ĐẠI HỌC CÔNG NGHỆ THÔNG TIN - ĐHQG TPHCM

BÁO CÁO ĐỒ ÁN CUỐI KÌ

ĐỀ TÀI:
DEEP REINFORCEMENT LEARNING FOR AUTOMATED STOCK TRADING

MÔN HỌC: TRÍ TUỆ NHÂN TẠO

LỚP: CS106.M11

GIẢNG VIÊN: LƯƠNG NGỌC HOÀNG

	Họ và tên	MSSV
1	Phan Đại Dương	19520482
2	Huỳnh Văn Hùng	19521564
3	Hồ Mỹ Hạnh	19521470
4	Phan Trọng Hậu	19520077

Mục lục

I. Giới thiệu:

Hiện nay, chứng khoán ngày càng phát triển mạnh mẽ, thu hút nhiều nhà đầu tư tham gia. Khi tham gia vào các sàn giao dịch, mỗi nhà đầu tư cần có sự tìm hiểu kỹ lưỡng cùng một chiến lược giao dịch đúng đắn, hiệu quả. Tuy nhiên, trong thị trường chứng khoán phức tạp và đầy biến động, việc xây dựng một chiến lược là điều không dễ dàng, đặc biệt là với những nhà đầu tư thiếu kinh nghiệm.

Học tăng cường sâu gần đây cũng được áp dụng trong lĩnh vực tài chính. Trong phạm vi đề tài này nhóm tiến hành tìm hiểu một số thuật toán phổ biến hiện nay: Proximal Policy Optimization (PPO), Advantage Actor Critic (A2C), và Deep Deterministic Policy Gradient (DDPG). Bên cạnh đó, nhóm còn tìm hiểu thêm một chiến lược kết hợp các thuật toán nói trên. Chiến lược này cố gắng kế thừa ưu điểm của từng thuật toán trên từng xu hướng thị trường, từ đó tìm ra chiến lược giao dịch tối ưu, tối đa hóa lợi nhuận.

Nhóm sẽ đánh giá các thuật toán trên tập dữ liệu là 30 Dow Jones Stocks (từ năm 2009 đến năm 2020) được thu thập từ Yahoo Finance. Các thuật toán sẽ được đánh giá, so sánh với chỉ số Dow Jones Average (Dow Jones Average Index - DJIA) và Minimum variance portfolio, một phương pháp cơ bản dựa trên lợi nhuận kỳ vọng và phương sai (dùng để ước tính tỉ lệ rủi ro).

II. Mô tả bài toán:

1 Mô hình MDP:

Với tính chất ngẫu nhiên (stochastic) của thị trường chứng khoán năng động,phức tạp như hiện nay, nhóm cài đặt mô hình MDP như sau:

- State $\mathbf{s} = [\mathbf{p}, \mathbf{h}, b]$: một vector bao gồm giá cổ phiếu (stock prices) $\mathbf{p} \in \mathbb{R}_+^D$, số lượng cổ phiếu (stock shares) $\mathbf{h} \in \mathbb{Z}_+^{\mathbf{D}}$, số dư tài khoản(balance) $b \in \mathbb{R}_+$. Trong đó D là số cổ phần, $\mathbb{Z}_+$ là tập số nguyên không âm.t
- Action **a**: một vector gồm các hành động được thực hiện lên D cổ phần.
- Reward *r(s,a,s')*: Phần thưởng nhận được ngay khi thực hiện hành động$a$ ở trạng thái s và chuyển sang trạng thái *s'*.
- Policy $\pi(s)$:chiến lược giao dịch tại trạng thái s (phân phối xác suất của các hành động tại trạng thái s.)

- Q-value $Q_\pi(s, a)$: Phần thưởng kì vọng khi thực hiện hành động a ở trạng thái s theo chiến lược π

Ở mỗi trạng thái, một trong ba hành động có thể thực hiện lên cổ phần d (d=1,. . . ,D) trong danh mục đầu tư (porfolio):

- Bán $\mathbf{k}[d] \in [1, \mathbf{h}[d]]$ cổ phiếu dẫn đến $\mathbf{h_{t+1}}[d] = \mathbf{h_t}[d] - \mathbf{k}[d]$.
- Giữ thì $\mathbf{h_{t+1}}[d] = \mathbf{h_t}[d]$.
- Mua $\mathbf{k}[d]$ cổ phiếu dẫn đến $\mathbf{h_{t+1}}[d] = \mathbf{h_t}[d] + \mathbf{k}[d]$.

