3C1-06

強化学習を用いた株式取引エージェントの評価

A Reinformcent Learning Agent for Stock Trading: An Evaluation

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 describes an evaluation of a reinforcement learning agent for stock-trading, which we have proposed. We used stock prices for 23 years (from 1983 to 2005) to learn and to evaluate and we show the experimental results. We then show the results of the second Kaburobo contest held from December 2005 to March 2006.

1. はじめに

強化学習 [3] は，試行錯誤に基づいた機械学習の手法であり，自律型エージェントの行動学習に適している．筆者らは，これまでに，強化学習アルゴリズムの**オンライン型** profit sharing（OnOS）[1, 2] を**株式取引エージェント**の行動学習に応用する方法を提案した [6]．本論文では，1983 年から 2005 年まで 23 年間のデータを用いて，[6] で提案した強化学習を用いた株式取引エージェントの評価を行う．

強化学習を用いた学習には，非常に多くの試行錯誤が必要である．しかしながら，[6] では，トレーニング・データの期間がわずか 1ヶ月であり，強化学習を行うにはデータの数が少なすぎる．そこで，本論文では，Yahoo!ファイナンス [4] の株価時系列データから取得可能な 1983 年 1 月から 2005 年 12 月まで 23 年分のすべてのデータを取得し，これをトレーニング・データとテスト・データに分割して用いた．

カブロボ・プログラミング・コンテスト（略称：**カブロボ・コンテスト**）[5] 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．2005 年 1 月から 2 月に開催された第 1 回のコンテストに続き，2005 年 12 月から 2006 年 3 月にかけて第 2 回のコンテストが開催され，多数のチームが参加した．

コンテストの参加者は，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，日経 225 銘柄の中から主催者が選んだ 100 銘柄[*1]を対象として**仮想証券会社**と取引し，所持金 500 万円[*2]からの運用成績を競う．

本論文では，まず，強化学習を用いた株式取引エージェントについて説明する．続いて，評価実験の結果を示すとともに，第 2 回カブロボ・コンテストに参加した結果を紹介し，考察する．

2. 強化学習を用いた株式取引エージェント

本研究で構築したロボットの概要を図 1 に示す．ロボットは，**強化学習エージェント**に行動を決定させ，**投資エージェント**を通して仮想証券会社との取引を行う．

本研究では，取引の対象を同業種の 2 銘柄に絞り，2 銘柄の価格比に着目した**レシオ取引**を行う．2 銘柄だけ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手法の一つである．1 組だけのレシオ取引，かつ，一定量の成り行き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「購入」（主取引銘柄を購入し，副取引銘柄を売却する）と「売却」（主取引銘柄を売却し，副取引銘柄を購入する）だけにできる．[6] ではこれに「様子見」を加えて行動を 3 種類にしていたが，学習を簡単にし，かつ，積極的に売買を行うために，本論文では「買い」と「売り」の 2 種類だけとした．

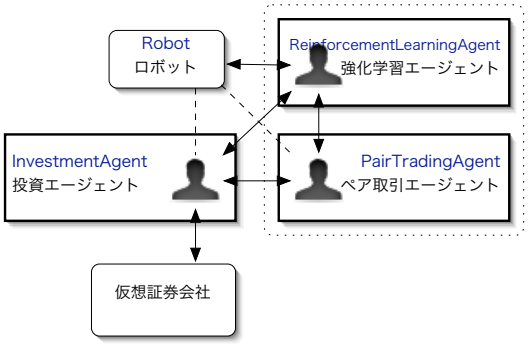

図 1: 本研究で実装したロボットと各種エージェントの関係．点線の枠内は本研究で構築したエージェント．

強化学習における**状態**は，2 銘柄の価格比を補正した**レシオ**と**割引ゴールデン・クロス**を用いて表現している [6]．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 [3] によって表現している．

強化学習における**報酬**には，総資産額の前日比を用いる．総資産額の前日比を，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．

強化学習アルゴリズムには，**オンライン型** profit sharing（OnPS）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，カブロボに適した強化学習アルゴリズムである．

