2007－MPS－63 (15)
2007／3／4

強化学習を用いた株式取引エージェントの構築

児玉 吉晃　　謝 孟春
和歌山工業高等専門学校

近年，カブロボの開催などにより，大規模なファンドの世界では一般的に用いられていたシステムトレードが個人投資家の間でも注目を集めるようになった．そのため自律的にタスクを遂行するエージェントの実現が強く望まれている．本研究では，強化学習を用いて売買ルールを自律学習する株式取引エージェントを構築し，エージェントがより多くの利益を得るように学習方法や状態空間の構築などの仕組みを検討する．構築した株式取引エージェントを用いて評価実験を行い，その有効性を示す．

Construction of the Stock Trading Agent Using Reinforcement Learning

Yoshiaki Kodama　Mengchun Xie
Wakayama National College of Technology

System trading is an investment method to buy and sell according to rules of a buying and selling automatically. Recently, it attracts the attention among individual investors in the world of a large scale fund through KabuRobo. It is important that an agent accomplishes a task autonomously. In this research, we construct a stock trading agent which learns buying and selling rules using reinforcement learning. Then, we examine learning algorithm and construction method of state space for getting more profit. Finally, the effectiveness of stock trading agent is verified by the testing experiments.

1. はじめに

近年，カブロボ（カブロボ・プログラミング・コンテスト）の開催などにより，大規模なファンドの世界では一般的に用いられているシステムトレードが個人投資家の間でも注目を集めるようになった[1]．しかし，システムトレードでは不確実な株式を対象としているため，売買ルールは設計者の経験則によって作成されることがほとんどである．そのため自律的にタスクを遂行するエージェントの実現が強く望まれている．その中で，環境との相互作用を通じて，適応的な学習が可能な強化学習が注目されている[2]．強化学習とは，試行錯誤を通じて最適な行動戦略を獲得する機械学習手法である[3]．

本研究では，強化学習を用いて売買ルールを自律学習する株式取引エージェントを構築し，エージェントがより多くの利益を得るように学習方法や状態空間の構築などの仕組みを検討する．

また，強化学習において，「次元の呪い」を招かずに，多次元のデータを状態入力データとして容易に扱うことができる「SOMを用いた状態入力ベクトルのクラスタリング法」を提案する．最後に，構築した株式取引エージェントを用いて評価実験を行い，その有効性を示す．

2. 強化学習を用いた株式取引エージェントの構成

2.1 株式取引エージェント

本研究で構築した株式取引エージェントの構成を図1に示す．株式取引エージェントは，強化学習を用いて学習し，行動決定を行い，仮想証券会社との取引を行う．

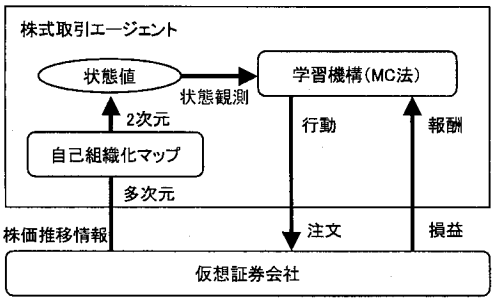

図1 株式取引エージェントの構成

株式取引エージェントは，仮想証券会社から取得した株価推移情報を基に取引対象銘柄の中から1つの銘柄を選択し，全資金分の成り行き注文により終値で取引を行う．その後，翌営業日以降に反対売買を行い，終値で取引を決済する．この1連の流れを1エピソードとして扱う．

2.2 SOM を用いた状態入力ベクトルのクラスタリング法

エージェントの強化学習を行うために，株価推移情報を状態値として表現する必要がある．しかし，株価推移情報は多次元の情報で，そのまま状態入力に用いると「次元の呪い」と呼ばれる状態空間の爆発を招き，強化学習が困難となる．そこで，本研究では，SOM を用いた状態入力ベクトルのクラスタリング法を提案する．

