情報処理学会　研究報告
Technical Report

2006－MPS－59　(10)
2006／5／25

AHPを導入したProfit Sharingエージェントによる株式売買に関する研究

吉本昌弘　藤森成一　佐々木将士　東海大学

本研究は、株式売買において、AHPを導入した強化学習エージェントが、Profit Sharingを用い、その行動に対する評価から、行動ルールの重みを更新し、ルールの重みに従って行動を決定し売買を行う。Profit Sharingを用いることで、目標達成時にのみスタートからゴールまでに選択したルールを一括して更新することができる。強化学習だけでは売買効率が悪いため、AHPをエージェントに導入することで、選択肢に対する人間の主観的評価を数値化し、強化学習の効率を向上させる。AHPからの重みを学習の進行につれて減衰させる合成比減衰法により、知識導入による強化学習への悪影響を抑制でき、有用かつ効率的な売買を可能にする。

Research on stocks buying and selling by Profit Sharing agent who introduces AHP

Masahiro Yoshimoto　Seiichi Fujimori　Masashi Sasaki　Tokai University

When stocks are bought and sold, the reinforced learning agent who introduces AHP uses Profit Sharing, the weight of the action rule is updated from the evaluation to the action, the action is decided according to the weight of the rule. The rule selected from the start to the goal only at accomplishment of the goal can be updated in bulk by the use of Profit Sharing. Man's stocks trading subjective evaluation to choices is expressed numerically by introducing AHP into the agent because the buying and selling efficiency is insufficient, and the reinforced learning is improved by introducing AHP, because the adverse effect on reinforced learning by the knowledge introduction can be controlled by the synthetic ratio attenuation method that attenuates weight from AHP as the learning progresses, a useful, efficient buying and selling is enabled.

1．はじめに

人工知能ではニューラルネットワークの研究が長年さかんに進められてきたが、近年はそれに代わるものとして強化学習[1][2][4]が注目されている。ニューラルネットワークでは、教師データに依存して学習を進めていくものが多いが、強化学習では教師データを必要とせず、自らが試行錯誤を繰り返し、その環境に最適な行動を経験により構築していく学習法である。これにより、エージェントは、予測不可能な環境に対しても適応していくことが期待できる。本研究は、人工知能エージェントに株式売買を行わせ、得られた経験からエージェント自身が自己を改善させ、株式売買利益を出すことを目指す。

2．強化学習

強化学習エージェントは、図2．1に示すように、未知の環境において、自らが置かれている状態を観察し、行動する。そして、その結果を評価する。これが強化学習エージェントの学習法である。このエージェントは目的を達成したときにのみ報酬を得ることができ、学習を繰り返し、経験を得ることによって目標達成までの最善の方法を自ら構築していくことが可能である。エージェントは状態認識器で現状態を認識し、学習器に保持している状態と行動がセットになった、状態－行動ルールの中から候補を選び出し、行動選択器でそれらの候補の中から一つを選んで行動を起こす。その行動に対する評価を報酬として受け取り、選択したルールの重みを更新する。従って、エージェントはルールの重みが大きくなる行動を選択することになる。

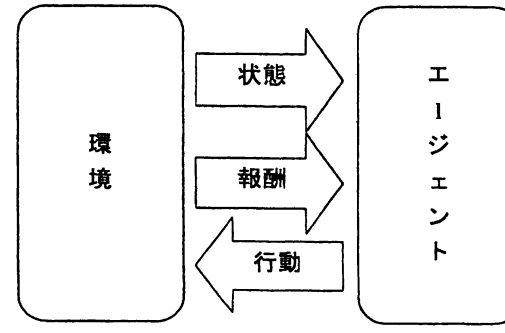

図2．1　強化学習モデル

3．行動選択法

ソフトマックス手法(ルーレット選択 Roulette)[2]は、重みが軽いルールはあまり選択されず、逆に重みが重いルールは頻繁に選択されるように各ルールをランク付けする方法である。ルーレット選択法は確率的に政策を

自然に実現する枠組みであり、非マルコフ決定過程環境における行動選択として有効な手法である。

グリーディ手法(Greedy)[2]は一番重い重み(価値)を持つルールの行動を決定的に選択する方法である。すなわち、重みが一番重いルールが最も有効である。この手法では、有効なルールしか選択されない。このため、エージェントは早い段階で学習を終了させることができるが、そうでない場合は学習の早さが改善されることはない。

