Vol.2022-MPS-141 No.19
2022/12/13

深層強化学習による株式投資戦略における最適な機会損失評価のための適応的な報酬設計

井上修一[†1]　穴田一[†1]

概要：近年，機械学習の発展に伴い深層強化学習を用いた金融取引戦略を構築する研究が精力的に行われている．我々のこれまでの研究では，取引エージェントの「買い」「売り」，及び「様子見」といった行動に対し機会損失を考慮した報酬を与えていた．しかし，すべての期間で十分な利益を上げられているわけではなかった．これは，銘柄固有の価格変動の存在や金融市場の状態によって，最適な機会損失評価のため報酬計算に用いる期間が異なることによるものだと考えられる．そこで本研究では，機会損失評価のための期間が相場の局面によって適応的に変化する報酬を提案し，その有効性を確認する．

キーワード：深層強化学習，機械学習，強化学習，投資戦略

Adaptive Reward Design for Optimal Opportunity Loss Assessment in Stock Investment Strategies Using Deep Reinforcement Learning

SHUICHI INOUE[†1]　HAJIME ANADA[†1]

1. はじめに

近年，機械学習の発展に伴い，機械学習を用いた金融取引に関する研究が盛んに行われている．その中には，深層強化学習を用いて金融取引戦略を構築する研究が存在する．深層強化学習とは，エージェントが試行錯誤を通して設定された環境において報酬を最大化するための行動を学習するもので，株式市場やエージェントの資産などの状態を表す数値を入力とし，資産を増やすための投資行動を出力するモデルを指す．これらの研究では，金融商品の取引量や[1]，報酬の替わりに利益率の複利効果に着目したもの[2]など様々なアプローチがなされている中，我々はこれまでの研究[3]で，取引エージェントの「買い」「売り」，及び「様子見」といった行動に対し機会損失を考慮した報酬による深層強化学習モデルによる金融取引 AI を構築していた．しかし，すべての期間で十分な利益を上げられているわけではない．これは，銘柄固有の価格変動の存在や，金融市場の状態によって最適な機会損失評価のため報酬計算に用いる期間が変化することによるものだと考えられる．そこで本研究では，機会損失評価のための期間が相場の局面によって適応的に変化する報酬を提案し，その有効性を確認する．

2. 深層強化学習

強化学習[4]とはエージェントが試行錯誤を通して目的を達成する方法論である．エージェントは現在の環境の状態を観測後行動し，環境から報酬を受け取る．そしてエージェントは次の状態へ遷移する．これを繰り返し行い，エージェントは報酬を最大化する最適な行動系列を得るための方策を学習する．深層強化学習の 1 つである Deep Q Network（以下，DQN）は，強化学習のアルゴリズムである Q 学習における行動価値関数 Q に対してニューラルネットワークを関数近似器として用いたものである．本研究の学習の際にはエージェントが株式市場から状態S_tを受け取り，確率 ε でランダムな行動を，確率$(1-\varepsilon)$でニューラルネットワークが出力する行動価値$Q(s,t)$が最も高い行動a_tを選択する．そして環境(株式市場)から報酬r_tと次の状態S_{t+1}を受け取る．この一連の経験(S_t, a_t, r_t, S_{t+1})を replay Memory に保存し，ランダムに取り出しミニバッチ学習を行う．

3. 提案手法

3.1 状態変数

本研究ではエージェントが環境から受け取る状態変数として「所持金」「総資産」「評価損益」「株価の前日比」「株価の移動平均」の 5 種類を使用した．株価の前日比は急激な上昇や下落に，株価の移動平均は短期・中期・長期における株価の傾向に対応するために状態変数として設定した．ここで，移動平均とは代表的なテクニカル指標のひとつで，価格のトレンドから相場の方向性を見る手掛かりをつかむために使用されるものである．扱う株価は銘柄の 1 日の終

†1 東京都市大学
Tokyo City University

値ベースとした.

3.2 行動

エージェントは「買い」,「売り」,「何もしない」の 3 種類の行動から 1 日 1 回 1 つの行動を終値で選択する.「買い」では 100 株をその日の終値で購入し,「売り」ではそれまでに保持してきた株をすべて売却する.本研究では買い増しが可能であるほか,株式の現物取引を想定し,空売りなど信用取引は考慮していない.

