DEIM 2021 E21-3

ポートフォリオマネジメント問題における 分散型強化学習を用いた低リスク投資行動の学習

佐藤　葉介†　　張　　建偉††

† 岩手大学大学院総合科学研究科〒 020-0066 岩手県盛岡市上田 4 丁目 3 - 5
†† 岩手大学理工学部〒 020-0066 岩手県盛岡市上田 4 丁目 3 - 5
E-mail: †{g0319083,zhang}@iwate-u.ac.jp

あらまし　金融市場は景気や政局などの多く複雑な要因が関わり変動するため，正確な予測や取引戦略の構築が困難である．近年，深層学習により金融市場における投資行動を獲得する研究が盛んである．複雑な環境の中で保有する資産価値が低下するリスクを防ぎつつ利潤を最大化させるような投資行動を学習することが主な研究課題である．一方，分散型強化学習は強化学習における行動価値関数を離散分布に拡張したものであり，とりうる行動により期待される Q 値を分布で表すことで単一の Q 値では表現できなかったリスクを学習できる．本研究では低リスクな投資行動を分散型強化学習により学習することを試みた．分散型強化学習はポートフォリオマネジメントにおける投資行動の学習には筆者の知る限りでは応用されておらず，本論文が最初の応用例となる．分散型強化学習を用いた実験では，多くの先行研究で利用されている DQN よりも標準偏差に関して優れた評価値を得ており，結果について両側検定を行ったところ評価値であるシャープレシオについて統計的有意差が得られた．

キーワード　多変量時系列データ，強化学習，分散型強化学習，低リスク投資行動，ポートフォリオマネジメント

1 はじめに

近年，深層学習を用いた金融市場に関する研究が盛んに行われている [1, 2]．金融市場は景気や為替などの経済的要因や政局などの経済外的要因など複雑な要因が関わり変動するため，確実な将来の状態予測や取引戦略の組み立てが困難な金融市場における投資行動の学習に関する研究はこれまでに多数されてきた [3–6]．特に近年は高い特徴表現力を持つ深層モデルを用いた取引エージェントの研究がされている [7–9]．多くの研究では利潤を増加させつつ保有する資産価値が減少するリスクへの対処をするという 2 つの課題に対して様々な手法が提案されてきた [10–13]．ほとんどの深層強化学習を用いた先行研究では Deep Q Network(DQN) が提案手法のベースや比較手法として用いられている．これらの研究で用いられる評価値は，利潤の増大を測るためにテスト期間に得られた資産の多さを利益率で表し，その標準偏差をどれだけ安定して利益を得られるかを調べるために利用し，これら 2 つの評価値を統合した，取ったリスクに対してリターンの大きさを示すシャープレシオ [14] が主に用いられている．

一方，分散型強化学習 [15] は深層強化学習における行動価値関数の各行動の評価値を値の分布に拡張した手法であり，ベンチマークにおいて DQN，Double DQN(DDQN)，Dueling Network より優れた結果を残している．分散型強化学習による行動価値関数では，ある行動で得られる報酬の期待値だけでなく定義した報酬の値の範囲で，各報酬が得られる期待値を離散分布で学習することができる．モデルの出力を人が観察しリスク操作をすることが可能な利点があり，行動価値関数の出力と手動で設定した歪度の要素積を計算することで，ある程度学習の対象ごとの性質に合わせた行動の選択をする改善手法も提案されている [16]．

本研究では先に述べた金融市場の不確実性から起こる資産価値低下のリスクに対して，筆者の知るところで金融分野で応用されていない分散型強化学習を適用することで，ポートフォリオマネジメント問題 [17] における低リスクな行動を獲得することを目的とする．日経 225 を構成する銘柄に対してバックテストを行い，DQN と比較してシャープレシオや得られた利益率などについて評価を実施した．特に評価値の標準偏差について DQN よりも提案手法の方が優位な結果を得られた．評価値の標準偏差が小さいほど安定した投資行動を期待できると考えられ，DQN に比べて低リスクな行動を学習できていると考えられる．また，結果に対して検定を行ったところシャープレシオについて統計的有意差が得られた．

2 関連研究

ポートフォリオマネジメントは利益の最大化やリスクの最小化を目的とした金融資産の分配をする問題であり，強化学習による適切な投資行動の学習が試みられている．価値ベースの手法として Shin [10] らは DQN を用いて 8 種類の暗号通貨と USD のポートフォリオマネジメントを学習させており，最も資産価値の減少を抑えた行動を学習できているため低リスクだとしている．方策ベースの手法として Xiong [11] らは Deep Deterministic Policy Gradient (DDPG) を用いて株価，保有株数，残高の状態からそれぞれの株に対する売却，保持，購入の行動を学習している．約 10 年分の Dow Jones 30 stocks に

ついてバックテストを行い比較手法より利益とシャープレシオについて優位な結果を残した．また，Ye [13] らも DDPG を用いており，ニュース記事と株価の推移を前処理したデータからポートフォリオの割り当てを学習している．ベンチマークにおいては比較手法より優位な利益を出し，シャープレシオも概ね優位な結果を残した．Direct Reinforcement Learning [18] は Fuzzy Learning と Recurrent Neaural Network を組み合わせた方策ベースの投資行動学習手法であり，時系列データのみで投資行動を決定できる．CNN ベースの DDPG を構築し 12 種類の暗号通貨のポートフォリオマネジメントを行った研究も存在する [19]．Jiang [20] らは暗号通貨のポートフォリオマネジメントを行う EIIE フレームワークを開発し UBAH などの比較手法より portfolio value やシャープレシオについて優位な結果を得ている．EIIE フレームワークは方策ベースの手法であり，資産の潜在的な成長性から直近の予測を行う IIE のアンサンブル学習である．著者 [21] らは分散型強化学習を初めて投資行動の学習に応用した．データセットは日経 225 を構成する銘柄を用いており，それぞれの銘柄に対して投資行動を学習し，結果を評価している．比較手法である DQN に比べて分散型強化学習は，評価値の平均値について多少優位な結果を出した．最終的な資産額の標準偏差についても分散型強化学習が優位な結果を出しており，得られる資産の大きさにおける安定性という意味で低リスクな手法だと言える．