この研究では、学習期間にルーレット選択法を用い、学習期間以降の実際の売買ではグリーディ手法により選択を行うことにより、学習結果にのみ依存した売買をエージェントに行わせる。これにより、学習中に報酬を得られず学習が進まないという状況を確率的に回避し、実際の売買シミュレーションでの学習結果を測ることができる。

4．Profit Sharing

Profit Sharing[1][2][4]は学習分類子システムの学習メカニズムとして研究され、強化学習の学習アルゴリズムとして非マルコフ決定過程環境での有用性が期待されている。Profit Sharingは独立に行われる学習である。このアルゴリズムはスタートからゴールまでのルールの履歴を記憶しておき、ゴール到達時に式(1)、式(2)に従って履歴にあるルールの重みを順番に更新する。

$$Q(a_t,s_t) \leftarrow Q(a_t,s_t)+\beta\left(r(t)-Q(a_t,s_t)\right) \quad \cdots(1)$$

$$r(t)=r(t+1)/episode \quad (t=0,\ldots,episode-1) \quad \cdots(2)$$

$Q(a_t,s_t)$は時刻tで実行したルールの重み、β ($0<\beta \leqq 1$)は割引率、$r(t)$は報酬関数である。現状態から次状態への遷移を1ステップとし、スタートからゴール到達まで、最終状態への到達による強制終了までを1エピソード*episode*と呼ぶ。本研究で作成する状態ではエピソード中に同じ状態に出会う確率が高く、その場合同一のルールを何度も更新してしまう。追跡問題ならば同じ状態に何度も遭遇してしまうことは、ゴール到達への遅延を招く要因として迂回ルールとすることが有効だが、株式売買の場合はそうとは言うことができない。その為、ルールの更新は初回訪問法とし、エピソード中に同一のルールを2度以上利用した場合、そのルールに関しては更新を1回とする。

5．階層化意志決定法

階層化意志決定法（Analytic Hierarchy Process, AHP）[1]は、その利用の度合を調節することにより、強化学習の効率を向上させることが示されている。AHPは人間の主観的評価により選択肢に重みを付けることができる意志決定法である。その重みの大きさから選択肢の優先順位を知ることができ、選択肢の決定に役立てることができる。状態認識器から得られる状態に基づいて、各行動の重み付けをAHP器で行い、行動選択器ではその重みに従ってルーレット選択法により行動選択する。AHP器は状態認識器から与えられる情報より、適切な行動が優先されるように候補となる行動群を重み付けする。エージェントは、AHP器と学習器の重みを合成し、その合成された重みに基づいて行動を選択する。提案法では状態認識器からの状態と実行された行動の対のルールは常に学習器に送られる。よって、AHP器の重みのみが利用される場合であってもエージェントは学習する。

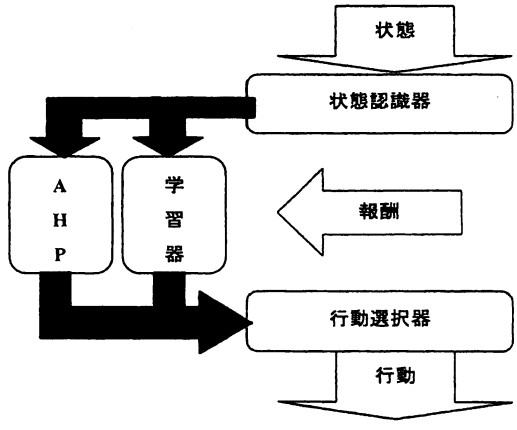

図5．1　AHPを導入したエージェント

6．階層構造

株式売買でのエージェントの目標は、学習期間500週の中で“買いと売り”、または“売りと買い”を一度行い、その銘柄の500週間における平均株価の30%以上を利益として回収することである。ここでは単に「利益を上げること」と表現する。

AHPの階層構造を、「利益を上げること(profit)」、評価基準を「株価テクニカル指標」[3][5]、代替案を「買い(buy)」と「待ち(wait)」、「売り(sell)」と「待ち(wait)」としている。「株価テクニカル指標」の詳細は「株価変化率(price)」、「出来高変化率(volume)」、「勝ち数サイコロジカル(logic1)」、「値動きサイコロジカル(logic2)」、「終

