Reinforcement Learning in Portfolio Management with Sharpe Ratio Rewarding Based Framework

Zhenqiang Liu

763838923@qq.com

School of Management, New York institute of Technology NY, NY, US, 10044

Abstract—Portfolio management is a financial operation which aims at maximizing the return or optimizing the Sharpe Ratio. One widely used portfolio management strategy, Mean-Variance Optimization, also known as Modern Portfolio Theory, mainly profits by focusing on finding out the expected return and variance of stocks based on historical data to maximize Sharpe Ratio. Yet, it is not easy and accurate to simply predict future return and variance based on a formula. So, in this paper, two Models-free framework, Sharpe Ratio reward based Deep Q-Network (DQN-S) and Return reward (DQN-R) are proposed to overcome the limitations above. Deep Q-learning was employed to train a neural network to manage a stock portfolio of 10 stocks. Stock price was defined as environment of NN, weight of portfolio was defined as action of neural network agent, and reward was indicated to train the model. Traditional portfolio allocation strategy Mean Variance Optimization (MVO) and Naïve Portfolio Allocation (NPA) were also introduced as benchmark to evaluate the performance of reinforcement learning. Moreover, the extensiveness of DQN-S was discussed. The result shows that the MVO is dominating the NPA with a 5% higher annual return and 0.5 higher of Sharpe ratio, although the MDD is slightly higher, indicating the superiority of Sharpe Ratio oriented strategy.

Keywords-Portfolio Management; Deep Q-Network (DQN); Model-free reinforcement learning; Sharpe Ratio; Mean Variance Optimization (MVO)

1. INTRODUCTION

Portfolio management is always a complex asset allocation in the purpose of getting maximum return while maintain a certain risk exposure. During the process of the asset allocation many methods were utilized. The traditional portfolio allocation strategy, Mean Variance Optimization (MVO), is universally applied in industry due to its feature of straightforward and efficiency. However, with the development of machine learning and the computation ability of computer, another path — reinforcement learning is more and more feasible to be applied.

ICEMME 2022, November 18-20, Nanjing, People's Republic of China

DOI 10.4108/eai.18-11-2022.2327121

There are some works implemented the reinforcement learning in portfolio management. It was illustrated that deep reinforcement learning framework can play a significant role in cryptocurrency filed by achieving 4-fold return in 50 days with commission fee [1]. With Convolutional neural Network, Recurrent Neural Network, and Long Short-Term Memory as the algorithms and 30 minutes as the trading period. However, this trading strategy is not suitable for asset which are less active than cryptocurrency market which are much more energetic and volatile. Moreover, in the scenario like this, commission fee will change the outcome significantly, a little tunning of commission fee might upend the algorism and the result.

Blue chip stocks, known as representative of quality, reliability, and ability to profit even in worst scenario, were strong preferred by institutions like pension fund which affects a big population financially. In addition to that, those stocks possess a very different features, for example the different Beta, variance, and expected return in comparison with tech stock, crypto currency, and commodity. However, there were scarce research in this domine. In this paper, the model-free framework based on Reinforcement Learning is presented to optimizing the portfolio management. The deep Q-learning is the core of the framework in the goal of maximizing Sharpe ratio and the return. The traditional portfolio allocation strategy Mean Variance Optimization (MVO) and Naïve Portfolio Allocation (NPA) is also utilized as benchmark for the purpose of comparison.

Although Blue chip companies are important topic affecting a huge population who are rely on the pension fund, nevertheless, it is understudied. A Model-free Reinforcement Leaning framework will enable the strategy automatic development which will outperform the traditional strategy due to the simplicity and fixed formular of traditional strategy. By implementing the framework in the Blue Chip stocks, the efficacy of the model will be tested in the environment of stable stocks. Financial institution like pension fund will benefit from the extra powerful framework. In this paper, the application of reinforcement learning in 10 specific Blue-chip stocks will be implemented to find out the potential of this strategy in the domain of less volatile stock market. In addition, since these 10 stocks are the most popular ones among the Blue-chip stocks, it is typical and representative for the category.

2. REINFORCEMENT LEARNING

Reinforcement learning is a robust machine learning subcategory where agent can learn through the environment they are in and reward from their action. Comparing to the supervised learning and unsupervised learning, the main difference is that RL is learning from interaction. In this environment, the agent does not have any prior knowledge, but establishes his own knowledge according to the reward R, which can be negative or positive.

Reinforcement learning can be classified into two categories, model-based framework and model free framework. Methods for solving reinforcement learning problems that use models and planning are called model-based methods, as opposed to simpler model-free methods that are explicitly trial-and-error learners—viewed as almost the opposite of planning [2]. Model-free learners directly sample the underlying MDP in order to obtain knowledge about unknown models in the form of value function estimates (Q values). These rewards for each relative state are helping agent to decide what will be the best option in specific scenario. Also, there will be

an Epsilon greedy setting to make agent balance the exploring which make agent to be innovative and the exploiting which makes the agent to action optimally.

In Q-learning, the Q values are defined below.

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma max_{a'} Q(s',a') - Q(s,a)] \quad (1)$$

where $\alpha \in [0,1]$ is the learning rate determining how many loss is learned, and $\gamma \in [0,1]$ is the discount factor.

