China A-Share Market Portfolio Management Based on Deep Reinforcement Learning

Junwu Zhou, Kai Jing

North China University of Water Resources and Electric Power, Zhengzhou, Henan, China
DOI: 10.32629/memf.v5i4.2573

Abstract: This paper uses the 50 constituent stocks in the Shanghai Composite 50 Index from January 1, 2010 to January 1, 2024 as a data set to study the application of investment portfolios in China's A-share market. The innovations of this paper mainly involve the following aspects. First, this paper introduces five algorithms, A2C, DDPG, SAC, TD3, and PPO, and compares the five algorithms by cumulative yield, maximum drawdown, and Sharpe ratio as evaluation indicators. The comparison shows that compared with the other four algorithms, the PPO algorithm is more in line with the specific situation of China's A-share market. Secondly, this paper collects data from the Shanghai Composite 50 Index and compares it with the reinforcement learning method in terms of cumulative yield. The comparison shows that the deep reinforcement learning method has played a huge role in improving the yield of the Shanghai Composite 50 Index investment portfolio.

***Keywords*:** SSE 50 Index, investment portfolio, A2C, DDPG, SAC, TD3, PPO, cumulative rate of return, maximum drawdown, Omega ratio, Sharpe ratio

1. Introduction

Portfolio management in the A-share market is one of the key issues in the financial field, which involves how to effectively allocate investment funds to maximize returns and control risks. As the financial market becomes increasingly complex, traditional portfolio management methods have become difficult to adapt to the rapidly changing market environment. Therefore, it has become particularly important to explore new investment strategies and methods.

In recent years, the advancement of science and technology has enabled artificial intelligence to be widely used in various fields. In the field of financial technology, deep learning algorithms are often used by domestic and foreign scholars for stock price prediction.[1]Although reinforcement learning is an emerging direction in recent years and its application in the financial field is relatively rare, its basic principles and operating modes are very suitable for decision-making activities in the financial field. Therefore, in existing research, reinforcement learning is often used in quantitative trading and asset portfolio management.[1][8]

In recent years, scholars have conducted extensive research on deep learning and reinforcement learning in the financial market, mainly summarizing their achievements in quantitative trading. Qi Yue et al. (2018) used the deterministic policy gradient DDPG algorithm to build an investment portfolio management model, reducing the overall risk by controlling the investment ratio of each stock. In addition, they also used the Dropout method to effectively avoid overfitting.[9] The research by Fu Feng and Wang Kang (2020) shows that the annual return rate can reach 17.53% by using the reinforcement learning SAC algorithm for financial portfolio management. Wang Wuyu, Zhang Ning, et al. (2021) developed a new intelligent portfolio optimization algorithm that can flexibly improve portfolio results based on the changing market environment and various risk factors, and make corresponding adjustments according to different situations to better meet customer needs.

The innovations of this paper are: (1) introducing five algorithms, namely A2C, DDPG, SAC, TD3, and PPO, and comparing them; (2) collecting data of the Shanghai Stock Exchange 50 Index and comparing them with the reinforcement learning method in terms of cumulative returns.

2. Model building

2.1 Stock Market Definition

The process of stock trading is modeled as a Markov decision process (MDP). MDP is defined as a tuple, where is the state space, A is the action space, P($S_{t+1} | S_t, a_t$) represents the probability of $S_t \in$ S going to the next state in $a_t \in$ A, and r($S_t,, S_{t+1}$) represents the direct reward of taking action in state, and reaching the new state S_{t+1} at the same time. Then, we

formulate the trading objective as a maximization problem:

Action: The action space describes the allowed operations of the agent to interact with the environment. Usually, it includes three actions: a $\in \{-1, 0, 1\}$, in which −1, 0, 1 represent selling, holding, and buying a stock.

Reward: r(s, a, s′) is the incentive mechanism for the agent to learn better actions. The change in portfolio value when taking an action in state s and transitioning to a new state s' is given by the reward function r(s, a, s') = v'-v, where v' and v represent the portfolio values in states s' and s, respectively.

State: The state space describes the observations that the agent receives from the environment. Trading agents observe many different features in order to better learn in an interactive environment.

Environment: SSE 50 Index constituent stocks and SSE 50 Index.

2.2 Algorithm Principle

Since the A2C, DDPG, SAC, TD3, and PPO algorithms all belong to the Actor-Critic network, the specific process of the Actor-Critic network is introduced below.

The Actor-Critic framework is a prominent approach in reinforcement learning, encompassing two integral components: the Critic and the Actor networks.

Critic Network: The Critic serves as the evaluative component, assessing the quality of actions taken under the current policy. It utilizes the temporal difference (TD) error to quantify the discrepancy between the predicted and actual returns. The loss function for the Critic is typically the mean squared error of the TD error, mathematically represented as:

$$L_{Critic} = \mathbb{E}\left[(\delta_t)^2\right] \tag{1}$$

where δ_t denotes the TD error at time step t.