連絡先: 松井 藤五郎（matsui@ia.noda.tus.ac.jp, とうごろう.jp）

∗1　第 1 回コンテストでは 40 銘柄だった．
∗2　第 1 回コンテストでは 1,000 万円だった．

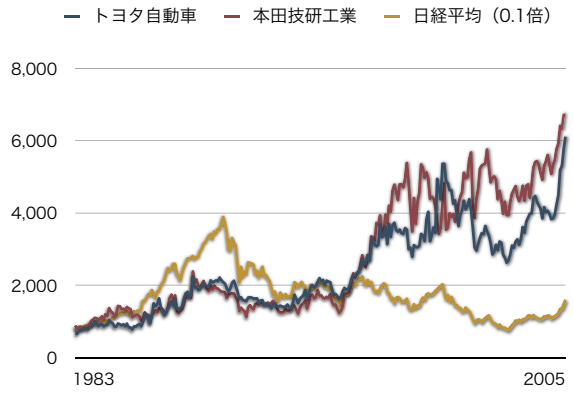

図 2: 1983 年 1 月から 2005 年 12 月までの調整後終値と日経平均株価の推移．日経平均株価は 0.1 倍して表示している．

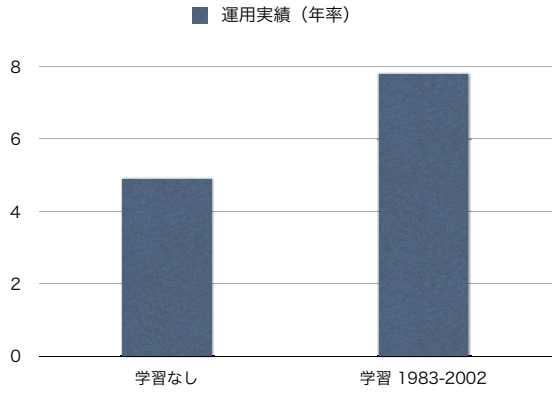

図 3: 2003 年 1 月から 2005 年 12 月までの 3 年間で評価したときの運用実績．

3. 評価実験

3.1 実験方法

本論文では，**トヨタ自動車**（以下，トヨタ）を主取引銘柄，**本田技研工業**（以下，ホンダ）を副取引銘柄とした[*3]．Yahoo! ファイナンス [4] の株価時系列データから取得可能な 1983 年 1 月から 2005 年 12 月まで 23 年分のすべてのデータを取得し，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．トヨタとホンダの調整後株価の推移を図 2 に示す．また，注文量は両銘柄の最小売買単位である 100 株とした．

特徴量 $\vec{\phi}$ を決める動径基底関数は，タイル・コーディングと同様に格子状に配置した．ここでは，幅が 0.25 である 9×9 の格子を 10 枚用意し，原点を基点として幅内でランダムにずらして重ねて配置した．強化学習のパラメータは，[6] と同様に，温度パラメータを $\tau = 0.1$，ステップ・サイズ・パラメータを $\alpha = 0.1$，割引率パラメータを $\gamma = 0.9$ とした．

行動を乱数によって選択するため，101 回評価を行い，その平均を求めた．強化学習を行う場合には，乱数の種を変えて学習と（101 回の）評価を 30 回繰り返し，すべての平均を求めた．

3.2 実験１：学習しない場合との比較

まず，23 年間のデータを 1983 年から 2002 年まで 20 年分のトレーニング・データと 2003 年から 2005 年まで 3 年分のテスト・データに分け，学習した場合と学習しない場合の運用実績を調べた．

結果を図 3 に示す．学習していないエージェントが運用した場合，3 年後の平均総資産額は 5,771,668.3 円で，年率は 4.9 %だった．OnPS を用いて学習を行ったエージェントが運用した場合，3 年後の平均総資産額は 6,268,261.4 円で，年率は 7.8 %だった．

3.3 実験２：テスト・データと同じ傾向を示すトレーニング・データからの学習

一般に，トレーニング・データとテスト・データは同じ性質を持つことが望ましい．内閣府が発表している月例経済報告 [7] によると，日本経済は 2002 年 3 月から本原稿執筆時の 2006 年 4 月まで 51 ヶ月連続で景気が拡大している．しかし，1980 年代後半のバブル景気とその崩壊により株価は大きく変動している．したがって，実験 1 のトレーニング・データ（1983 年から 2002 年）とテスト・データ（2003 年から 2005 年）の傾向はかなり異なっていると考えられる．

