Comparison of Actor-Critic Reinforcement Learning Models for Formulating Stock Trading Strategy

Chendi Li

School of Mathematics and Physics, Xi'an Jiaotong-Liverpool University, Jiang Su, 215000, China
Chendi.Li22@student.xjtlu.edu.cn

Abstract. The development of a stock trading strategy holds significant importance inside investing organizations. Investors can enhance their trading performance and attain their financial objectives by comprehending market behaviour, creating returns, effectively managing risk, and making well-informed judgments. Nonetheless, formulating a proficient plan within the intricate and perpetually evolving stock market poses a formidable undertaking. Machine learning and deep learning have been widely recognized as useful methodologies for stock trading. This research presents a comparative analysis of four distinct reinforcement learning models. The algorithms under consideration are Deep Deterministic Policy Gradient (DDPG), Proximal Policy Optimization (PPO), Advantage Actor Critic (A2C), and the ensemble method. These models are tested within the context of the actual stock market environment. The findings indicate that the ensemble strategy outperforms the other three algorithms, suggesting that the ensemble model holds promise to improve the effectiveness of stock trading techniques.

Keywords: Reinforcement Learning, Stock Trading Strategy, Deep Learning.

1 Introduction

A strategy for engaging in stock trading can be defined as a systematic framework comprising rules and concepts that guide investors in their decision-making about purchasing and selling stocks within the stock market. The stock trading strategy encompasses three significant decisions: the purchase of stocks, the retention of stocks, and the sale of stocks. The primary objective of stock trading is to generate increased financial returns. The future price of stocks is commonly forecasted based on historical data. These strategies may be derived from various analysis methodologies, market observations, experiential knowledge, and investing objectives to enhance the stock trading performance of investors [1,2].

In the modern financial market, people always employ managers with high-quality experience to make effective investment decisions quickly; however, hiring managers

B. H. Ahmad (ed.), *Proceedings of the 2023 International Conference on Data Science, Advanced Algorithm and Intelligent Computing (DAI 2023)*, Advances in Intelligent Systems Research 180,
https://doi.org/10.2991/978-94-6463-370-2_22

always cost money, and the decisions made by the managers might be wrong. In recent years, numerous methodologies have been devised for forecasting stock trends. Reinforcement learning stands out as the most often favoured approach. There exist multiple fundamental rationales for the utilization of reinforcement learning within the domain of the stock marketplace. Firstly, it is worth noting that obtaining multisource heterogeneous financial data is relatively straightforward. This data encompasses high-frequency trading data, a wide range of technical data, macroeconomic data, industry policy and environment policy, market news, and even data about social media.

Furthermore, there has been a significant advancement in the study of intelligent algorithms. Significant development has been observed in intelligent computer systems, with breakthroughs transitioning from linear models such as the support vector machine and neural network to more advanced deep learning models and strategies for reinforcement learning [3]. They have already been effectively employed in image recognition and natural language processing. In certain scholarly articles, the authors suggest that these sophisticated algorithms can effectively capture the dynamic fluctuations inside the financial market, accurately replicate the stock trading process, and autonomously generate investment judgements. The proliferation of advanced computing equipment, including Graphics Processing Units (GPUs), large-scale servers, and other devices, has significantly contributed to the availability of substantial storage capacity and computational power for the effective utilization of financial big data. The combination of advanced computer hardware, sophisticated algorithms, and extensive financial data has provided decision support for the computerized trading of stocks. This technique has experienced increasing acceptability among professionals in the business. Therefore, the impact of financial technology is reshaping the financial market and modifying the framework of the finance industry. [4,5].

The objective of this study is to provide a comprehensive overview of three reinforcement learning algorithms: Proximal Policy Optimization (PPO), Advantage Actor Critic (A2C), and Deep Deterministic Policy Gradient (DDPG). The author intends to train these algorithms in the stock market setting and then analyze and evaluate the results of the tests conducted on the three algorithms and ensemble approaches. The evaluation will be conducted by considering the return for the risk, which will be quantified utilizing the ratio of risk to return.

2 Method

2.1 Advantage Actor-Critic (A2C)

The A2C algorithm is a widely employed actor-critic algorithm with significant importance inside the ensemble method [6]. The primary objective of this approach is to optimize the policy gradient updates. A2C accomplishes the reduction of policy gradient variance by utilizing an advantage function. In contrast to the exclusive estimation of the value function, the critic network undertakes the estimation of the advantage function, enabling the appraisal of an action to encompass not only its present efficacy but also its capacity for enhancement. This approach mitigates the issue of large variance in the policy network, hence enhancing the model's robustness.