金融市場に関する研究では投資行動の学習に限らずリスクを考慮した手法が提案されているが，関連研究や過去の研究では分散型強化学習をポートフォリオマネジメントに対して適用した例は存在しない．著者らの先行研究 [21] では 1 つの銘柄に対して投資行動を学習したが，全ての銘柄を同時に扱うような投資行動は学習していない．本研究は分散型強化学習をポートフォリオマネジメントにおける投資行動の学習に初めて応用する．

3 手　　法

3.1 問題設定

本研究では金融市場における投資行動をマルコフ決定過程 $M(\mathcal{S}, \mathcal{A}, R, P)$ とする．S は状態空間を表し，状態 s と $s \in \mathcal{S}$ のような関係がある．A は行動空間を表し，行動 a と $a \in \mathcal{A}$ のような関係がある．投資行動の対象資産は東京証券取引所第一部に上場している企業の株式と無リスク資産とし状態 s に含まれる．また，行動 a は保持している無リスク資産で株式を “購入”，保持している株式の “売却”，資産の売買を行わない “保持” の 3 つを取りうる．本手法では日足データを元に取引行動を行う．日足には始値，高値，低値，終値が含まれており，エージェントが購入または売却するときは終値を用いる．

$R : S \times A \to \mathbf{R}$ は報酬関数を表しており，時刻 t に得られる報酬は r_t とする．本手法では初期状態あるいは株式の購入時から売却時までの資産額の増減を報酬値に利用している．売却したときのみ即時報酬を与えるとそれまでの過程が評価されないため，報酬が得られたときに初期状態あるいは株式の購入を行ったときから売却したときまでの各状態 s に対して遅延報酬を与える．このとき，過去の状態に遡るにつれて割引率を適用することで偏りが現れる可能性を排除する．さらに各報酬値に対して DQN と同様に reward clipping を適用する．

遷移関数 $P : S \times A \times S \to [0, 1]$ は状態 s のとき行動 a を取り状態 s' へ遷移する状態遷移確率を表す．方策 $\pi(a \mid s)$ は状態 s の時の行動 a をとる確率を表す関数である．行動価値関数 $Q^\pi(s, a)$ は状態 s のとき方策 π に従い行動 a を取ったときに得られる期待報酬値を定義する．

$$Q^\pi(s, a) = \mathbf{E}[R(s, a)] + \gamma \mathbf{E}_{P,s}[Q^\pi(s', a')]$$

最適方策 π^* を学習し最適行動価値関数 $\mathbf{E}[Q^*(s, a)]$ の戻り値を最大化するような行動を学習することが目的となる．最適方策とは任意の初期状態 $s \in \mathcal{S}$ から期待報酬を最大化することである．最適方策の学習にはいくつか手法が存在するが，Q 学習では以下の更新式により最適行動価値関数を学習する．

$$Q_\theta(s, a) \leftarrow \mathbf{E}[R(s, a)] + \gamma \mathbf{E}_P[\underset{a'}{max} Q_\theta(s', a')]$$

分散型強化学習の拡張元手法である DQN は上記の Q 学習をベースとしている．

3.2 分散型強化学習

森村 [22] らは期待リターンの再帰式であるベルマン期待方程式のリターンを分布に拡張した分布ベルマン方程式を定義している．分布ベルマン方程式を解くことでリターン分布を推定できるが分布ベルマン方程式は汎関数の自由度を持つため一般に推定は困難であるため近似が必要となる．

Bellemare [15] はリターン分布を多項分布で近似した categorical DQN を提案している．リターン分布は直感的には複数個の bin と呼ばれる 1 つの報酬値が得られる期待値を表すものが連続している．ハイパーパラメータとして設定した数だけの bin 数でリターン分布が構成される．近似リターン分布の bin 数 $M \geqq 2$ と，近似リターン分布の上限 Q_{max} と下限 Q_{min} をハイパーパラメータとして定め，bin 間隔 Δ_z を

$$\Delta_z := \frac{Q_{max} - Q_{min}}{M - 1}$$

のように定数として定め，各 bin に対するリターン代表値 z_m , $m \in \{1, ..., M\}$ を

$$z_m := Q_{min} + (m - 1)\Delta_z$$

とする．状態 s と行動 a の入力に対するリターン分布を表現する M 次元ベクトル $(q_1(s, a)$,..., $q_M(s, a))$ を出力する深層モデル $\mathcal{S} \times \mathcal{A} \to \mathbf{R}^M$ を用いて，推定リターン分布 $\hat{P}$ を

$$\hat{P}(C = z_m | s, a) := \frac{exp(q_m(s, a))}{\sum_{m'=1}^{M} exp(q_{m'}(s, a))} ,$$