Ở thời gian t, một hành động được thực hiện thì giá cổ phiếu sẽ cập nhật ở thời gian t+1. Theo đó, giá trị các danh mục đầu tư sẽ thay đổi từ "porfolio value 0" đến "porfolio value 1" hay "porfolio value 2", "porfolio value 3" được minh hoạ ở Hình 2. Lưu ý là giá trị danh mục đầu tư (porfolio value) được tính bằng $b + \mathbf{p^T}.\mathbf{h}$

2 Kết hợp một số ràng buộc trong giao dịch chứng khoán:

Dưới đây là một số ràng buộc có thể ảnh huưởng tới việc giao dịch chứng khoán trong thực tế:

- Tính thanh khoản(Market liquidity).
- Số dư tài khoản không âm $b \geq 0$: Dựa trên hành động tại thời gian t, các cổ phần sẽ được chia ra làm 3 tập: bán S, mua B, giữ H, trong đó $S \cup B \cup H = \{1, \ldots, D\}$ và chúng không trùng lặp với nhau. Gọi $\mathbf{p}_t^B, \mathbf{k}_t^B = [p_t^i, k_t^i : i \in B]$ là vecto giá tiền và số lượng cổ phiếu trong tập B. Tương tự cho tập S và H.Từ đó ta có thể biểu diễn ràng buộc số dư không âm như sau:
$$b_{t+1} = b_t + (\mathbf{p}_t^S)^T \mathbf{k}_t^S - (\mathbf{p}_t^B)^T \mathbf{k}_t^B \geq 0$$
- Chi phí giao dịch: bằng 0.1% giá của mỗi giao dịch (mua hoặc bán):
$$c_t = \mathbf{p}^T \mathbf{k}_t \times 0.1\%$$
- Mức lo ngại rủi ro sự sụp đổ của thị trường: Nhóm sử dụng financial turbulence index $turblulence_t$ giúp đo được sự dịch chuyển giá cả với quy mô lớn:

$$turblulence_t = (\mathbf{y_t} - \mu)\mathbf{\Sigma^{-1}}(\mathbf{y_t} - \mu)' \in \mathbb{R}$$

Trong đó $\mathbf{y_t}$ là tỉ suất lời cổ phiếu ở thời gian t, μ là trung bình tỉ suất ở quá khứ và $\mathbf{\Sigma}$ là ma trận hiệp phương sai của tỉ suất lời quá khứ. Khi $turblulence_t$ cao hơn một ngưỡng cho trước có nghĩa là thị trường đang có thể gặp vấn đề nghiêm trọng, ta cần tạm dừng mua cổ phiếu và bán tất cả cổ phiếu hiện có. Ngược lại thì việc giao dịch sẽ tiếp tục tiến hành.

3 Mục tiêu là tối ưu hóa lợi tức nhận được:

Nhóm định nghĩa hàm reward là sự thay đổi giá cả các danh mục đầu tư (porfolio value) khi thực hiện một hành động a tại trạng thái s và chuyển sang trạng thái s'. Mục tiêu là thiết kế ra một chiến lược tối đa hoá được giá trị của sự thay đổi đó:

$$r(s_t, a_t, s_{t+1}) = (b_{t+1} + \mathbf{p_{t+1}}^T\mathbf{h_{t+1}}) - (b_t + \mathbf{p_t}^T\mathbf{h_t}) - c_t$$

Trong đó: $\mathbf{h}_{t+1} = \mathbf{h}_t - \mathbf{k}_t^S + \mathbf{k}_t^B$ và $b_{t+1} = b_t + (\mathbf{p}_t^S)^T\mathbf{k}_t^S - (\mathbf{p}_t^B)^T\mathbf{k}_t^B$

