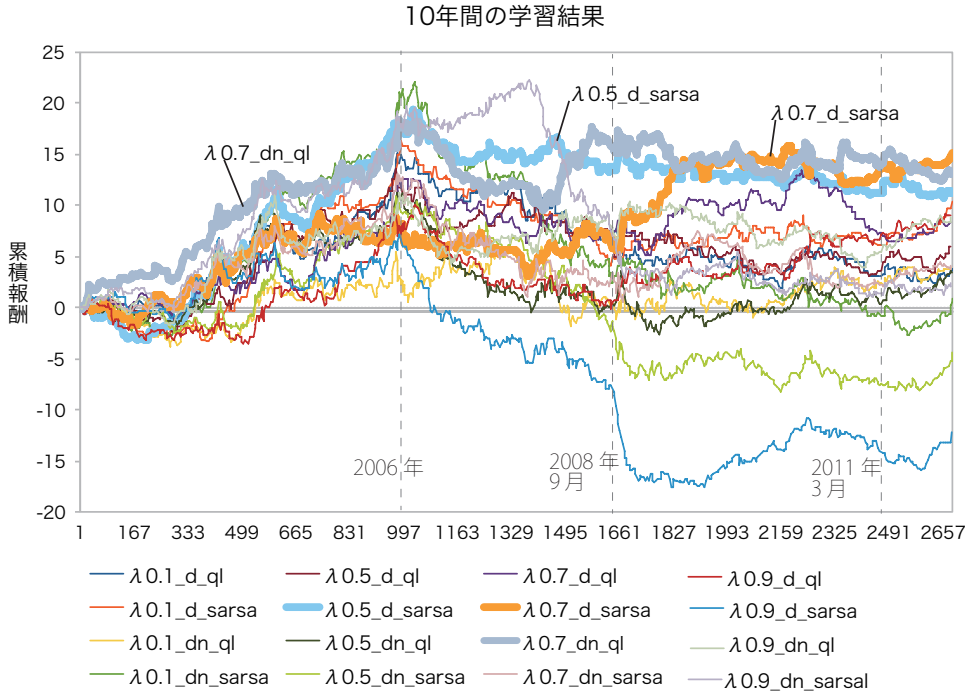

図 1: 10 年シミュレーションの結果 (累積報酬)

いる。「d」の場合はセンチメント指数は利用しておらず、「dn」の場合にのみ利用している。アルゴリズムは ql が Q(λ) でもう 1 つが sarsa(λ) である。全体的に累積報酬は 1000 日程度 (2006 年) までは上昇傾向であるが、その後にピークを迎え、そこからは減衰傾向になっている。これは、2006 年の村上ファンド事件やライブドア事件、そしてアルゴリズム取引の開始時期と時を同じくしており、急激な環境の変化が生じたと考えられる。そして、その変化に学習が追いついていないことが原因で報酬が減衰傾向になっていると考えられる。また、2008 年 9 月のリーマン・ショックの影響で、累積報酬が急激に落ち込んでいるパラメータもある。このような急激な環境の変化に対応するためには、動的に α を変化させる方法が効果的かもしれない。2011 年の震災時はそれほど大きな影響を報酬には与えていない。

一方で全体とは異なる動きをしたパラメータもいくつかあり、太い線で示したパラメータ (λ0.5_d_sarsa) は、2006 年までの学習結果がよくその後も減少の少ない安定した学習が行えている。また、λ0.7_d_sarsa の線は、リーマン・ショックまでは、比較的緩やかに上昇し、その後さらなる上昇が確認できており、最終的に一番累積報酬が高くなっているパラメータである。線の名前に「dn」を含んでいるものは、離散と連続状態を扱った方法であり、λ0.7_dn_ql は、2008 年リーマン・ショックの影響は受けておらず、連続値を含んだパラメータの中では最も累積報酬が高くなっている。これは、センチメント指数を利用している効果の現れであると考えられる。