$$\forall m \in \{1, ..., M\}$$

として求める．このとき categorical DQN の行動価値の推定値 $\hat{Q}$ は

Algorithm 1 投資行動学習アルゴリズム

```
1: T ⇐ 0
2: done ⇐ True
3: episodes ⇐ 0
4: stack_memory ⇐ [ ]
5: env ⇐ InitializeEnvironment()
6: agent ⇐ InitializeAgent()
7: mem ⇐ InitializeReplayMemory()
8: while episodes < MAX_EPISODES do
9:    if done is True then
10:      state ⇐ env.reset()
11:      done ⇐ False
12:      episodes ⇐ episodes + 1
13:   end if
14:   action ⇐ agent.act_epsilon_greedy(state)
15:   next_state, reward, done ⇐ env.step(action)
16:   reward ⇐ reward_clipping(reward)
17:   stack_memory.append([state, action, reward, done])
18:   if action is selling then
19:      for i = 0 to len(stack_memory) do
20:         Apply discount reward to each stacked data
21:      end for
22:      for item in stack_memory do
23:         mem.append(item)
24:      end for
25:   end if
26:   if episodes > 1 then
27:      if T%REPLAY_FREQUENCY = 0 then
28:         agent.learn(mem)
29:      end if
30:      if done is True then
31:         evaluate_model(agent)
32:      end if
33:   end if
34:   T ⇐ T + 1
35:   state ⇐ next_state
36: end while
```

$$\hat{Q}(s,a) \triangleq \sum_{m=1}^{M} z_m \hat{P}(C = z_m | s, a)$$

となる．$\hat{Q}$ は DQN と同様に近似分布ベルマン行動最適作用素 $\hat{D}$ [15] を適用して現在の推定リターン分布 $\hat{P}$ から目的分布 $\hat{P}_n^{target}$ を求める．experience replay により得た経験 n を用いて目的分布 $\hat{P}_n^{target}$ と現在の推定分布 $\hat{P}(\cdot\,|s_n, a_n)$ との差異が小さくなるように深層モデルの重みを更新する．

学習は DQN と同様に experience replay を取り入れる [23]．学習において 1 ステップ更新されるごとに，replay memory に現在の状態，次の状態，評価値を保存する．n ステップに一度，バッチサイズだけ replay memory からデータを取り出し Target-Network の重みを学習する．

3.3 提 案 手 法

提案手法のアルゴリズムを Algorithm1 に示す．また，提案手法におけるデータフローを図 1 に示す．基本的な流れは一般的な強化学習を踏襲している．本手法では 225 銘柄に対するポートフォリオマネジメントを行っており，エージェントはどの銘柄に対してどの程度購入するか，それぞれの銘柄に対して売却を行うタイミング，保持しておくべき銘柄はどれかといった行動を学習する．3.1 節に従い Algorithm 1 のように遅延報酬をエージェントに与える．使用する深層モデルの入出力を図 2 に示す．入力次元数は状態 s として，n 日分の 225 銘柄の日足データを用いるため $n \times 225 \times 4$ となる．出力次元数は分散型強化学習における離散分布を構成する bin 数 M として 3 種類の行動を各銘柄に対して学習するため $M \times 3 \times 225$ となる．

提案手法のアルゴリズムを Algorithm1 に示す．基本的な流れは一般的な強化学習を踏襲している．1 行目から 7 行目にかけて変数の初期化を行う．T はタイムステップを表しており 1 ステップ進むごとに 1 日だけ時間を進める．$done$ はエポックの終了判定に，$episodes$ はエピソード数のカウントに使用する．$stack\_memory$ は遅延報酬を与えるまでのスタックである．env，$agent$ はそれぞれ環境とエージェントである．mem は replay memory に利用する．9 行目から 13 行目にかけて初期化処理を行っている．1 エポックが終了したときやプログラム開始時の初期化に利用する．14 行目から 16 行目にかけてはエージェントの行動と環境のインタラクション，報酬の獲得が行われる．17 行目から 25 行目にかけて 3.1 項に従い Algorithm 1 のように遅延報酬をエージェントに与える．26 行目に示すように 1 エポックだけデータ取得期間を得てから，27 行目から 29 行目で，指定した頻度でエージェントの学習を行う．31 行目に示すようにエポックの終了ごとにモデルの評価を実施する．34, 35 行目で次のステップへの変数の更新を行う．

4 実　　験

categorical DQN による分散型強化学習を用いて，バックテストにより金融市場における投資行動を学習する実験を行った．DQN と比較して最終的な資産額，利益率の平均，利益率の標準偏差，シャープレシオについて評価した．

4.1 データセットと前処理の手法

データセットは東京証券取引所第一部に上場しており，日経 225 に含まれる 225 銘柄を利用した．期間は 2010 年 1 月 4 日から 2019 年 12 月 30 日までの 10 年間の日足データを利用し，10 年分のデータが存在しない銘柄については，データが存在する年から 2019 年 12 月 30 日まで利用する．そのうちより新しい 1 年のデータを評価に用いて，より過去のデータを学習に用いる．データが存在しない期間がある銘柄については，その期間についてその銘柄を投資行動の対象とせず学習を行う．市場は過去の状態の影響を受けて将来の状態が決定していると考えられるため，評価において未来の情報を学習していないモデルを用いるようにする．

日足には始値 (open price)，高値 (high price)，安値 (low price)，終値 (close price) の 4 つの変数が含まれ，それぞれの