Do đó hàm reward có thể viết lại như sau:

$$r(s_t, a_t, s_{t+1}) = r_H - r_S + r_B - c_t$$

với $r_H = (\mathbf{p}_{t+1}^H - \mathbf{p}_t^H)^T\mathbf{h}_t^H$, $r_S = (\mathbf{p}_{t+1}^S - \mathbf{p}_t^S)^T\mathbf{h}_t^S$, $r_B = (\mathbf{p}_{t+1}^B - \mathbf{p}_t^B)^T\mathbf{h}_t^B$

Trong đó, r_H, r_S, r_B là giá trị thay đổi porfolio value có được từ việc giữ, bán, mua cổ phiếu từ thời gian t đến thời gian t+1. Hàm reward trên cho thấy mục tiêu của ta là cần phải tối đa hoá giá trị tăng thêm của porfolio bằng việc mua và giữ cổ phiếu được dự đoán sẽ tăng giá ở mốc thời gian tiếp theo. Giá trị giảm bớt của porfolio value cũng cần được tối thiểu hoá bằng cách bán cổ phiểu sẽ giảm giá ở mốc thời gian tiếp theo.

Ta cần phải kết hợp turbulence index với hàm reward để phòng trường hợp thị trường bị sụp đổ. Khi chỉ số vượt ngưỡng cho trước:

$$r_{sell} = (\mathbf{p_{t+1}} - \mathbf{p_t})^T\mathbf{k_t}$$

Mục đích là để tối thiểu hoá sự sụt giảm porfolio value bằng việc bán hết tất cả cổ phiếu trước khi bị rớt giá do thị trường bị sụp đổ.

Mô hình được khởi tạo như sau: p_0 là thị giá cổ phiếu tại thời điểm 0, b_0 là lượng tiền vốn ban đầu. h và $Q_\pi(s, a)$ đều là 0. $\pi(s)$ là phân phối đồng đều cho tất cả các hành động đối với mỗi trạng thái. sau đó $Q_\pi(s, a)$ sẽ được cập nhập thông qua việc tương tác với môi trường chứng khoáng cho trước. Chiến lược tối ưu sẽ đạt được từ Bellman Equation:

$$Q_\pi(s_t, a_t) = \mathbb{E}_{s_{t+1}}[r(s_t, a_t, s_{t+1}) + \gamma\mathbb{E}_{s_{t+1}\sim\pi(s_{t+1})}[Q_\pi(s_{t+1}, a_{t+1})]]$$

Mục tiêu là tối đa hoá $r(s_t, a_t, s_{t+1})$ trong một môi trường thị trường

chứng khoán và nhóm tiến hành sử dụng phương pháp học tăng cường sâu để giải quyết bài toán trên.

III. Môi trường thị trường chứng khoán:

Nhóm tiến hành sử dụng OpenAI để cài đặt môi trường để giả lập thị trường chứng khoán và huấn luyện mô hình.

1 Không gian trạng thái(state space):

Một vector 181 chiều gồm 7 phần được sử dụng để biểu diễn một trạng thái của môi trường thị trường chứng khoán: $[b_t, p_t, h_t, M_t, R_t, C_t, X_t]$. Mỗi thành phần được định nghĩa như sau:

- $b_t \in \mathbb{R}_+$: Số dư trong tài khoản tại thời điểm t.
- $p_t \in \mathbb{R}_+^{30}$: Giá đóng cửa điều chỉnh (adjusted close price) của mỗi cổ phần.
- $h_t \in \mathbb{Z}_+^{30}$: Số cổ phiếu đang nắm giữ của mỗi cổ phần.
- $M_t \in \mathbb{R}^{30}$: Moving Average Convergence Divergence(MACD)
- $R_t \in \mathbb{R}_+^{30}$: Relative Strength Index (RSI)
- $C_t \in \mathbb{R}_+^{30}$: Commodity Channel Index (CCI)
- $X_t \in \mathbb{R}^{30}$: Average Directional Index (ADX)

2 Không gian hành động:

Với mỗi cổ phần, không gian hành động (action space) được định nghĩa là $\{-k, \ldots, -1, 0, 1, k\}$ với k và $-k$ là số cổ phiếu có thể mua và bán. $k \leq h_{max}$ với h_{max} là siêu tham số cho trước quy định số lượng cổ phiểu tối đa cho mỗi hành động mua. Do đó kích thước của không gian hành động là $(2k+1)^{30}$.

