A-035

深層強化学習による売買と様子見時の機会損失を考慮した株式投資戦略の構築
Building Stock Investment Strategies with Trading and Stand-off Opportunity Loss Using Deep Reinforcement Learning

井上 修一† 穴田 一†
Shuichi Inoue Hajime Anada

1. はじめに

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に関する研究が盛んにおこなわれている．その中には，2016 年に囲碁のプロを打ち負かした AlphaGo で話題になった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．ここでの深層強化学習は，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利益を最大化するための行動を学習するものである．これらの研究では，金融商品の売買数も含めた最適化[1]に着目したものなど[2]様々なアプローチがなされているが，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．我々のこれまでの研究[3]では，取引エージェントの売買行動に対する報酬に機会損失を考慮していたが，下げ相場の際に買いすぎてしまう結果が見られた．これは，「何もしない」行動について評価するための報酬を設定していなかったからだと考えられる．そこで，本研究では取引エージェントの「買い」，「売り」，「何もしない」のすべての行動に対し，機会損失を考慮した報酬を深層強化学習に組み込み，株式投資において限られた資産の中で利益を最大化するための最適な売買タイミングを学習するモデルの構築を目的とする．

2. 深層強化学習

強化学習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．図 1 に示すように，エージェントは行動を起こし，環境から報酬と次の状態を受け取る．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[3]の 1 つである Deep Q Network（以下，DQN と略す）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を受け取り，確率 ε でランダムな行動を，確率$(1-\epsilon)$でニューラルネットワークが出力する行動価値$Q(s,t)$から最も高い行動a_tを選択する．そして環境(株式市場)から報酬と次の状態を受け取る．この一連の経験を Experience Replay Memory に一定数記録しておき，過去の経験をランダムにサンプリングしミニバッチ学習を行う．

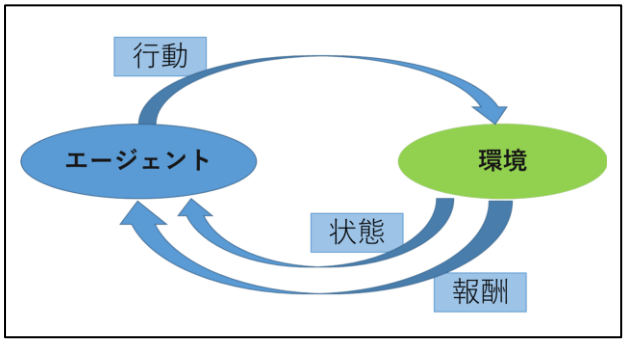

図 1　強化学習の流れ

2.1 提案手法

2.1.1 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以下の 5 種類計 7 つを使用した．

- 所持金（初期保有量:10,000,000￥）
- 総資産
- 評価損益
- 株価の増減率（前日比）
- 株価移動平均（5 日，25 日，75 日）

急激な上昇や下落に対応するため，株価の前日比を状態変数として設定した．また，短期から中長期における株価の傾向をエージェントに与えるために，株価の 5 日，25 日，75 日移動平均を使用した．移動平均とは代表的なテクニカルチャートのひとつで，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．扱う株価は銘柄の 1 日の終値ベースとした．

2.1.2 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する．「売り」はそれまでに保持してきた株をすべて売却する．本研究では株式の現物取引を想定しており，空売りなど信用取引は考慮していない．

2.1.3 報酬

エピソード終了時にまとめて報酬を与えると報酬を受け取るまでの時間が長くなり，学習が進まなくなることを考慮し，我々のこれまでの研究ではエージェントの選択した「買い」，「売り」の 2 つの行動に機会損失を考慮した報酬を与えていた．本研究では新たに「何もしない」行動についても報酬を与えることで，エージェントのすべての行動に対して最適な行動であるかを評価する．さらに，我々のこれまでの研究では学習時の「買い」と「売り」の機会損失の計算に過去 n 日間の終値を用いて計算をしていた．本研究では，エージェントの選択した良い行動をより評価

†東京都市大学大学院総合理工学研究科　Graduate School of Integrative Science and Engineering, Tokyo City University

Copyright © 2022 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

するために，学習時のみ「買い」と「売り」の報酬において，過去だけでなく未来の終値を利用する．t日目の報酬R_tを以下のように定義する．

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

ここで，P_{sell}は売却時株価を，P_{buy}は購入時株価を，S_{all}は売却株数を，a_tは t 日目の行動を表す．α_t，β_t，γ_tはそれぞれ「売り」，「買い」，「何もしない」を選んだことにより発生する機会損失を考慮した適正度を表す項であり，以下のように表される．