3.3 報酬

本研究の目的は,運用期間において高利益率の取引戦略を構築することである.このためには日々のエージェントの行動の評価を報酬としてフィードバックしていくことが重要である.我々の研究では,エージェントの「買い」,「売り」,「様子見」のすべての行動に機会損失を考慮した報酬を与えることで,エージェントの選択した行動が最適な行動であるかを評価する.t日目の報酬R_tを以下のように定義する.

$$R_t = \begin{cases} \left(\dfrac{P_{sell} - P_{buy}}{\overline{P_{buy}}}\right) S_{all} + \alpha_t & if\ a_t\ is\ sell \\ \beta_t & if\ a_t\ is\ buy \\ \gamma_t & if\ a_t\ is\ hold \end{cases}$$

ここで,α_tはt日目のエージェントの行動,P_{sell}は売却時株価,$\overline{P_{buy}}$は平均購入時株価,S_{all}は売却株数を表す.α_t, β_t, γ_tはそれぞれ「売り」,「買い」,「様子見」を選んだことにより発生する機会損失を表す項であり,α_t, β_tは以下のように表される.

$$\alpha_t = \frac{\left(P_{sell} - max(L_n(t))\right) + \left(P_{sell} - min(L_n(t))\right)}{P_{sell}}$$

$$\beta_t = \frac{\left(max(L_n(t)) - P_{buy}\right) + \left(min(L_n(t)) - P_{buy}\right)}{P_{buy}}$$

ここで,$L_n(t)$はt日目におけるサイズ$2n$のリストであり,前後n日間の終値が格納されている.$max(L_n(t))$, $min(L_n(t))$はそれぞれリスト$L_n(t)$の最高値と最安値を表す.この情報を参照することで,「もっと安く買えた」「もっと高く売れた」といった機会損失を表現している.このように,過去n日間と未来n日間値動きを考慮して正しい売買行動をエージェントに学習させる.次に,「様子見」の機会損失を表すγ_tを次に示す.

$$\gamma_t = \begin{cases} \dfrac{(max(L_n(t)) - P_{hold})}{P_{hold}} & if\ S_t > 0\ and\ t < t_{max(L_n(t))} \\ \dfrac{(P_{hold} - max(L_n(t)))}{P_{hold}} & if\ S_t > 0\ and\ t \geq t_{max(L_n(t))} \\ \dfrac{(P_{hold} - min(L_n(t)))}{P_{hold}} & if\ S_t = 0\ and\ t < t_{min(L_n(t))} \\ \dfrac{(min(L_n(t)) - P_{hold})}{P_{hold}} & if\ S_t = 0\ and\ t \geq t_{min(L_n(t))} \end{cases}$$

ここで,P_{hold}は様子見時株価を,S_tはt日目の保有株式数を,$t_{max(L_n(t))}$はリスト$L_n(t)$の最高値を付けた日を,$t_{min(L_n(t))}$はリスト$L_n(t)$の最安値を付けた日を表している.エージェントが株式保有状況に応じた機会損失を考慮する「様子見」を学習するために,「様子見」の報酬は株式を保有しているかどうかで場合分けしている.株式保有時の「様子見」の報酬は,高く売るために待つ「高値待ち」を学習するような設計,株式非保有時の「様子見」の報酬は,安く買うために待つ「安値待ち」を学習するような設計としている.

3.4 報酬パラメータ$L_n(t)$

我々のこれまでの研究の投資戦略評価実験より,銘柄固有の価格変動の大きさや相場の状態によって,最適な機会損失評価のための期間が異なることが考えられる.そこで,機会損失評価期間の決定に銘柄の価格変動率を表すボラティリティを用いる.ボラティリティとは変化率の標準偏差を表すもので,過去の株価から計算されるものをヒストリカルボラティリティ(以下,HV)と呼ぶ.HVを以下に示す.

$$HV = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(Q_i - \overline{Q})^2} \times \sqrt{K}$$

$$Q_i = \ln C_i - \ln C_{i-1}$$

ここで,C_iはi日目の終値,Nは標準偏差を計算する期間,Kは扱う銘柄の年間取引日数を表している.このHVを用いて,機会損失を計算する$L_n(t)$におけるパラメータnを次のように表す.

$$n = D(1 - (HV)_{norm})$$

ここで,Dは基準となるリストの長さである.正規化したHVを用いることにより,HVが大きいとき,つまり市場が不安定な時はnが小さくなり,より直近の株価の変動を参考に機会損失を評価し,HVが小さいときには,長期的な視点で機会損失を評価することが可能となる.

4. 評価実験

提案手法の有効性を確認するため,実際の株式データを利用した実験を行う.

発表時に詳細な結果と考察を述べる.

参考文献

[1] 和田 裕貴, 長尾 智晴, “深層強化学習による株式売買戦略の構築”, 情報処理学会第 79 回全国大会, Vol.2017, No.1, pp.345-346 (2017).
[2] 松井藤五郎, 後藤 卓, 和泉 潔, 陳 ユ : ”複利型強化学習の枠組みと応用”, 情報処理学会論文誌, Vol.52, No.12, pp.3300–3308 (2011).
[3] 井上 修一, 穴田一, “深層強化学習による機会損失を考慮した株式投資戦略の構築”, 研究報告数理モデル化と問題解決（MPS）, Vol. 2022-MPS-137, No.4 (2022).
[4] Sutton,R.S., Barto,A.G.:Reinforcement Learning, MIT press, (1998)