値と短期移動平均の乖離率(premium1)」、「短期移動平均と長期移動平均の乖離率(premium2)」とした。

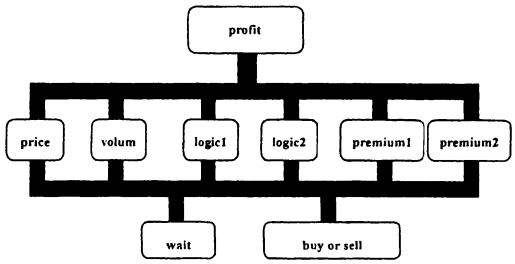

図６．１　株式売買に対するAHPの階層構造

７．AHPの自動更新設定

エージェントはテクニカル指標[3][5]結果から、項目がどの大きさにあるかを状態として認識する。指標は%で算出され、状態の要素として利用することが可能である。エージェントは“high”と“low”の二段階、項目によっては“middle”を加えた三段階に分ける。

表７．１　テクニカル指標要素の詳細

	high	low	middle
price	10%以上	以外	
volume	30%以上	以外	
logic1	75%以上	25%以下	以外
logic2	70%以上	30%以下	以外
premium1	20%以上	-20%以下	以外
premium2	40%以上	-40%以下	以外

８．AHP器と学習器の重み合成

AHP器[1]は学習器と同様に状態観測器から得た状態に基づき各行動の重みを算出して行動選択器に受け渡す。これより、学習初期において良いとは言えない学習器の性能を補うことができると考えられ、AHP器からの重みを用いることで、状態ごとに適切な行動を選択することができる。本研究は、合成比 *rate* を用いてAHP器と学習器から得られる重みを合成する。

$$TWs = rate \cdot AHPWs + (1 - rate) \cdot LMWs \quad (0 \leqq rate \leqq 1) \qquad \cdots(3)$$

$$rate = A \cdot rate \quad (0 \leqq A \leqq 1) \qquad \cdots(4)$$

TWs は最終的に行動選択器に送られる各行動の重みであり、*AHPWs*はAHP器から得られる各行動の重み、*LMWs*は学習器からの各重みである。ただし、合成の計算を行う前に *AHPWs* および *LMWs* ともに行動郡の合計が１になるようにする。式(３)、式(４)に従いAHP器の重みの利用を除々に減衰させることにより、最終的には学習器のみの重みを利用する。

９．実験結果、並びに考察

本研究では株価データとして“始値”、“高値”、“安値”、“終値”、“出来高”を週足で1052週間用意し、テクニカル分析後に1000週間が残るようにした。各銘柄のグラフは終値と 26 週移動平均線[3][5](26Week Moving Average, 26WMA)52週移動平均線(52WMA)を表示。

学習期間は第１週から第500週までとし、残りの第501週から第1000週までを実際の売買シミュレーション用とする。学習エピソードはエージェントが売買を１回行うか、第500週に到達することによる強制終了によってカウントされる。エージェントが“買い”か“売り”を判断した場合、データには終値を用い、“買値”と“売値”は翌週の終値となる。学習期間以降は、学習結果に従い売買を第1000週まで繰り返す。

表９．１　エージェントのパラメータ

学習エピソード	100,000回
最大学習ステップ	500ステップ
学習率	0.1
割引率	0.9
報酬	1.0
ルーレット選択	Profit sharing
グリーディ選択	実際の売買時
目標達成条件	学習期間の株価の平均の三割を利益とし回収