IV. Các thuật toán:

1 Advantage Actor Critic (A2C):

A2C là một thuật toán actor-critic điển hình trong học tăng cường sâu. Thuật toán này được đề xuất để cải thiện các lần cập nhật policy gradients. A2C sử dụng advantage function để giảm phương sai policy gradients trong quá trình học. Thay vì chỉ ước lượng value function, mạng critic sẽ ước lượng advantage function. Do đó, việc đánh giá một hành động không chỉ phụ thuộc vào Q-value của hành động đó mà còn phụ thuộc vào độ tốt của nó so với giá trị kỳ vọng của trạng thái. Do đó phương sai của các lần cập nhật policy gradients cũng giảm giúp cho quá trình học của thuật toán được ổn định hơn.

Hàm mục tiêu của A2C:

$$\bigtriangledown J_\theta(\theta) = \mathbb{E}[\sum_{t=1}^{T} \bigtriangledown_\theta log\pi_\theta(a_t|s_t)A(s_t,a_t)],$$

Trong đó $\pi_\theta(a_t|s_t)$ là mạng policy(mạng actor), $A(s_t,a_t)$ là advantage function:

$$A(s_t,a_t) = Q(s_t,a_t) - V(s_t)$$

hoặc:

$$A(s_t,a_t) = r(s_t,a_t,s_{t+1}) + \gamma V(s_{t+1}) - V(s_t)$$

2 Deep Deterministic Policy Gradient (DDPG):

DDPG là thuật toán kết hợp cả đặc điểm của Q-learning lẫn policy gradient và sử dụng mạng nơron để ước lượng hàm. DDPG học trực tiếp các mẫu thông qua policy gradients. Thuật toán sẽ ánh xạ các trạng thái thành các hành động một cách xác định (deterministic) do đó nó phù hợp cho môi trường có không gian hành động liên tục (continous action space).

DDPG sử dụng replay buffer R để lưu trữ các trạng thái dịch chuyển (s_t,a_t,s_{t+1},r_t) và một tập N trạng thái dịch chuyển trong R sẽ được lấy ra để cập nhật Q-value y_i :

$$y_i = r_i + \gamma Q'(s_{t+1},\mu'(s_{i+1}|\theta^{\mu'},\theta^{Q'})), i = 1,...,N$$

Mạng critic sau đó được cập nhật bằng cách tối thiểu hoá loss function $L(\theta^Q)$ là độ lệch kì vọng giữa hai ra của mạng target Q' và Q:

$$L(\theta^Q) = \mathbb{E}_{s_t,a_t,s_{t+1},r_t \sim buffer}[(y_i - Q(s_t,a_t|\theta^Q))^2]$$

Và policy gradients cho mạng actor định nghĩa như sau:

$$\bigtriangledown J_\theta(\theta) = \mathbb{E}_{s_t}[\bigtriangledown_\theta \mu_\theta(s) \times \bigtriangledown_a Q^\mu(s,a)|_{a=\mu_\theta(s)}]$$

3 Proximal Policy Optimization (PPO):

PPO được đề xuất để kiểm soát việc cập nhật policy gradients và đảm bảo rằng policy mới sẽ không quá khác so với policy trước đó. PPO đơn giản hoá mục tiêu của Trust Region Policy Optimization (TRPO) bằng thêm một "clipping term" vào hàm mục tiêu.

Giả sử tỉ lệ giữa policy mới và policy cũ là:

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$$

Hàm mục tiêu của PPO sẽ là:

$$J^{CLIP}(\theta) = \hat{\mathbb{E}}_t[min(r_t(\theta)\hat{A}(s_t,a_t), clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}(s_t,a_t))]$$

Trong đó $r_t(\theta)\hat{A}(s_t,a_t)$ là mục tiêu của policy gradient thông thường, $\hat{A}(s_t,a_t)$ là ước lượng của advantage function. $clip(r_t(\theta), 1-\epsilon, 1+\epsilon)$ giới hạn $r_t(\theta)$ trong đoạn $[1-\epsilon, 1+\epsilon]$. PPO khiến cho sự thay đổi policy không quá lớn do đó độ ổn định khi huấn luyện cũng được mạng policy cũng được cải thiện.