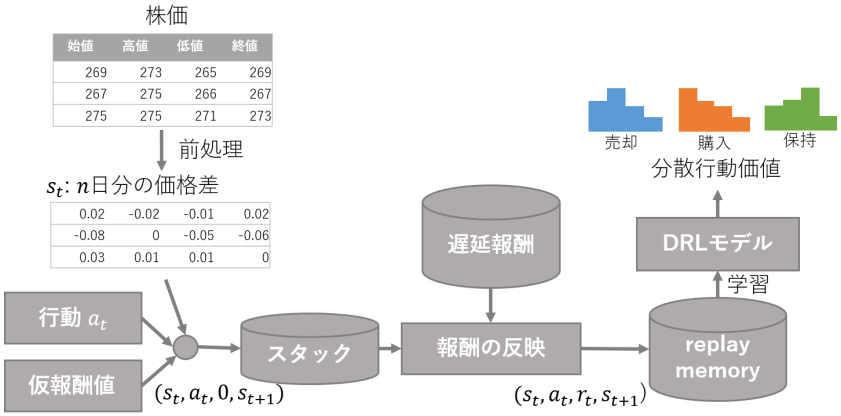


図 1: 提案手法のデータフロー

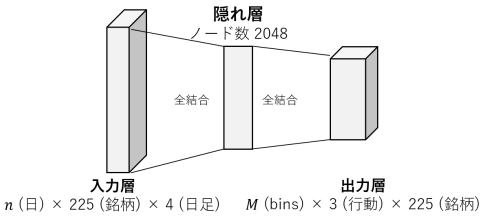


図 2: 提案手法のモデル

Algorithm 2 エージェント行動アルゴリズム

```
ActionValues ⇐ Model(s_t)
StockRatioForBuy ⇐ [ ]
Sum ⇐ 0
for i = 1 to 225 do
    Action = GetMaxValAction(ActionValues[i])
    if Action is sell then
        sell stock i
    end if
end for
for i = 1 to 225 do
    Action ⇐ GetMaxValAction(ActionValues[i])
    if Action is buy then
        StockRatioForBuy[i] ⇐ Max(ActionValues[i])
        Sum+ = StockRatioForBuy[i]
    else
        StockRatioForBuy[i] ⇐ 0
    end if
end for
for i = 1 to 225 do
    StockRatioForBuy[i] / = Sum
    SubRiskFreeAsset ⇐ RiskFreeAsset×StockRatioForBuy[i]
    buy stock i using SubRiskFreeAsset
end for
```

変数について前処理を行う．本手法では前日からの値動き，すなわち差分を学習させる．さらに DQN と同様に複数ステップの情報をまとめて 1 つの状態とする．1 銘柄における，あるステップ t における n 日分の時系列データは以下のようになる．

$$v_t^{diff} = v_t - v_{t-1}$$

$$\boldsymbol{v_t} = (f(v_t^{open\ diff}), f(v_t^{high\ diff}), f(v_t^{low\ diff}), f(v_t^{close\ diff}))$$

$$\boldsymbol{e_t} = (\boldsymbol{v_t}, \boldsymbol{v_{t-1}}, ..., \boldsymbol{v_{t-n+1}})$$

v_t^{diff} はステップ t における前ステップとの差分を表す．f は引数として得た値に対して正規化を行う関数とする．225 銘柄分の $\boldsymbol{e_t}$ を以下のようにベクトル化したものが状態 $\boldsymbol{s_t}$ となる．$\boldsymbol{s_t}$ はステップ t における環境から観測される状態とする．replay memory には $\boldsymbol{s_t}$ を保存する．

$$\boldsymbol{s_t} = (\boldsymbol{e_t^1}, \boldsymbol{e_t^2}..., \boldsymbol{e_t^{225}})$$

次状態 $\boldsymbol{s_{t+1}}$ はステップ t+1 において同様に計算したものである．r_t は 3.1 節で述べたように報酬が得られてから与えられるため，それまで $(\boldsymbol{s_t}, \boldsymbol{a_t}, 0, \boldsymbol{s_{t+1}})$ の組をスタックする．ここで，$\boldsymbol{a_t}$ は 225 銘柄それぞれに対する行動を集めたベクトル，0 は仮の報酬値を表す．報酬が得られてからスタックしたデータに割引率を適用した報酬を与え，replay memory に保存する．

4.2 日経 225 データを用いた投資学習実験

実験環境は Open AI Gym を利用して構築した．モデルや分散型強化学習 (DRL) に利用するパラメータは固定し投資行動の学習を行った．初期状態として投資エージェントは無リスク資産である ¥1,000,000 を所有する．環境から観測される状態は日足であり，4.1 節のように前処理を行い replay memory に保存する．学習データを用いて 1 epoch だけ replay memory を構築してから学習を開始し，設定した頻度で experience replay による学習を行う．次に学習したモデルとリセットした初期資産を用いて評価期間について投資行動を行う．資産額は無リスク資産と保持している株式の時価額の和で計算する．エージェントが売却を行ったときの資産額を $Asset\_sell$，購入したの資産額を $Asset\_buy$ としたとき，売却時の利益率は $Asset\_sell/Asset\_buy$ により計算される．1 回の実験では最終的な資産額，評価期間における利益率の平均値，利益率の標準偏差，シャープレシオによる評価値を求める．評価は 5 エポック学習してから実施した．シャープレシオ [14] は評価期間における利益率の平均値を利益率の標準偏差で割った値とする．実

表 1: 各評価値の平均値

	DQN	DRL (提案手法)
(a) 最終資産額 (円)	**1,085,988**	1,075,999
(b) 平均利益率	**1.000387**	1.000354
(c) 利益率標準偏差	0.0106	**0.01007**
(d) シャープレシオ	96.175	**100.153**

表 2: 各評価値の標準偏差

	DQN	DRL (提案手法)
(e) 最終資産額 (円)	109,665	**63,324**
(f) 平均利益率	4.24×10^{-4}	$\mathbf{2.44 \times 10^{-4}}$
(g) 利益率標準偏差	1.54×10^{-3}	$\mathbf{9.48 \times 10^{-4}}$
(h) シャープレシオ	12.841	**9.522**