SOM（Self-Organizing Map）は自己組織化マップといい，脳の持つ自己組織化，空間マッピングをモデル化した 2 層構造のニューラルネットワークである．SOM は多次元の入力層と 2 次元の出力層からなり，教師なし競合学習によって入力データの位相関係を保持した特徴マップを形成することで，予備知識なしに非線形なクラスタリングを行うことが可能である．

本手法は，多次元の状態入力ベクトルを SOM においてクラスタリングを行い，低次元ベクトルに圧縮し，状態値に変換する．低次元ベクトルは，SOM の発火ノード位置情報を用いて表現する．

また，SOM を用いて自律的に状態空間を構成することによって，連続値ベクトルを状態入力に用いる際，状態を離散的に表現する強化学習の枠組みにおいて容易に表現することもできる[2].

2.3 エージェントの行動

エージェントの強化学習における行動は，仮想証券会社への注文とする．注文は積極的に売買を行うために「購入」と「空売り」の 2 種類とする．「空売り」とは，株式を保有していないのに株式を売る取引であり，売却した株価よりも安い株価で買い戻すことにより，株価が下落していく局面でも利益を得られる手法である．この「空売り」を行うことにより株式取引エージェントは上げ相場でも下げ相場でも 50%の確率で利益を上げることができ，相場の影響を受けずに利益を上げられる．「空売り」は信用取引に当たるため，通常，資金の 3 倍程度の取引が可能である．しかし，本研究では「購入」と「空売り」の資産に与える影響に違いがでないように資金の 1 倍までの取引に制限する．

2.4 強化学習の状態と報酬

強化学習における状態は，SOM を用いた状態入力ベクトルのクラスタリング法によって，多次元の株価推移情報を 2 次元に圧縮して表現している．株価推移情報として過去数日間の株価と出来高を SOM に入力し，2 次元の発火ノード位置情報で状態を表現する．SOM は変化する相場環境に追従するため，オンラインで学習を行う．

強化学習における報酬は，総資産額の前日比を用いる．保有している株式の資産評価は終値を用いて行う．

2.5 強化学習のアルゴリズム

強化学習においては，モンテカルロ法（MC 法）と ε-greedy 法を用いる。MC 法は行動価値推定型の学習を行う環境同定型のアルゴリズムであり，profit sharing などの経験強化型のアルゴリズムと比べて学習に多くの試行錯誤が必要となる．しかし，経験強化型のアルゴリズムでは，「0 に近い確率で発生し多くの報酬を獲得する」といったルールを学習初期に経験すると，その経験に固執し適切な価値に修正できないという欠点がある．このことから，企業の不祥事などによって株価の暴落，乱高下が発生する可能性のある株式取引への適用には環境同定型のアルゴリズムの MC 法が適していると考える．

また，本研究では SOM を用いて状態入力ベクトルのクラスタリングを行っているため，SOM をオンラインで学習させている．このことにより，状態空間の変化が学習の初期以外にも起こりうる．強化学習から見た場合，この状態空間の変化によってエージェントは同じ状態であるにもかかわらず状態値が変化することになる．一般的な MC 法では，行動価値関数を得られた報酬の平均化で決定するために，状態空間の変化に追従することが難しい．そこで，本研究では式(1)を用いて価値関数を更新する．

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left[r - Q(s,a)\right] \quad (1)$$

ここで，Q は行動価値関数，α は学習率，r は報酬である．

3. 評価実験と考察

3.1 実験内容

評価実験では，株価の推移と売買の結果を比較するためにトヨタ自動車（7203.T）1 銘柄を対象銘柄とした．トレーニング期間を 2000 年からの 5 年間，テスト期間を 2005 の 1 年間とした．株式取引エージェントにトレーニング期間の中から無作為に選んだ営業日においての取引を 10 万エピソ

ード学習させ，テスト期間においての取引エージェントの性能を評価する．

株式取引エージェントの性能の評価にはシステムトレードにおいて一般的な性能指標である年率とシャープレシオを用いた．年率とは1年間の利回り，シャープレシオとはリスク（月次毎の資産変動の標準偏差）に対する利益率の割合である．初期資金は5,000万円とする．SOMのマップサイズは，ノード数がトレーニング期間内の営業日数の約半分となるように25×25とし，状態数を625とした．株価推移情報は過去2日間のものを使用した．