4 Thuật toán kết hợp:

Thuật toán kết hợp được sử dụng để chọn thuật toán có hiệu suất tốt nhất trong ba thuật toán A2C, DDPG, PPO dựa trên tỉ lệ Sharpe. Thuật toán gồm các bước sau:

- Bước 1: Chọn n để huấn luyện lại ba tác tử của ba thuật toán mỗi n tháng. Ở đây nhóm chọn n là 3.
- Bước 2: Đánh giá cả ba tác tử bằng 3 tháng sau khoảng thời gian huấn luyện để chọn được tác tử có hiệu suất tốt nhất (có tỉ lệ Sharpe cao nhất). Hệ số Sharpe được tính như sau:

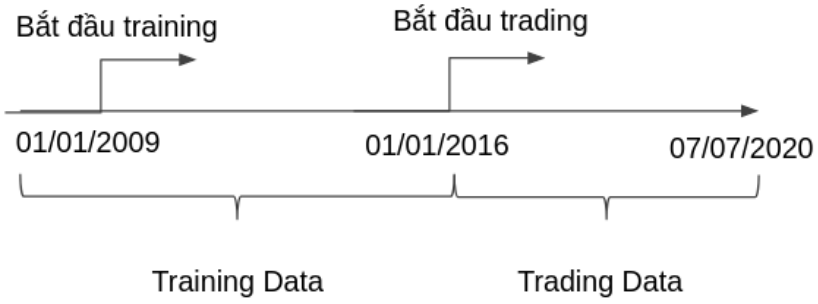
  $$Sharperatio = \frac{\bar{r}_p - r_f}{\sigma_p}$$

  Trong đó $\bar{r}_p$ là tỉ suất lợi nhuận kỳ vọng, r_f là tỉ suất lợi nhuận phi rủi ro và σ_p là độ lệch chuẩn của tỉ suất lợi nhuận vượt quá của danh mục.
- Bước 3: Sau khi chọn được tác tử tốt nhất, sử dụng nó để dự đoán và giao dịch ở quý tiếp theo.

V. Đánh giá hiệu suất

1 Tiền xử lí dữ liệu chứng khoán

Nhóm chọn 30 cổ phần từ Dow Jones bắt đầu từ 01/01/2016 làm nhóm cổ phần để khảo sát giao dịch. Bộ dữ liệu dùng để đánh giá bắt đầu từ 01/01/2009 đến 07/07/2020 của 30 cổ phần Dow Jones. Để tải xuống bộ dữ liệu , nhóm sử dụng trang web Yahoo Finance - trang web cung cấp dữ liệu chứng khoán, tin tức tài chính, báo cáo tài chính,.. Nhóm sử dụng thư viện FinRl để cài đặt các thuật toán trên và tải xuống bộ dữ liệu. Bộ dữ liệu được



Hình 1: Chia bộ dữ liệu

chia như trong hình 1. Dữ liệu từ ngày 01/01/2019 đến 01/01/2016 được dùng để huấn luyện và từ ngày 02/01/2016 đến ngày 07/07/2020 được dùng để đánh giá hiệu suất giao dịch của các thuật toán. Để việc đánh giá dữ liệu giao dịch một cách tốt hơn, nhóm tiếp tục huấn luyện các thuật toán trong các giai đoạn giao dịch, điều này giúp các thuật toán thích ứng tốt hơn khi thị trường giao dịch biến động.