験では無リスク資産の利子率は 0 とする．モデルの初期状態や方策が ε -greedy 法でありランダムな要素を含むため 100 回実験を行い各評価値の平均と標準偏差を求める．

DRL と DQN の共通パラメータとして，予備実験により，モデルは 3 層全結合とし隠れ層のノード数を 2048，Q 学習の割引率を 0.9，replay frequency を 4，experience replay におけるバッチサイズを 16，Adam- ε を 1.5×10^{-2}，1 状態に含める日数を 5 日とした．DRL のパラメータは，モデルの最終層の bin 数を 71，Vmax を 10，Vmin を-10 とした．

最終的な資産額，評価期間における利益率の平均値，利益率の標準偏差，シャープレシオによる評価値で，提案手法である DRL と比較手法である DQN について比較する．平均値における評価結果を表 1 に，標準偏差における結果を表 2 に示す．

平均値の結果のうち最終資産額，平均利益率，シャープレシオは値が大きいほど優れており，平均値のうち利益率の標準偏差と，各評価値の標準偏差の結果については値が小さいほどばらつきが少なく優れている．100 回実験を行ったところ，評価値の平均値では，(c) と (d) に示すように DRL の方が評価値の標準偏差とシャープレシオ，標準偏差は (e) から (h) に示す結果について提案手法が上回る結果となった．特に標準偏差の結果については大きな差があり，(e) 最終資産額では DRL は DQN の 57.7%，(f) 平均利益率は 57.6%，(g) 利益率の標準偏差は 61.6%，(h) シャープレシオは 74.2%となっている．

評価値の平均値については多少の優劣はあるものの大きな差はなかった．しかし，評価値の標準偏差について DQN より DRL の方が大きく優れた結果となったといえる．

5 考　　察

5.1 最終資産額のヒストグラムの比較

本実験では DQN と DRL について 100 回実験を行い，評価値の平均と標準偏差を計算した．それぞれの実験結果について，最終資産額のヒストグラムを図 3 に示す．最終資産額の平均では DQN は DRL の 1.0092 倍となったが，ヒストグラムを観察してわかるように DQN の方がばらつきが大きく，DQN の標準偏差は DRL の 1.73 倍である．DRL は比較的最終的な資産額が小さかった一方，全体的には安定した投資行動を学習したと考えられる．DQN は DRL に比べるとより利益が大きかった反面，損失を出した結果も多かった．DRL の評価値の標準偏差では結果的に DQN よりも優位となった．

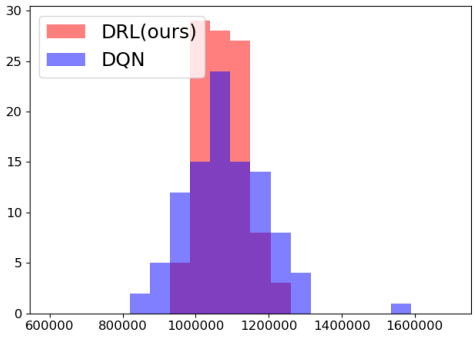


図 3: 最終資産額における実験結果のヒストグラムの比較

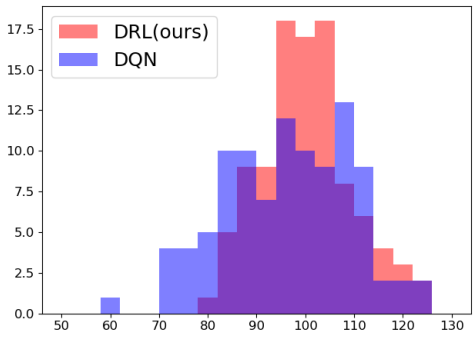


図 4: シャープレシオにおける実験結果のヒストグラムの比較

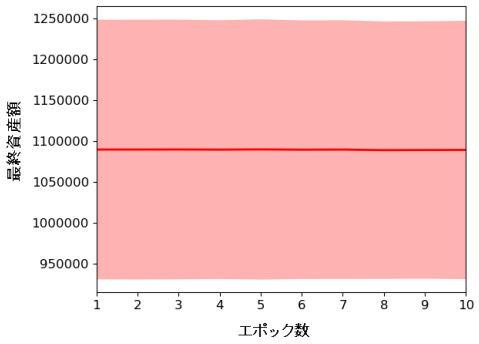


図 5: (DQN) 学習エポック数ごとの最終資産額の変化

シャープレシオのヒストグラムを図 4 に示す．表 1 が示すように，DRL の方が平均値は大きく，優れていることを示しており，ヒストグラムを観察しても DRL の方がシャープレシオの値が大きい方に分布していることが観察できる．また，ばらつきに関しても DRL の方が小さいことがわかる．

5.2 エポック数による評価値の変化

本実験では 5 エポック学習したモデルを利用して評価を行っている．前実験として，10 エポックほど学習を行いそれぞれのエポック終了時の評価値を観察し，実験でどのエポック数を評価するか決定した．