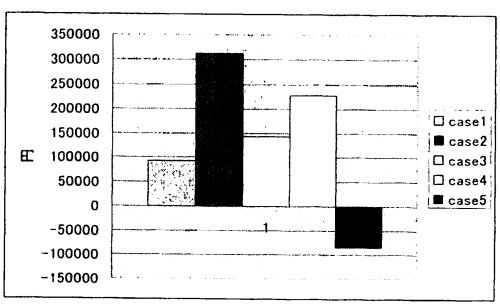

図９．１　全13社における平均利益

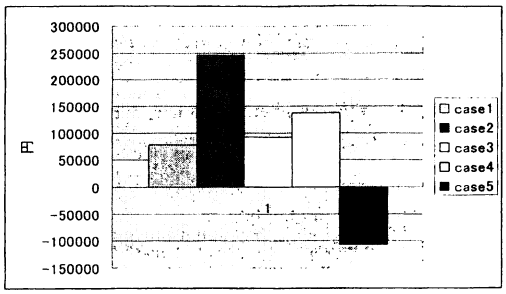

図９．２　全 13 社の実際の売買での利益の見込み

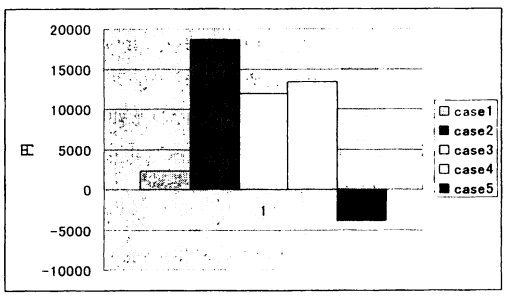

図９．３　全 13 社の売買１回当りの平均利益

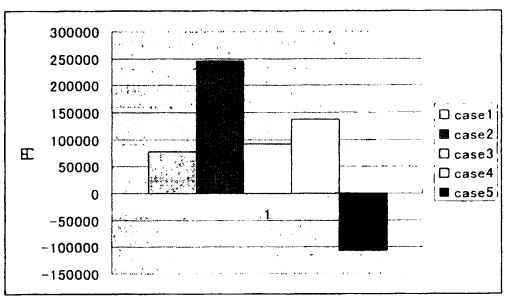

図９．４　売買１回当りの実際の利益見込み

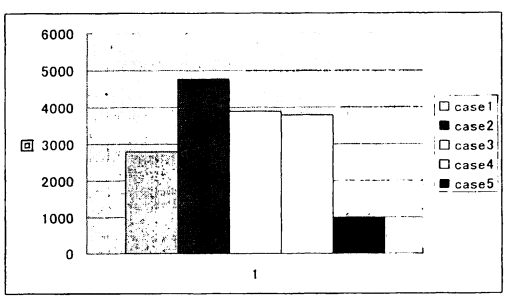

図９．５　学習期間における目標達成回数

case1 は減衰率 0.0 の強化学習のみによる売買結果、case2 から case5 は減衰率がそれぞれ 0.9、0.99、0.999、0.9999 の AHP を導入した強化学習エージェントの売買結果である。case2 から case4 にかけて AHP の導入が Profit Sharing の効率を高めている。しかし case5 では学習は悪化している。case5 については学習期間における AHP への長期依存は Profit Sharing に悪影響を及ぼしてしまうことを意味している。

学習期間における目標達成回数からも、case2 から case4 にかけて AHP の導入が強化学習の効率を高めていることがわかる。case5 では学習効率は悪化している。やはり、長期の AHP 器への依存は危険と言える。

１０．結論

本研究の有効性である、テクニカル指標の AHP 器への適応性はテクニカル指標をパーセンテージの度合で数段階に分けることより、状態数が膨大になり、学習が終わらないことを避けることができ、テクニカル指標は強化学習に対して有効であることがわかった。

また、Profit Sharing は AHP 器との相性の良さがマルチエージェント環境において示されていたが、AHP 器の導入によって学習期間における目標達成回数が増加したことから、シングルエージェントでも AHP と Profit Sharing の相性が良いことを示すことができた。

学習の進行度合は売買結果に大きく関係することから、学習効率を高めることに成功した本研究は、AHP を導入した強化学習エージェントの株式売買への応用において、十分に価値のある成果であると考えられる。

１１．参考文献

[1]片山　謙吾、奥石　尚宏、成久　洋之：「強化学習エージェントへの階層化意志決定法の導入　―追跡問題を例に―」、人工知能学会誌、Vol.19,No. 4, pp.279-291 (2004).

[2]Sutton, R. S. and Barto, A. G.: "Reinforcement Learning": An Intoduction, MIT Press, Cambrige, MA(1998), (邦訳：強化学習、三上　貞芳、皆川　雅章共訳、森北出版、pp.61(2000)).

[3]鳥海不二夫：「株式自動売買ソフトウェア株ロボを作ろう！」秀和システム(2005).

[4]宮崎　和光、木村　元、小林　重信：「Profit Sharing に基づく強化学習の理論と応用」、人工知能学会誌、Vol.14, No.5, pp.800－807(1999).

[5]システック株式会社：「TELECHART－W　Ver.3.5」株式会社システック(2000).