2 So sánh hiệu suất

2.1 Phân tích quá trình huấn luyện của thuật toán kết hợp và các chỉ số dùng để đánh giá thuật toán:

Từ Table 1, chúng ta có thể thấy thẩm định chỉ số Sharpe của DDPG là cao nhất từ ngày 04/01/2016 đến ngày 05/04/2016, vì vậy chúng ta sử

dụng thuật toán DDPG để tính các giao dịch từ ngày 05/04/2016 đến ngày 05/07/2016. Tương tự như vậy, thẩm định chỉ số Sharpe của DDPG là cao nhất từ ngày 01/06/2020, nên chúng ta sử dụng DDPG để tính các giao dịch từ ngày 06/01/2020 đến 07/07/2020. 5 chỉ số được dùng để đánh giá kết quả giao dịch từ các thuật toán:

- Lợi nhuận tích lũy (Cumulative return)- đây là nghĩa tiếng Việt của thuật ngữ Cumulative return - một thuật ngữ được sử dụng trong lĩnh vực kinh doanh, được tính bằng cách trừ đi giá trị cuối cùng của danh mục đầu tư so với giá trị ban đầu của nó và sau đó chia cho giá trị ban đầu, phản ánh lợi nhuận cuối giai đoạn giao dịch của thuật toán.
- Lợi nhuận hàng năm (Annual Return) là lợi nhuận mà một khoản đầu tư được cung cấp trong một khoảng thời gian, được biểu thị bằng tỷ lệ phần trăm hàng năm theo thời gian.
- Biến động hàng năm(Annual volatility) là độ lệch chuẩn hàng năm của lợi tức danh mục đầu tư
- Tỉ lệ Sharpe(Sharpe ratio) là một thước đo xem lợi nhuận thu được là bao nhiêu trên một đơn vị rủi ro khi đầu tư vào một tài sản hay đầu tư theo một chiến lược kinh doanh, là 1 chỉ số được sử dụng rộng rãi để kết hợp lợi nhuận và rủi ro với nhau
- Mức giảm (Max drawdown) là tỷ lệ phần trăm lỗ tối đa trong thời gian giao dịch. Annualized volatility và Max drawdown đảm bảo độ an toàn của thuật toán.