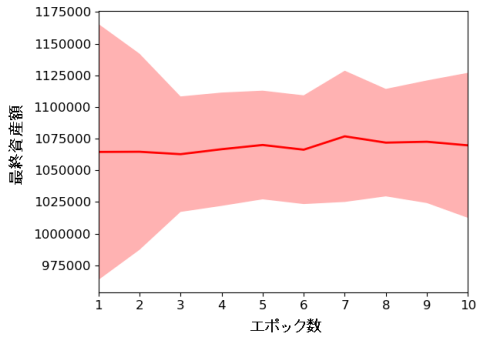


図 6: (DRL) 学習エポック数ごとの最終資産額の変化

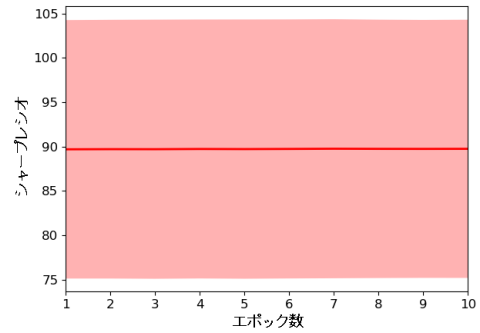


図 7: (DQN) 学習エポック数ごとのシャープレシオの変化

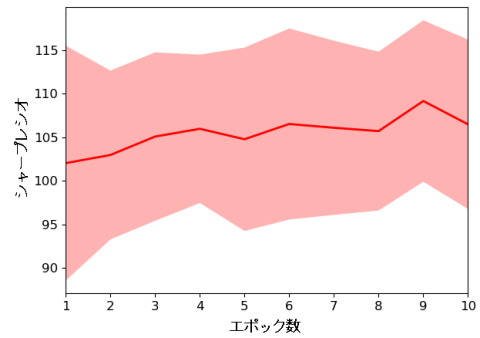


図 8: (DRL) 学習エポック数ごとのシャープレシオの変化

図 5〜図 8 は赤い線が平均値，薄い赤が標準偏差 $\pm 1\sigma$ を示している．最終的な資産額では，DQN(図 5) の結果はエポック数が増加してもほとんど変化が無かった．一方，DRL(図 6) は 3 エポックで大幅な標準偏差の減少がみられ，平均値は微小な上昇傾向が続いた．シャープレシオについて観察すると，DQN(図 7) はほとんど変化が現れなかった一方，DRL(図 8) は全体的に上昇傾向がみられた．また，DRL の標準偏差は学習が進むにつれて減少傾向がみられた．よって本実験では最終的な資産額を優先し，比較的安定した評価値の推移がみられる 5 エポックほど学習し評価値を算出することとした．

5.3 有意性の検証

本研究では各評価値について平均値と標準偏差を計算しているが，結果の有意性について調べる．DRL と DQN により得られた結果を，各評価値について 2 つの群として代表値の有意

表 3: シャピロ-ウィルク検定による各評価値の p 値

	DQN	DRL (提案手法)
最終資産額 (円)	1.132×10^{-3}	7.327×10^{-3}
利益率の平均値	4.370×10^{-2}	4.293×10^{-2}
利益率の標準偏差	9.350×10^{-5}	0.560
シャープレシオ	0.406	0.217

表 4: 各評価値の有意性の検定結果

評価値	p 値
最終資産額 (円)	0.596
利益率の平均値	0.614
利益率の標準偏差	0.0611
シャープレシオ	0.018

性を検定する．まず，DRL と DQN の平均値と標準偏差それぞれの評価値の結果について，シャピロ-ウィルク検定により標本に正規性があることを帰無仮説としテストした．有意水準を 0.05 とすると表 3 に示す p 値のように，利益率の標準偏差と DRL におけるシャープレシオ以外のデータについて帰無仮説が棄却され正規性を持たないことがわかった．

DQN と DRL の結果にはデータ間の対応が無い．表 3 の結果から最終資産額，利益率の平均，利益率の標準偏差の結果についてはウィルコクソンの順位和検定を行った．シャープレシオについて F 検定を行ったところ 2 群間は p 値が 0.005 となり有意水準 1%で棄却され，不等分散であることが示されたため，Welch の t 検定を行った．これらの検定は両側検定で実施した．各検定結果を表 4 に示す．

Welch の t 検定の結果，シャープレシオの結果に関しては有意水準 5%で統計的有意差があることがわかった．その他の結果については統計的有意差が存在しないことがわかった．よって 4.2 節の結果のうち，利益率の標準偏差とシャープレシオの結果に関しては統計的有意差が存在する．

5.4 学習した投資行動

本実験ではポートフォリオマネジメントを学習している．本節ではモデルが学習した行動の一例を示すとともに，実験結果や銘柄と比較し，学習の傾向を考察する．

bin 数 51 についてテスト期間におけるポートフォリオの推移の 1 例を図 9 に示す．積み上げ面グラフであり，銘柄ごとのポートフォリオのうち占める額と無リスク資産を統合している．図に示した色は銘柄ごとに異なっており，最も上に積み上げられた面グラフは無リスク資産を表している．

一見すると特定の銘柄がポートフォリオを占める割合が増加しているが，株価は変化せず購入し続けていて増加しているのかなど，要因が判別できない．そこで，学習 5 エポックの結果について，ポートフォリオへの寄与度を計算した．本アルゴリズムでは特定の銘柄について，持ち株が徐々に増加することはあっても徐々に減少することはない．そのためそれぞれの銘柄について，持ち株がある期間に株価の上下を判別し，テスト期間中には複数回取引を行うことがあるがその上下幅の和を計算

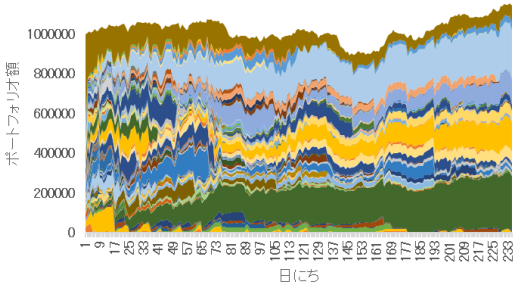


図 9: 学習エポック数によるポートフォリオ推移

表 5: ポートフォリオ増加に寄与した銘柄

寄与度順位	銘柄コード
1	9009
2	6702
3	4043
4	8729
5	1928
6	4911
7	8354
8	9437
9	9301
10	4183