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

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

$$\gamma_t = \begin{cases} \frac{A_{t+1} - A_t}{A_t} & if\ SH_t > 0 \\ 0 & if\ SH_t = 0 \end{cases}$$

ここで，PL_tは t 日目における前後 n 日間の終値が格納されているリストで，$max(PL_t)$，$min(PL_t)$はそれぞれPL_tの最大値，最小値を表す．このPL_tに保存された過去の情報を参照することで，図 2，図 3 に示すように評価日の価格に対して「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．我々のこれまでの研究では過去 n 日分の終値がPL_tに格納されていたが，学習時に評価日後 n 日分もPL_tに含むことで，過去だけでなく未来の値動きを考慮して正しい行動をエージェントに学習させる．γ_tにおけるSH_tはエージェントが t 日目に保有している株式数であり，A_tは t 日目の総資産である．本研究では先物取引を想定しているため，学習時には，評価日の総資産と比較して翌日の総資産が高ければ正の報酬を，低ければ負の報酬を与える．これにより，株式を保持している場合に相場が上がることを期待して売買を様子見した行動についての機会損失を考慮した報酬を表現している．

3. 実験

提案手法の有効性を示すため，実際の株式データを利用した実験を行った．対象は日本の株式市場の代表的な株価指数の一つである日経平均株価と日経プライム市場における 17 業種区分：電気・精密に含まれる企業銘柄を対象とした．対象株式データ使用期間は 2012 年 1 月～2021 年 12 月の 10 年間を扱う．学習期間を 5 年間，テスト期間を 1 年間として図 4 に示すような交差検証を行った．ニューラルネットワークのパラメータはそれぞれバッチサイズ 20，メモリーサイズ 200，隠れ層 3，隠れニューロン数 100，最

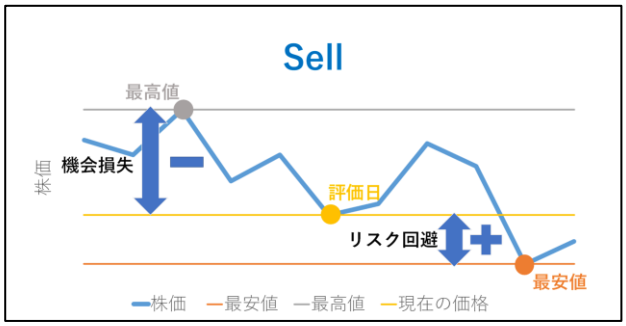

図 2　学習時における売却に対する機会損失

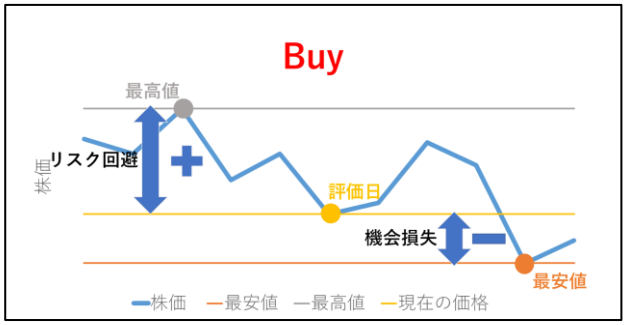

図 3　学習時における売却に対する機会損失

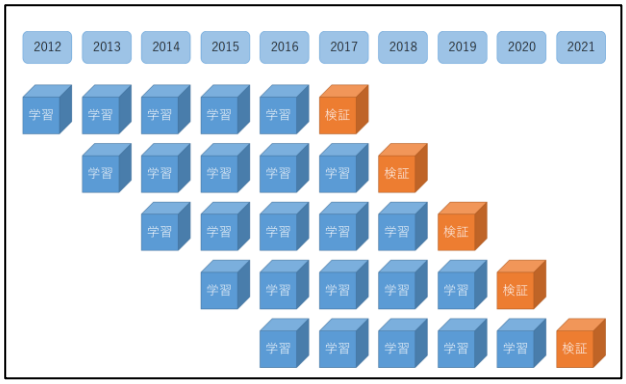

図 4　交差検証

適化手法には Adam optimizer，活性化関数には ReLU 関数と Softmax 関数を，損失関数には Huber 損失を用いた．

4. 実験結果

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] 近藤 巧麻, 松井 藤五郎, “複雑な環境における複利型深層強化学習を用いた金融取引戦略”, 人工知能学会全国大会(第 34 回), (2020).
[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)
[5] Jinho Lee, Raehyun Kim, Yookyung Koh, Jaewoo Kang. :Global Stock Market Prediction Based on Stock Chart Images Using Deep Q-Network, IEEE Access(Volume：7),pp.16726–167277 (2019)

Copyright © 2022 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.