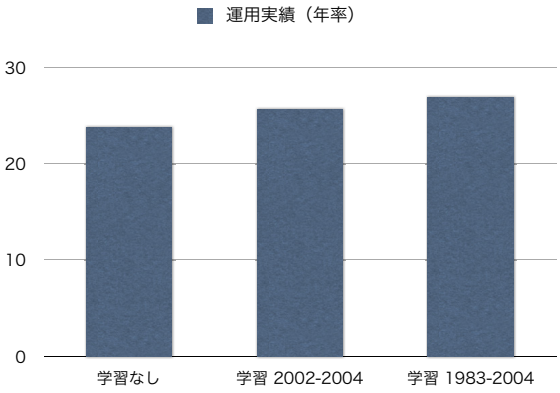

図 4: 2005 年 1 月から 12 月までの 1 年間で評価したときの運用実績．

そこで，テスト・データを 2005 年の 1 年間とし，景気拡大局面の 2002 年から 2004 年までの 3 年間をトレーニング・データとした場合と，バブル景気とその崩壊を含む 1983 年から 2004 年までの 22 年間をトレーニング・データとした場合の運用実績を調べた．

結果を図 4 に示す．学習していないエージェントが運用した場合，1 年後の平均総資産額は 6,187,811.9 円で，年率は 23.8 %だった．2002 年から 2004 年までの 3 年間学習したエージェントが運用した場合，1 年後の平均総資産額は 6,280,598.0 円で，年率は 25.6 %だった．1983 年から 2004 年までの 22 年間学習したエージェントが運用した場合は，1 年後の平均総資産額は 6,344,560.4 円で，年率 26.9 %だった．

4. 第２回カブロボ・コンテストの結果

2005 年 12 月 19 日から 2006 年 3 月 31 日にかけて行われた**第 2 回カブロボ・コンテスト**に，本ロボットで参加した．コンテストに参加したロボットは，1983 年から 2005 年までのすべてのデータを用いて学習させたものである．エントリーの際の

*3 [6] では，イオンを主取引銘柄，セブン・イレブン・ジャパンを副取引銘柄としていたが，セブン・イレブン・ジャパンが持ち株会社に移行したため，過去の株価データとの整合性がなくなった．

表 1: 第 2 回カブロボ・コンテストの結果．T はトヨタ自動車，H は本田技研工業を表す．

日付	T 終値	H 終値	行動	T 株数	H 株数	総資産額
2 日	6,210	6,850	—	0	0	5,000,000
3 日	6,120	6,730	売却	−100	100	4,995,000
6 日	6,190	6,750	売却	−200	200	4,981,000
7 日	6,210	6,900	購入	−100	100	5,005,000
8 日	6,230	6,910	売却	−200	200	5,003,000
9 日	6,270	6,960	購入	−100	100	4,998,000
10 日	6,290	7,050	売却	−200	200	5,009,000
13 日	6,340	7,150	売却	−300	300	5,021,000
14 日	6,420	7,130	売却	−300	400	5,017,000
15 日	6,340	7,180	売却	−300	400	5,031,000
16 日	6,300	7,080	購入	−200	300	5,012,000
17 日	6,290	7,080	売却	−300	400	5,011,000
20 日	6,400	7,150	売却	−300	400	5,006,000
22 日	6,340	7,120	売却	−300	400	5,012,000
23 日	6,320	7,180	売却	−300	400	5,042,000
24 日	6,350	7,220	購入	−200	300	5,047,000
27 日	6,390	7,330	購入	−100	200	5,072,000
28 日	6,360	7,290	売却	−200	300	5,072,000
29 日	6,390	7,370	購入	−100	200	5,089,000
30 日	6,470	7,400	売却	−200	300	5,080,000
31 日	6,430	7,290	売却	−300	400	5,049,000

参加者ニックネームは**とうごろう**，ロボット名は**リッチー**[*4]である．本ロボットの運用実績の詳細をカブロボのウェブ・サイト [5] で見ることができる．

3 月 2 日にロボットを登録し，翌 3 月 3 日から 3 月 31 日まで[*5]株式取引を行った．コンテスト中のロボットの振る舞いとその結果を表 1 に示す．「売却」行動は主取引銘柄であるトヨタを売却し副取引銘柄であるホンダを購入し，「購入」はその逆を行う．この行動は，前日の大引け後に選択し，注文したものであり，その日の始値で取引が行われる．