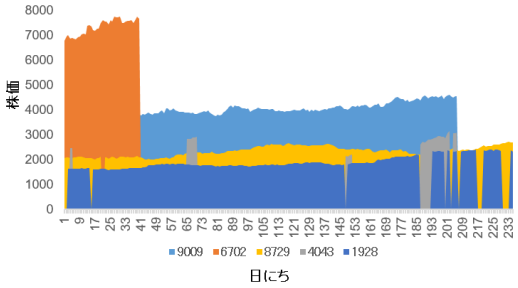


図 10: 高寄与度銘柄における投資期間内の株価推移 (1〜5 位)

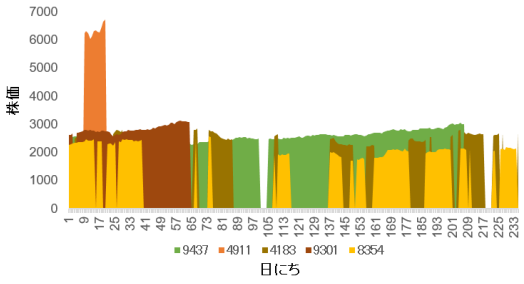


図 11: 高寄与度銘柄における投資期間内の株価推移 (6〜10 位)

し銘柄順でソートすることで寄与度を求めた．株価のみを計算に用いることで単に持ち株が増加したからポートフォリオが増加したというファクターを除外している．ただし持ち株と株価が増加するような要因でポートフォリオが増加したというファクターも除外される．

表 5 に寄与度順の銘柄コードを，図 10 と図 11 に寄与度が高い銘柄について持ち株が存在したときの株価を示す．最も寄与

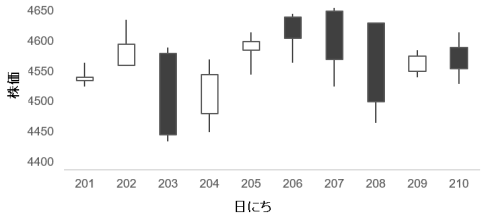


図 12: 銘柄コード 9009 における売却行動時周辺の株価推移

度が高い銘柄コード 9009 は明らかに上昇傾向の区間のみで取引できており，ポートフォリオの増加に寄与していることがわかる．1〜5 位についてはどれも上昇傾向の区間で取引できている．6 位以下もそれより上の順位ほどではないが緩やかな増加傾向がみられた．

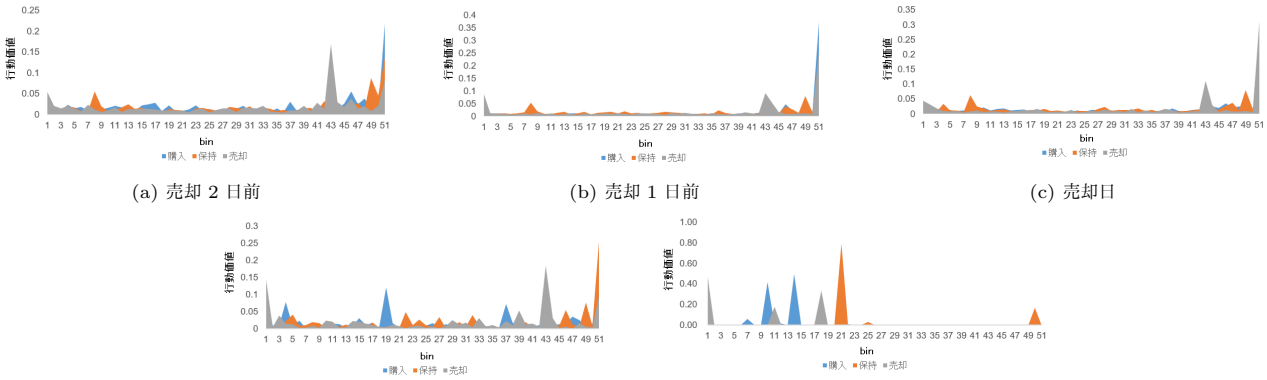
銘柄コード 9009 について，売却を行った日にち周辺のデータを観察する．206 日目に売却を行って以降，テスト期間において購入の行動を行っていない．このとき購入時から売却時まで株価は約 1.3 倍となっていた．まず，図 12 に売却を行った日にち周辺の終値を示す．横軸は日にち，縦軸は株価を示している．このうちモデルに入力された区間は 202 から 206 日である．203 日で株価が下がったのち回復し，208 日で再び下がるような値動きをしている．

次に，図 13 に売却を行った周辺の日にちの銘柄コード 9009 の各行動の分散行動価値を示す．図に示す行動価値は 9009 のみであり，実際はモデルは各銘柄の各行動について分散価値を出力している．面グラフで示しており，売却の面グラフが最も手前に表示されている．縦軸は bin に対応する行動価値，横軸は bin 数を表している．横軸はより値が大きいほど正の値として扱われる．売却の分散行動価値のうち，ほとんどの局面で z_{43} が最も大きい．売却日のみ，売却の分散行動価値の z_{51} が大きいことがわかる．207 日，208 日にかけて株価は降下しており，いずれの日にちでも保持の行動価値が最大であった．特に大きく株価が下落している 208 日では購入や売却の行動価値が低いことがわかり，学習モデルは安定志向な行動をとっていると捉えることができる．