The A2C algorithm utilizes numerous instances of a single agent to compute gradient updates by employing distinct data samples. Each agent functions autonomously in order to engage with identical surroundings. After the completion of gradient computation by each agent in every iteration, a coordinator is utilized to send the average gradients from all agents to a global network. This process enables the coordination of the worldwide network with the actor and critic networks. A global network enhances the diversity of information used for training. The application of synchronized gradient updating has yielded improved cost-effectiveness, quicker performance, and superior outcomes when employed with large batch sizes. The A2C model is highly suitable for stock trading due to its inherent stability.

2.2 Deep Deterministic Policy Gradient (DDPG)

The current methodology involves the utilization of a DDPG to optimize investment returns. The DDPG method can be considered an enhanced version of the Deterministic Policy Gradient (DPG) algorithm, incorporating key principles from both Q-learning and policy gradient [7,8]. The DDPG algorithm distinguishes itself from the DPG approach by including neural networks as function approximators. The development of the DDPG algorithm was motivated by the need to tackle the Markov Decision Process framework within the stock trading domain. The DDPG method differentiates itself from the DQN algorithm through the utilization of policy gradient techniques for the direct acquisition of knowledge from observations instead of relying on Q-value tables for indirect learning.

Furthermore, the DQN approach is impeded by the challenge known as the curse of dimensionality. The primary aim of this methodology is to establish a deterministic correlation between states and actions, with the purpose of better accommodating the continuous action space of the environment. The effectiveness of the DDPG algorithm in addressing continuous action spaces has been demonstrated, making it a feasible alternative for the stock trading domain.

2.3 Proximal Policy Optimization (PPO)

PPO, or Proximal Policy Optimization, is a powerful algorithm that controls policy gradient updates to maintain stability and prevent drastic policy changes. It achieves this by updating the policy network in small steps, ensuring that the updated policy remains close to the original policy [9]. This is accomplished using a surrogate objective function, which measures the similarity between the updated and old policies. By constraining the policy update, PPO effectively prevents large policy deviations that could hinder the learning process. This stability is crucial in domains like stock trading, where maintaining a consistent and controlled policy is important for making informed decisions. In addition to its stability, PPO is known for its speed and simplicity in implementation and tuning. These factors make PPO an attractive choice for stock trading, as it allows for efficient training and optimization without requiring excessive computational resources or complex parameter tuning. Overall, PPO's ability to control

policy updates, stability, and simplicity make it a suitable algorithm for training policy networks in the stock trading domain.

The proposed methodology employs a rolling window approach to select the most proficient trading agent with the highest Sharpe ratio, thereby enhancing risk-adjusted returns. This choice is motivated by the fact that each trading agent exhibits sensitivity to specific market trends. While one agent may excel in bullish markets, it may falter in bearish conditions, while another may be better suited for volatile markets. The Sharpe ratio serves as a metric to gauge an agent's performance relative to the level of investment risk undertaken. Hence, the selection process prioritizes the trading agent that maximizes returns while accounting for the escalating risk associated with market fluctuations.

2.4 Ensemble Strategy

The Ensemble Strategy leverages the strengths of the DDPG, PPO, and A2C algorithms, integrating them into a cohesive framework to create a resilient and adaptable trading agent [10]. By employing a rolling window period of 3 months, the ensemble strategy ensures the inclusion of the most recent agents, enabling the capture of evolving market dynamics.

Subsequent to training each agent using their respective algorithms, performance validation is conducted using a rolling window approach spanning three months. This iterative assessment allows continuous evaluation of the agents' efficacy across diverse market conditions, ultimately enabling the identification of the most proficient agent With the Sharpe statistic being maximized.

The Sharpe statistic is a metric used to measure the risk-adjusted return of an agent, indicating their capacity to generate returns relative to the level of risk they have taken on. The objective is to optimize returns while effectively mitigating the increasing risk inherent in the market by choosing the agent with the highest Sharpe statistic.

Optimal agent, it is deployed to forecast and execute trades for the subsequent quarter. This selection is based on the expectation that the chosen agent's past performance and sensitivity to various market trends will position it well to perform in the prevailing market conditions.

The Ensemble Strategy, through amalgamating the strengths of multiple algorithms and selecting the most proficient agent, seeks to generate consistent and adaptable trading decisions that optimize returns while effectively managing risk within the ever-changing landscape of the stock market.