コンテスト終了時の総資産額は 5,049,000 円となり，単純計算年率利益率は 16.0 %であった．

5. 考察

実験 1 の結果から，本研究で構築した強化学習を用いた株式取引エージェントは，長期間の運用においても学習の効果があると考えられる．また，学習したエージェントをトレーニング・データと異なる期間のデータに対して適用しても学習の効果が得られることがわかる．

実験 2 の結果は，当初の予想とは異なり，テスト・データと同じ傾向を持つトレーニング・データから学習するよりも，バブル景気とその崩壊を含む全トレーニング・データから学習した方が運用実績が高かった．これは，3 年分ではトレーニング・データの数が少なすぎるからだと考えられる．強化学習には数多くの試行錯誤が必要であることから，テスト・データとは異なる傾向を持つトレーニング・データであっても，より多くのデータを用いて学習するほうがいいと考えられる．また，図 2 からわかるように，本論文で対象としたトヨタ株とホンダ株は，バブル崩壊の影響をそれほど受けず，全体的に上昇傾向だったことも影響していると考えられる．

∗4　強化学習の第一人者 Richard S. Sutton のニックネームと金持ちを表す Rich から命名した．

∗5　4 月 1 日と 2 日は土日だったため，3 月 3 日から 4 月 2 日までの 1 ヶ月間とみなすことができる．

第 2 回カブロボ・コンテストにおいて，3 月 2 日の終値の価格比は 1.103，3 月 31 日の終値の価格比は 1.134 であった．ペア取引の観点から見ると，1 ヶ月後に価格比が開くため，運用開始時点では，主取引株であるトヨタを売却し，副取引株であるホンダを購入しておくことが望ましい．表 1 を見ると，本ロボットはトヨタを売却しホンダを購入する傾向にあったことがわかる．これは，ペア取引の理論に一致した振る舞いであり，学習の成果であると考えられる．

また，表 1 からは，20 日から 23 日にかけてのように，運転可能金額の不足により注文が成約できていない事態が生じていることがわかる．このような事態は，学習中にも同様に生じていると考えられる．本論文では，状態をレシオと割引ゴールデン・クロスだけで表現し，状態が運転可能金額や持ち株数を含んでいないことから，このような事態が生じると行動が次状態や報酬に適切に反映されない．これは，今後の課題である．

6. おわりに

本論文では，トヨタ自動車株と本田技研工業株を対象にしたペア取引の戦略を OnPS アルゴリズムで学習する株式取引エージェントに対し，1983 年から 2005 年までの 23 年間の株価データを用いて学習と評価を行った．

1983 年から 2002 年までの 20 年間をトレーニング・データ，2003 年から 2004 年までの 3 年間をテスト・データに用いた実験では，学習を行わないエージェントに比べていい運用実績を示した．トレーニング・データをテスト・データ（2005 年）と同じ傾向を示す期間（2002 年から 2004 年まで）だけに限定した実験では，当初の予想に反して，トレーニング・データを限定しないほうがいい運用実績を示した．強化学習を用いた学習には多くの試行錯誤が必要となることから，データの質よりも量が必要となると考えられる．

今後は，学習した知識（パラメータ）が他の銘柄の組み合わせに対しても適用できるかを調べたい．

参考文献

[1] 松井藤五郎, 犬塚信博, 世木博久．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（第 16 回）論文集, pp. 2D3–03, 2002.

[2] Tohgoroh Matsui, Nobuhiro Inuzuka, and Hirohisa Seki. Online profit sharing works efficiently. In V. Palate, R. J. Howlett, and L. Jain, editors, *Proceedings of the 7th International Conference on Knowledge-Based Intelligent Information & Engineering Systems*, Vol. 2773 of *Lecture Notes in Artificial Intelligence*, pp. 317–324, 2003.

[3] Richard S. Sutton and Andrew 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.

[4] Yahoo!ファイナンス, 2006. http://quote.yahoo.co.jp/.

[5] カブロボ・コンテスト, 2004–2006. http://kaburobo.jp/.

[6] 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, pp. 1D4–01, 2005.

[7] 内閣府. 月例経済報告, 2002–2006.