6 結　　論

本研究では分散型強化学習を用いて日経 225 を構成する銘柄における金融市場において，ポートフォリオマネジメントを実施する投資行動の学習を行い DQN と比較した．比較手法に比べて提案手法は 8 つのうち 6 つの項目で優位な結果を得ており，特に評価値の標準偏差については DQN に対して大きく優れている．さらに実験結果について両側検定を行ったところ，シャープレシオの結果に関して統計的有意差が存在することがわかった．学習エポック数の観点では，DQN は 1 エポックを学習した時点で評価値が収束しており，学習の速さという点では DRL よりも優れていると言える．一方，DRL は 5 倍程度の学習で大幅な安定性を得られている点が優位である．

今後の展望としては，比較手法が少ないため類似研究が多く採用している DDPG [24] などの手法との比較と，DRL がどの



(a) 売却 2 日前　(b) 売却 1 日前　(c) 売却日

(d) 売却 1 日後　(e) 売却 2 日後

図 13: 売却日周辺の分散行動価値

ような場面でどのような投資行動を学習していたのかを詳しく分析することが課題である．他にも別の金融商品に対する有用性の検証も考えられる．

文　献

[1] Nicholas Tung Chan and Christian Shelton. An electronic market-maker. In *AI Memo 2001-005*, 2001.

[2] J. B. Heaton, N. G. Polson, and J. H. Witte. Deep learning in finance. In *arXiv:1602.06561*, 2018.

[3] Ryota Ishihara. Topix trading ai using malti-layer neural networks and ga. In *The 19th JSAI Special Interest Group on Financial Informatics*, 2017.

[4] Ohki Kato and Hajime Anada. Construction of a financial transaction strategy tree using genetic programming by technical index. In *The 24th JSAI Special Interest Group on Financial Informatics*, 2020.

[5] Tsubasa Ueda and Takuo Higashide. Monetary policy analysis and investment strategy using artificial intelligence and economists' forecast distribution. In *The 18th JSAI Special Interest Group on Financial Informatics*, 2017.

[6] Junya Miyasaka and Hajime Anada. The investment decision making considering psychological factor. In *The 18th JSAI Special Interest Group on Financial Informatics*, 2017.

[7] Jia Wu, Chen Wang, Lidong Xiong, and Hongyong Sun. Quantitative trading on stock market based ondeep reinforcement learning. In *IJCNN 2019*, 2019.

[8] Hiroyuki Kobayashi, Kiyoshi Izumi, Hiroyasu Matsushima, Hiroki Sakaji, and Takashi Shimada. The construction of high frequency trading strategy viareinforcement learning. In *The 24th JSAI Special Interest Group on Financial Informatics*, 2020.

[9] Shota Tokoi and Hajime Anada. Trading system using deep reinforcement learning. In *The 22nd JSAI Special Interest Group on Financial Informatics*, 2019.

[10] Wonsup Shin, Seok-Jun Bu, and Sung-Bae Cho. Automatic financial trading agent for low-risk portfolio management using deep reinforcement learning. In *arXiv:1909.03278*, 2019.

[11] Zhuoran Xiong, Xiao-Yang Liu, Shan Zhong, Hongyang (Bruce) Yang, and Anwar Walid. Practical deep reinforcement learning approach for stock trading. In *NIPS 2018 Workshop on Challenges and Opportunities for AI in Financial Services*, 2018.

[12] Yifan Zhang, Peilin Zhao, Qingyao Wu, Bin Li, Junzhou Huang, and Mingkui Tan. Cost-sensitive portfolio selection via deep reinforcement learning. In *arXiv:2003.03051*, 2020.

[13] Yunan Ye, Hengzhi Pei, Boxin Wang, Pin-Yu Chen, Yada Zhu, and Bo Li Jun Xiao. Reinforcement-learning based portfolio management with augmented asset movement prediction states. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

[14] William F. Sharpe. The sharpe ratio. *The Journal of Portfolio Management*, pages 49–58, 1994.

[15] Marc G. Bellemare, Will Dabney, and Remi Munos. A distributional perspective on reinforcement learning. In *ICML 2017*, 2017.

[16] Will Dabney, Georg Ostrovski, David Silver, and Remi Munos. Implicit quantile networks for distributional reinforcement learning. In *ICML 2018*, 2018.

[17] Harry Markowitz. Portfolio selection, 1959.

[18] Y. Deng, F. Bao, Y. Kong, Z. Ren, and Q. Dai. Deep direct reinforcement learning for financial signal representation and trading. *IEEE Transactions on Neural Networks and Learning Systems*, 28(3):653–664, 2017.

[19] Zhengyao Jiang and Jinjun Liang. Cryptocurrency portfolio management with deep reinforcement learning. In *IntelliSys 2017*, 2017.

[20] Zhengyao Jiang, Dixing Xu, and Jinjun Liang. A deep reinforcement learning framework for the financial portfolio management problem. In *arXiv:1706.10059*, 2017.

[21] Yosuke Sato and Jianwei Zhang. Modeling low-risk actions from multilative time series data using distributional reinforcement learning. In *iCAST 2020*, 2020.

[22] Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Parametric return density estimation forreinforcement learning. In *UAI 2010*, 2010.

[23] Mnih Volodymyr, Kavukcuoglu Koray, Silver David, Rusu Andrei A., Veness Joel, Bellemare Marc G., Graves Alex, Riedmiller Martin, Fidjeland Andreas K., Ostrovski Georg, Petersen Stig, Beattie Charles, Sadik Amir, Antonoglou Ioannis, King Helen, Kumaran Dharshan, Wierstra Daan, Legg Shane, and Hassabis Demis. Human-level control through deep reinforcement learning. *Nature*, pages 529–533, 2015.

[24] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *arXiv:1509.02971*, 2015.