表 2: パラメーター別平均累積報酬と標準偏差

パラメータ	平均	標準偏差
λ0.1	4.787	2.373
λ0.5	4.502	4.781
λ0.7	**7.847**	3.199
λ0.9	2.916	5.886
離散 (d)	**5.116**	4.910
離散連続 (dn)	4.910	3.986
SARSA(λ)	4.650	**5.189**
Q(λ)	**5.375**	3.390

表 2 はパラメータ別に累積報酬の平均と標準偏差を計算したものである。λ は 0.7 をピークにそれより大きくても小さくても、累積報酬の値は低くなっており、また標準偏差も比較的小さいことから、連続・離散状態やアルゴリズムに関わらず、λ は 0.7 が良い値になっている。離散と連続状態では、平均値にあまり大きな違いはないが、離散状態だけを利用した方が平均値は高く、センチメント指数の効果はそれほど大きくないと考えられる。最後にアルゴリズムについては、Q(λ) の方が平均値は高いが、SARSA(λ) は標準偏差が大きくなっており、パラメータの変換に敏感に反応している。

適格度トレースである λ は、どれだけ過去に遡って価値を更新するかを決定するパラメータである。$\lambda = 1$ の場合は、エピソード終了時に報酬が得られるまでの軌跡となる状態行動すべてに対しての価値が更新される。したがって、λ が 1 に近づくと 1 回の報酬で得ら

れた価値が広く伝播するため、λ0.9_d_sarsa は、リーマン・ショックのような負に大きな影響が急速に伝播することで、累積報酬が 2008 年 9 月頃から急に減少していると考えられる。

図 2: 累積超過収益率

図 2 は、強化学習による効果を確かめるために累積報酬の高かった、λ0.7_d_sarsa、λ0.5_d_sarsa と λ0.7_dn_ql を対象に、ポートフォリオに組み込んだ場合の超過収益率を示している。最初の 2 つのパラメータは、sarsa であり、2006 年前後の環境の変化にうまく対応出来ているが、リーマン・ショックの際には λ0.5_d_sarsa の収益率が減少している。一方 λ0.7_d_sarsa は、リーマン・ショックの後に収益率を伸ばしており、短期的なトレンドの変化をうまく捉えながら学習できている。連続値を含んだ λ0.7_dn_ql は最初に多く収益を上げ、その後は、環境の変化が生じても大きな減少はなく、緩やかな上昇が確認できている。

5 おわりに

本研究では、強化学習を利用して、短期的なトレンドを捉えながら株を売買する戦略に対して、10 年間の株価データを用いて学習と評価を行った。10 年間に起こった急激な環境の変化にもに対応可能な学習パラメータを発見することができ、10 年間で 15%の超過収益率を達成できている。一方で、連続状態を含む結果に関しては、ガウス関数の中心をどのように設定するかによって、パフォーマンスが異なることが考えられるので、もう少し様々な値を用いて実験を行いながら、連続状態を近似するための方法については、改善が必要である。また、今回は MACD を軸にグループ化を行い、学習を実施したが、異なる軸を対象にした場合も今回発見したパラメータが適用出来るか確認したい。

参考文献

[Fama 92] Fama, E. and K. French (1992). “The cross-section of expected stock returns.” Journal of finance: 427-465.

[Fama 93] Fama, E. and K. French (1993). “Common risk factors in the returns on stocks and bonds.” Journal of financial economics 33(1): 3-56

[Sutton 98] Sutton, R.S. and Barto (1998), 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.

[Okada 12] 岡田克彦，中元政一，東高宏，羽室行信 (2012).「証券アナリストの格下げ記事により価値を失う企業の特徴分析センチメント解析と時系列パターン解析を中心として」, 人工知能学会論文誌，Vol.27, No.6，pp. 355-364.

[Takadama 03] 高玉 圭樹,「マルチエージェント学習 ―相互作用の謎に迫る」, コロナ社, 2003.