Actor Network: The Actor is the policy network that takes the current state as input and produces a distribution over actions or a deterministic action value. It leverages the feedback from the Critic to refine the policy. The optimization objective of the Actor is to maximize the expected return, given by:

$$J(\theta) = \mathbb{E}\left[R_t \mid \pi_\theta(s,a)\right] \tag{2}$$

Where $\pi_\theta(s,a)$ represents the policy with parameters θ, and R_t is the reward received at time t. The notation θ^* signifies the optimal policy parameters.

The interplay between the Critic and Actor networks is depicted in Figure 1, illustrating the iterative process of policy evaluation and improvement.

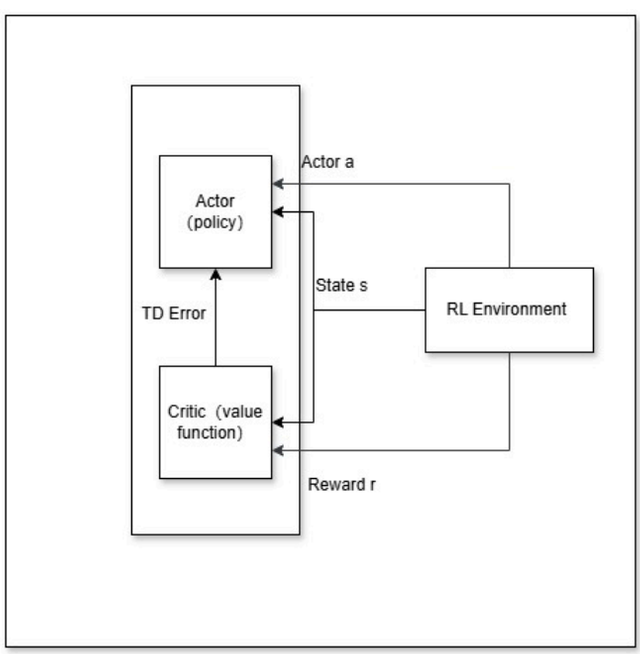

Figure 1. Actor-Critic network structure

3. Experiment procedure

3.1 Data preparation

The dataset is sourced from the Tushare website, comprising trading data of 50 stocks included in the SSE 50 Index in the A-share market. The raw data includes opening price, closing price, highest price, lowest price, and trading volume. Data from January 1, 2012, to December 31, 2021, is designated as in-sample data for training purposes. Data from January 1, 2022, to December 31, 2023, is used for validation and parameter adjustment. During this trading phase, the agent continues training to better adapt to the market conditions.

3.2 Adding technical indicators

For each stock, in order to better meet the market conditions, the following technical indicators are selected, as shown in Table 1.

Table 1. Technical indicators

Indicator name	Indicator significance	Indicator Type
20-period relative strength index (rsi_20)	Indicates the strength of the market's buying and selling forces within a certain period of time. The higher the value, the overbought, and the lower the value, the oversold.	Oscillators
Moving average indicator (macd)	It measures price trend and momentum by taking the difference between two exponential moving averages of different periods (usually 12 and 26 periods).	Trend indicators

3.3 Evaluation indicators

Cumulative return, maximum drawdown and Sharpe ratio are used as evaluation indicators to evaluate the model effect.

3.3.1 Cumulative rate of return

Rp refers to the total rate of return generated since the purchase of stocks to date, measuring the cumulative returns:

$$R_p = \frac{m}{n} \times 100\% \tag{3}$$

Here, m represents the profit, and n represents the principal.

3.3.2 Maximum Drawdown Rate

It refers to the maximum drop from the local high point to the next lowest point in a specified time period. It is used to describe the maximum possible loss of investment and is an important risk indicator. Assuming P_i is the net value of the product on the i-th day, and P_j is the net value of the product on a certain day after P_i, the calculation of the maximum drawdown rate is as follows:

$$Max\_drawdown = \frac{\max\left(P_i - P_j\right)}{p_i} \tag{4}$$

3.3.3 Sharpe Ratio

It refers to the excess return obtained per unit risk. The higher the ratio, the higher the excess return obtained by the strategy per unit risk, so the higher the Sharpe ratio, the better. The calculation of the Sharpe ratio is as follows:

$$Sharpe_ratio = \frac{R_p - R_f}{\sigma_h} \tag{5}$$

4. Experimental results and analysis

4.1 Comparative Experiment

This paper takes the Shanghai Composite 50 Index as the baseline, and then uses five algorithms, A2C, PPO, DDPG, SAC, and TD3, to draw the yield curve of the total return of the investment portfolio based on the baseline, as shown in Figure 2.