株式取引エージェントは行動確率，SOMの各ノードの初期荷重ベクトル，学習に用いる営業日の選択において乱数を用いるため，20回繰り返し，平均を求めた．

株式データはYahoo! ファイナンスから取得した株価時系列データの調整後終値を用いた[5]．調整後終値とは，株価が株式分割による影響を受けないように，分割実施前の終値を分割後の値に調整した終値のことである．実験に用いた期間の株価推移を図2に示す．トレーニング期間は，2002年までのバブル崩壊後の平成不況と呼ばれる時期と，その後の景気回復の時期であり，様々な相場傾向を含んだものとなっている．テスト期間での値動きは，前半に緩やかに下落し，その後大きく上昇している．このため，テスト期間における運用結果より，様々な相場傾向における評価が可能である．

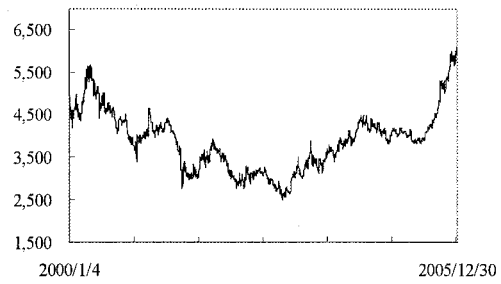

図2 実験に用いた期間における株価推移

3.2 実験1：学習の有効性

本研究で構築した株式取引エージェントを用いて，学習の有効性を示すために，テスト期間での運用結果（年率とシャープレシオ）の平均値を図3に示す．比較するために，学習なしの結果も併記している．年率は，学習ありでは29.50%に対して，学習なしでは−4.85%であった．シャープレシオは，学習ありの場合には5.98に対して，学習なしの場合は−1.11であった．

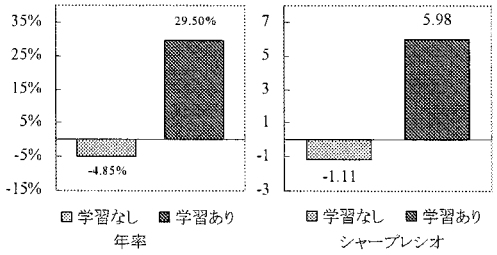

図3 学習の有無による運用結果比較

テスト期間におけるエージェントの資産推移の平均を図4に示す．この1年間の運用後の総資産額の平均は，学習ありでは6,474.88万円に対して，学習なしでは4,759.86万円であった．

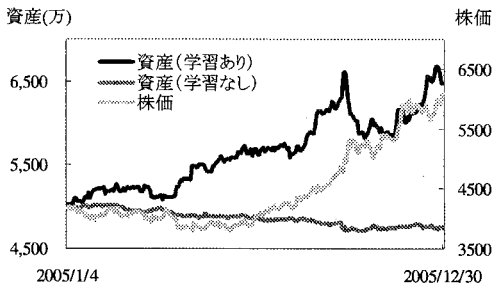

図4 資産推移

また図4より，エージェントは資産運用の結果，株価が下落している局面と上昇している局面の両方においてもほぼ安定して，利益をあげることができることが分かった．しかし，10月後半に大きく損失を出している局面がある．この原因として，株価の推移からでは推測不可能な外因の影響が大きい値動きや，トレーニング期間において経験したことのない未知の状況だったことが考えられる．

3.3 実験2：クラスタリングの有効性

提案したSOMマップを用いた状態入力ベクトルのクラスタリング法の有効性を示すために，運用結果と学習の収束状況から検証実験を行う．

テスト期間での運用結果（年率とシャープレシオ）の平均値を図5に示す．図5にはクラスタリングありとクラスタリングなしの結果をそれぞれに示している．

株価推移情報をそのまま状態入力とするクラスタリングなしの場合では，トレーニング期間から必要な状態数を計算すると約 900 億となった．この状態数では状態空間が爆発しており，行動価値関数を保持するメモリ空間を確保することが困難である．そこで，クラスタリングなしは，各次元に格子数が 50 となるように等間隔のグリッドを配置し，状態数を 625 万に圧縮して表現している．