2.2 Phân tích hiệu suất của các thuật toán :

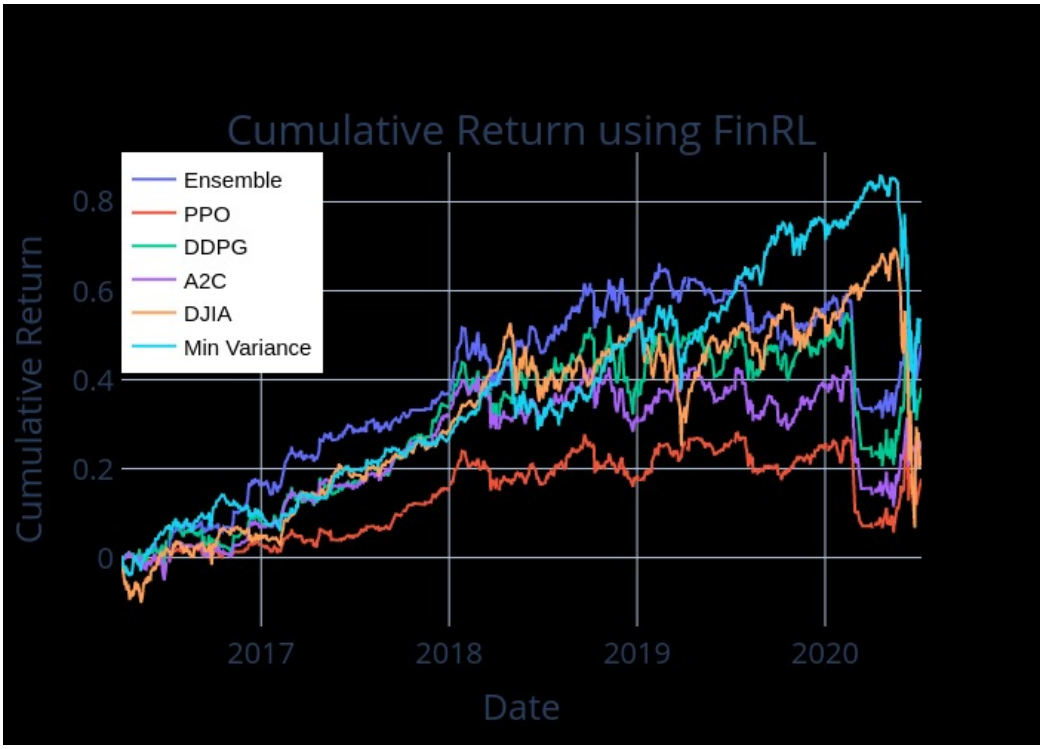
Bảng 1: Chỉ số Sharpe theo thời gian

Val Start	Val End	Model Used	A2C Sharpe	PPO Sharpe	DDPG Sharpe
2016-01-04	2016-04-05	DDPG	0.148703	0.046768	**0.154782**
2016-04-05	2016-07-05	DDPG	0.031038	0.039746	**0.087971**
2016-07-05	2016-10-03	DDPG	-0.053001	-0.053974	**-0.039166**
2016-10-03	2017-01-03	DDPG	0.371398	0.220361	**0.507461**
2017-01-03	2017-04-04	DDPG	0.163701	0.101813	**0.432919**
2017-04-04	2017-07-05	DDPG	0.067129	0.008525	**0.298768**
2017-07-05	2017-10-03	PPO	0.231195	**0.430857**	0.162756
2017-10-03	2018-01-03	DDPG	0.691829	0.637224	**0.771238**
2018-01-03	2018-04-05	DDPG	-0.070878	-0.146269	**-0.042566**
2018-04-05	2018-07-05	A2C	**0.076292**	-0.095003	-0.031633
2018-07-05	2018-10-03	DDPG	0.297151	0.482054	**0.591059**
2018-10-03	2019-01-04	DDPG	-0.202759	-0.302158	**-0.150033**
2019-01-04	2019-04-05	PPO	0.258359	**0.431699**	0.422720
2019-04-05	2019-07-08	A2C	**0.059885**	-0.114435	-0.209631
2019-07-08	2019-10-04	DDPG	-0.238005	-0.346411	**-0.128716**
2019-10-04	2020-01-06	A2C	**0.481637**	0.463694	0.152931
2020-01-06	2020-04-06	DDPG	-0.623757	-0.580020	**-0.517191**

Bảng 2: Bảng đánh giá hiệu suất

	Ensemble	A2C	PPO	DDPG	Min Variance	DJIA
Annual return	0.117502	0.069733	0.048797	0.099840	0.128817	0.098643
Cumulative returns	0.603452	0.331745	0.224445	0.498477	0.742574	0.539651
Annual volatility	0.125333	0.126137	0.108756	0.148477	0.162174	0.204247
Sharpe ratio	0.950294	0.598363	0.493311	0.716214	0.829370	0.563916
Max drawdown	-0.201184	-0.221916	-0.170107	-0.222787	-0.286422	-0.370862

Dựa vào bảng 2 và hình 2, thuật toán PPO thích ứng với rủi ro hơn với chỉ số Annual volatility là 0.108756 và Max drawdown là -0.170107 thấp nhất so với 3 thuật toán. Vì vậy, PPO rất tốt trong việc xử lý thị trường giảm giá. Thuật toán DDPG tạo ra nhiều lợi nhuận hơn so với 2 thuật còn lại với , chỉ số Annual return là 0.099840 và chỉ số Cumulative returns là 0.498477. Thuật toán A2C tương tư như thuật toán PPO nhưng kém hiệu quả hơn , có thể sử dụng A2C khi thị trường tăng giá trở lại. Tiếp theo là thuật toán kết hợp 3 thuật toán trên cho ra kết quả tốt hơn so với 3 thuật toán