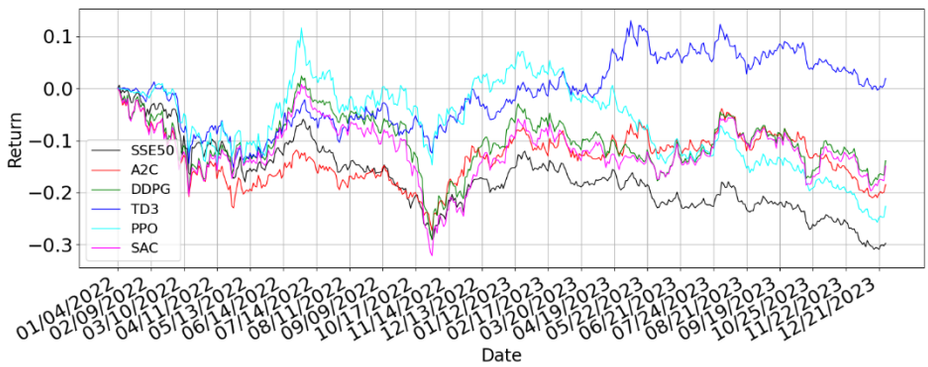

Figure 2. Cumulative rate of return

As shown in Figure 2, different models have different performances in the SSE 50 Index constituent stocks data set. TD3 has the highest total return compared to other algorithms; followed by the PPO algorithm, whose overall return on the portfolio ranks second; the returns of the A2C and DDPG algorithms are similar; among the five algorithms, the SAC algorithm has a relatively small return compared to the other four algorithms, but its return is still significantly improved compared to the baseline.

4.2 Comparison of evaluation indicators

In order to more intuitively demonstrate the superiority of the model with data, the following uses three risk indicators to compare the returns of the investment portfolio. As shown in Table 2, in the comparison of evaluation indicators of different models, the TD3 algorithm has the largest cumulative return rate, reaching 121.26%, which has exceeded the other four algorithms in terms of return. Although the returns of the DDPG and A2C algorithms are slightly lower than the PPO algorithm, at 107.10% and 108.49% respectively, their maximum drawdown rates are lower than the PPO algorithm, which shows that the risk resistance of the PPO strategy is not as good as that of the A2C and DDPG strategies. Next are the SAC algorithms, with cumulative returns of 102.61% respectively. Although these algorithms have relatively low cumulative returns compared to the above four algorithms, their cumulative returns are 58.98% compared to the baseline-Shanghai Composite 50 Index, which is a huge improvement.

Table 2. Result analysis

Performance evaluation indicators	Shanghai Composite 50 Index	A2C	DDPG	PPO	SAC	TD3
Cumulative rate of return	58.98%	108.49%	107.10%	112.85%	102.61%	121.26%
Maximum Drawdown Rate	-18.22%	-21.83%	-20.45%	-30.75%	-20.24%	-15.69%
Sharpe Ratio	1.37	1.23	1.5	1.52	1.54	1.72

5. Conclusion

The data in this article comes from the constituent stocks of the Shanghai Composite 50 Index in the A-shares of listed companies. Taking the Shanghai Composite 50 Index as the baseline, five algorithms, namely A2C, PPO, DDPG, SAC, and TD3, are used to compare the cumulative returns of the investment portfolio. The experimental results show that the investment portfolio calculated by reinforcement learning has better performance, and it also has a huge improvement in terms of risk avoidance and profit acquisition.

References

[1] SHAHI TB, SHRESTHA A, NEUPANE A, et al. Stock price forecasting with deep learning: a comparative study [J]. Mathematics, 2020, 8(9): 1441.
[2] JIY, LIEW WC, YANG L. A novel improved particle swarm optimization with long-short term memory hybrid model for stock indices forecast [J]. IEEE access, 2021(9): 23660-23671.
[3] Chen HQ, Liu YD, Zhou ZT, et al. A2C: Attention-augmented contrastive learning for state representation extraction. Applied Sciences, 2020, 10(17): 5902.
[4] Zhang FJ, Li J, Li Z. A TD3-based multi-agent deep reinforcement learning method in mixed cooperation-competition

environment. Neurocomputing, 2020 (411): 206-215.
[5] Cuschieri N, Vella V, Bajada J. TD3-based ensemble reinforcement learning for financial portfolio optimisation The 31st International Conference on Automated Planning and Scheduling. Guangzhou, China: The International Conference on Automated Planning and Scheduling, 2021: 6-14.
[6] Haarnoja T, Zhou A, Hartikainen K,et al. Soft Actor-Critic Algorithms and Applications. Available from: https://arxiv.org/abs/1812.05905.
[7] Weng Xiaojian, Lin Xudong, Zhao Shuaibin. Stock price rise and fall prediction model based on short-term memory network based on empirical mode decomposition and investor sentiment[J]. Computer Applications, 2022, 42(z2): 296-301.
[8] Liang Tianxin, Yang Xiaoping, Wang Liang, et al. Research and development of financial trading systems based on reinforcement learning[J]. Journal of Software, 2019, 30(3): 20.
[9] Qi Yue, Huang Shuohua. Portfolio management based on deep reinforcement learning DDPG algorithm [J]. Computers and Modernization, 2018 (5): 93-99.
[10] Fu Feng, Wang Kang. Portfolio management based on deep reinforcement learning SAC algorithm [J]. Modern Computer, 2020 (9): 45-48.
[11] Wang Wuyu, Zhang Ning, Fan Dan, et al. Intelligent portfolio optimization based on dynamic trading and risk constraints [J]. Journal of Central University of Finance and Economics, 2021 (9): 32-47.