3 Result

Based on the findings presented in Fig 1, it is evident that all three agents exhibit superior performance compared to both the DIJA and the Min-Variance in the stock market throughout the period spanning from January 2016 to September 2019. While the three actor-critic-based algorithms demonstrate a strong correlation with the fluctuations in the Dow Jones Industrial Average (DIJA), they also exhibit the capacity

to mitigate potential risks in the stock market. Fig 1 illustrates a significant decline in stock prices in 2020, yet the returns generated by the three actor-critic-based algorithms do not experience a substantial decrease. This phenomenon occurs due to the trading policy implemented by the agents, wherein the occurrence of a turbulence index beyond a certain level signifies an exceptional market condition. Consequently, the agents respond by divesting all presently owned shares and adopting a wait-and-observe approach until the market stabilizes, at which point they resume their trading activities.

Fig. 1. Result comparison [10].

4 Discussion

The findings from Table 1 show interesting insights into the performance of different agents in the given market conditions. The A2C agent demonstrates a higher level of adaptability to risk, as evidenced by its lowest annual volatility of 10.4% and maximum drawdown of -10.2% among the three agents. These results suggest that the A2C agent is well-equipped to handle a bearish market.

Table 1. Quantitative result comparison.

(2016/01/04-2020/05/08)	Ensemble	PPO	A2C	DDPG
Cumulative Return	70.4%	83.0%	60.0%	54.8%
Annual Return	13.0%	15.0%	11.4%	10.5%
Annual Volatility	9.7%	13.6%	10.4%	12.3%
Sharp Ratio	1.30	1.10	1.12	0.87
Max Drawdown	-9.7%	-23.7%	-10.2%	-14.8%

On the other hand, the PPO agent excels in following market trends and generates higher returns compared to the other agents. It exhibits the highest annual return of 15.0% and a cumulative return of 83.0% among the three agents. These findings indicate that the PPO agent is particularly suited for a bullish market, where it can capitalize on the upward trends and generate significant returns.

Interestingly, the DDPG agent performs similarly to the PPO agent but falls slightly short in terms of performance. The utilization of this approach may be regarded as a supplementary tactic to the PPO agent during a period of market growth where it can provide additional support and enhance the overall investment strategy.

Overall, results highlight the importance of selecting the right agent based on the specific market conditions. The A2C agent's adaptability to risk makes it a favourable choice in a bearish market, while the PPO agent's ability to follow trends and generate higher returns makes it a preferred option in a bullish market. While the DDPG agent may not exhibit the same degree of robustness compared to the PPO agent, it nevertheless holds potential as a complementary strategy in a market environment.

5 Conclusion

This study explores the feasibility of acquiring stock trading strategies using the utilization of three algorithms. Moreover, in order to effectively respond to dynamic market situations, this work involves the comparison of an ensemble method that autonomously identifies the most proficient agent for trading purposes, employing the Sharpe ratio as the determining criterion. In terms of the Sharpe statistic, the comparative research demonstrates that it achieves higher returns than the three independent algorithms, as well as the min-variance investment approach. This is achieved by effectively managing risk and return while also considering transaction costs. The comparative findings indicate the existence of the ensemble strategy has the potential to be a promising solution towards the development of an efficient stock trading strategy.

References

1. Li, Y., Ni, P., & Chang, V. Application of deep reinforcement learning in stock trading strategies and stock forecasting. Computing, 102(6), 1305-1322 (2020).
2. Fister, D., Mun, J. C., Jagric, V., & Jagric, T. Deep learning for stock market trading: a superior trading strategy? Neural Network World, 3,151-171 (2019).
3. Wu, X., Chen, H., Wang, J., Troiano, L., et al. Adaptive stock trading strategies with deep reinforcement learning methods. Information Sciences, 538, 142-158 (2020).
4. Yong, B. X., Abdul Rahim, M. R., & Abdullah, A. S. A stock market trading system using deep neural network. In Modeling, Design and Simulation of Systems: 17th Asia Simulation Conference, 356-364 (2017).
5. Li, A. W., & Bastos, G. S. Stock market forecasting using deep learning and technical analysis: a systematic review. IEEE access, 8, 185232-185242 (2020).
6. Mnih, V., Badia, A. P., Mirza, M., Graves, A., et al. Asynchronous methods for deep reinforcement learning. In International conference on machine learning, 1928-1937 (2016).
7. Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., et al. Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971 (2015).
8. Xiong, Z., Liu, X. Y., Zhong, S., Yang, H., & Walid, A. Practical deep reinforcement learning approach for stock trading. arXiv preprint arXiv:1811.07522, 1-7 (2018).
9. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017).
10. Yang, H., Liu, X. Y., Zhong, S., & Walid, A. Deep reinforcement learning for automated stock trading: An ensemble strategy. In Proceedings of the first ACM international conference on AI in finance, 1-8 (2020).