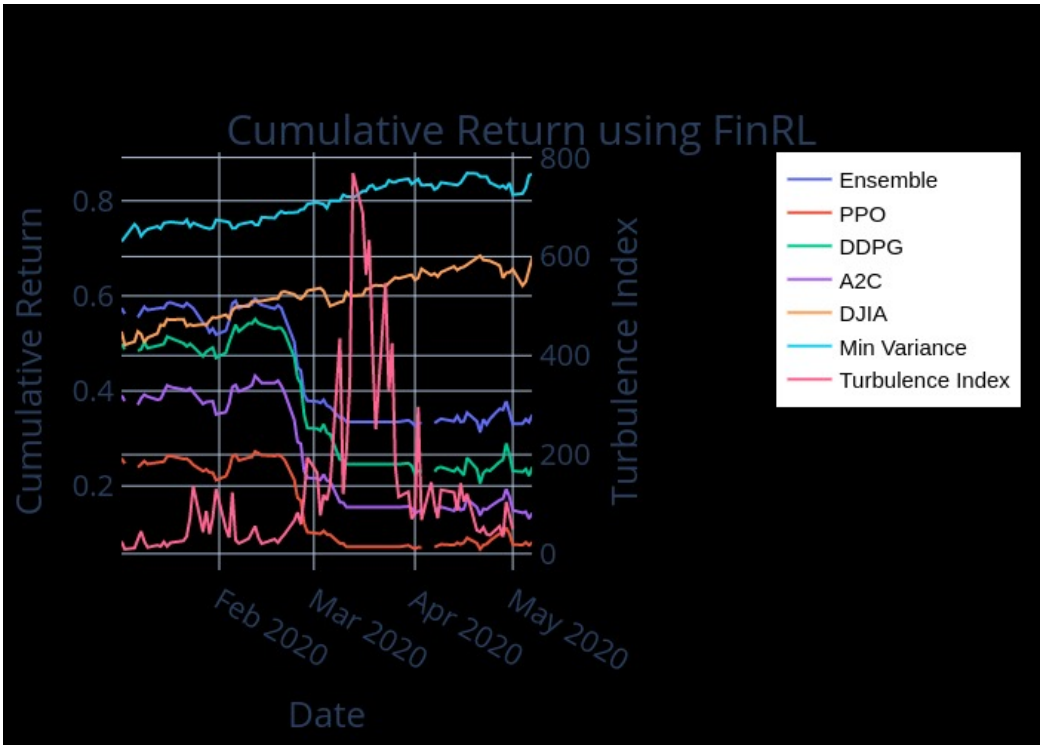


Hình 2: Lợi nhuận tích lũy của thuật toán kết hợp, 3 thuật toán A2C,PPO,DDPG, phương sai (min-variance) của chiến lược phân bổ danh mục đầu tư, chỉ số trung bình của công nghiệp Dow Jones(the Dow Jones Industrial Average) (giá trị danh mục đầu tư ban đầu là 1 triệu đô, từ ngày 05/04/2020 đến ngày 06/07/2020)

A2C,PPO,DDPG với Sharpe ratio cao nhất là 0.950294 và chỉ số Cumulative returns là 0.603452 cao hơn so với 3 thuật toán còn lại. Hiệu suất của 3 thuật toán đều an toàn hơn so với phương sai (min-variance) của chiến lược phân bổ danh mục đầu tư, chỉ số trung bình của công nghiệp Dow Jones(the Dow Jones Industrial Average).

2.3 Hiệu suất của các thuật toán khi thị trường bị sụp đổ vào quý 1 năm 2020:

Dựa vào hình 3, ta thấy 3 thuật toán A2C,PPO DDPG và thuật toán kết hợp 3 thuật toán trên, cho kết quả không được tốt lắm khi thị trường bị sụp đổ. Theo như bài báo của tác giả, kết quả Min variance và DJIA sẽ cho kết



Hình 3: Hiệu suất của các thuật toán khi thị trường sụp đổ vào quý 1 năm 2020:

quả xấu nhưng kết quả nhóm em thực nghiệm cho ra kết quả cao hơn.

3 Kết luận:

Dựa vào hình 2, ta thấy thuật toán kết hợp và 3 thuật toán A2C,PPO,DDPG,Ensemble ổn định hơn so với phương sai (min-variance) của chiến lược phân bổ danh mục đầu tư, chỉ số trung bình của công nghiệp Dow Jones(the Dow Jones Industrial Average) với các chỉ số Annual volatility và Max drawdown mang rủi ro cao trong quá trình giao dịch.