図 5 よりクラスタリングありの場合は，クラスタリングなしと比べて状態数が 1 万分の 1 であるにも関わらず，非常に良い運用結果であった．これは，SOM を用いたクラスタリングにより，状態が圧縮される際に効率的に一般化が行われているためと考えられる．予測問題においては，常にトレーニング期間とまったく同じ状態が起こるとは限らず，この際に一般化が重要となる．

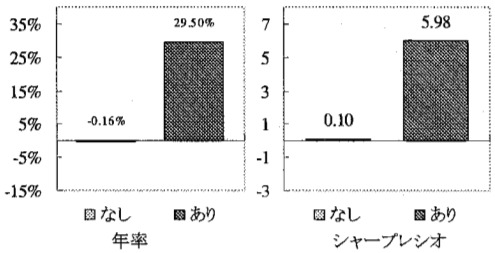

図 5 クラスタリングの有無による運用結果比較

1000エピソードごとの運用結果を示した学習の収束状況の平均を図 6 に示す．

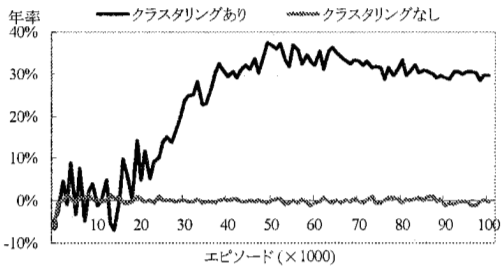

図 6 学習の収束状況

図 6 より，クラスタリングありの場合では学習の収束が遅いことが分かった．クラスタリングなしでは 1000 エピソードまでの間に学習が収束しているが，クラスタリングありでは学習の収束まで約 8 万エピソードを必要とした．これは，クラスタリングありの場合では強化学習の学習だけでなく，SOM の学習もオンラインで行っているため，SOM の収束に多くのエピソードが必要であることが原因だと考えられる．

4. まとめと今後の課題

4.1 まとめ

本研究は，強化学習を用いて株式取引エージェントを構築した．評価実験を行い，構築したエージェントが安定的に利益を上げられることを確認した．また，本研究で提案した「SOM を用いた状態入力ベクトルのクラスタリング法」により，状態空間の爆発を招くことなく，多次元のデータを状態入力データとして容易に扱うことができるだけでなく，クラスタリングによる効率的な一般化によりエージェントの性能を飛躍的に向上させることが可能であることが確認できた．

4.2 今後の課題

現段階では，エージェントはテクニカル分析的な情報のみである株価推移情報を基に売買を行っているが， SOM によるクラスタリングを行っていることから，ファンダメンタル的な情報やその他の様々な情報を扱うことが可能であり，エージェントの性能をより高められると考えられる．今後，この点について検証することが課題の 1 つである．

また，2007 年春に行われる第 2 回スーパーカブロボコンテストで検証を行う予定である．そのため，対象銘柄を複数に拡張した場合や異なる期間での評価実験をする必要がある．

参考文献

[1] http://www.kaburobo.jp/

[2] 岩崎秀樹,末田直道, “強化学習における自己組織化マップを用いた状態空間の自律的構成法”,The 19th Annual Conference of the Japanese Society for Artificial Intelligence (2005)

[3]三上貞芳,皆川雅章共訳,「強化学習」,森北出版(2000)

[4] 特高平蔵,藤村喜久朗,山川武烈,”自己組織化マップにおける応用事例集”,海文堂出版(2002)

[5] http://quote.yahoo.co.jp/

[6] 児玉吉晃,謝孟春, “株式売買エージェントへの強化学習の応用”,第 12 回高専シンポジウム講演要旨集,p148(2007)