Q-learning is powerful strategy since the reward as feedback will eventually lead the agent to be success, and professional in the game. However, the Q-table which Q-learning rely on can only cope a relative simply environment and limited possibility environment, when the environment is comprehensive enough that Q-table can not present and store all the situation, the strategy will be less accurate than straightforward scenario. Hence, Deep Q-Network (DQN) was introduced by Volodymyr Mnih to tackle the shortage of the Q-learning [3]. They employed the Q-learning in top of convolutional neural network, using the pixel as the input to build a function estimating the maximum future reward. With the neural network, the Q table can be replaced by an approximation function which can be more general in adapting the environment.

3. MEAN VARIANCE OPTIMIZATION

Mean variance optimization is an investment strategy normally used by portfolio manager, and it is trying to find the maximum expect return under the given risk mathematically. The variance of asset is utilized as a proxy for risk [4]. The model postulated that investors are risk aversion, rational investor tend to prefer other portfolio while there is a portfolio exist with same return but lower risk. In other worlds the portfolio makes no sense when other portfolio can provide same return and lower risk.

In MVO, the expected the return is presented below, where E(R) is the expected return, $w_i$is the weighting of asset i, and R_i is the expected return of asset i.

$$E(R) = \sum_i w_i R_i \quad (2)$$

Portfolio variance is presented below, where σ^2 represent the variance of the portfolio, and the ρ_{ij} represent the correlation coefficient between the return on asset I and j. The risk of portfolio or standard deviation is σ.

$$\Sigma^2 = \sum_i w_i^2 \sigma_i^2 + \sum_i \sum_{j \neq i} w_i w_j \sigma_i \sigma_j \rho_{ij} \quad (3)$$

Every possible outcome from the portfolio is within the area of the curve below（as shown in fig.1). The downward curve is dominated by the upward curve called the effective boundary, because the upward curve always has a higher expected return, resulting in the same risk. However, the most efficient point is where the straight line is tangent to the curve. The line between the tangent point and the market risk free point is known as Capital Allocation Line (CAL) which is the most efficient investment portfolio in the market. Any point on the line is

considered having a same optimal Sharpe ratio, hence equally optimized. The difference is that portion of the money invested in the stock. In addition, if the point goes beyond the tangent point, it means that we are shorting risk-free rate (borrowing money in RF) to invest in stocks. Since the short strategy is not considered in the RL model, we will not be conducting the short strategy here for the purpose of comparative. Instead, we will allocate all the capital in the stock mark and none in the risk-free market because the risk is not considered in RL model.

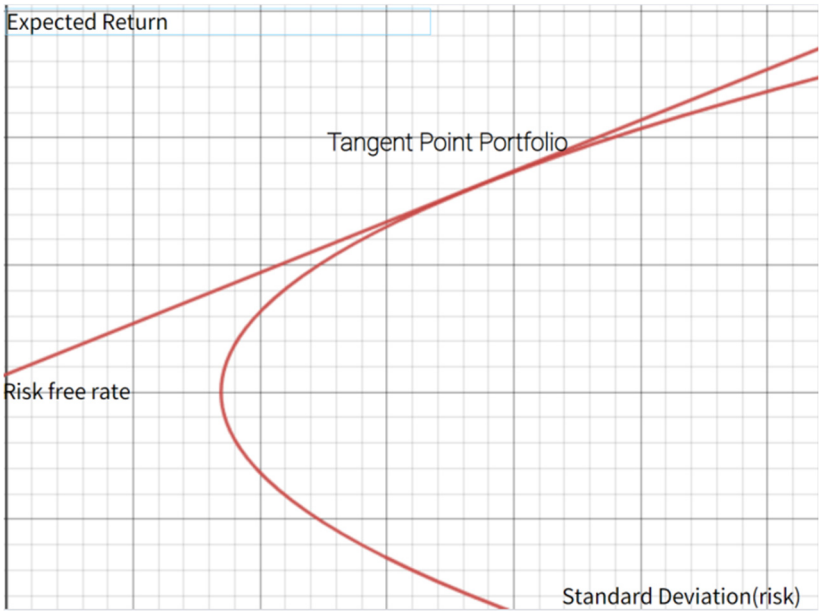

Figure.1 Efficient frontier and the Capital Allocation Line

4. METHODOLOGY

The flowchart presented in figure.2 is the main model and RL architecture of this project. The first part of this process is to clean up and normalize the data. Unlike other DQN models which might have a frank input like images or position states like mountain car project, the input of this model is complexion. The data is normalized into percentage increase or decrease at this step because percentage change is most direct way in evaluating the performance, and it is much more useful in financial domain.

Data set are split into training set and testing set which is used to reflect model performance. In this paper, the dataset dimension is 10* 5* 10* 15. It is constructed by 10 Blue chip stocks with 5 parameters: Open, Close, High, Low price and the volume which illustrated how liquidity of the stocks by showing how frequent the stocks are traded. In every action process, 10 sequential historical date is utilized as environment to make decision. The period length covered 15 years of daily close stock price [5].

The data is split into two datasets. Nov. 7 2006 to Nov. 6 2019 was identified as training dataset and the rest 2 years of data were utilized as testing dataset to verify the performance of the model.

Total return, Sharpe ratio, and maximum draw down (MDD) are referred as the assessment of performance. The total return is defined as profit in these two years divided by the investment. The Sharpe ratio is the difference between return and market risk free rate divided by standard deviation of daily profit. The MDD stands for the maximum loss between the value peak and the lowest point until next peak.

In the RL model, each environment state S is a multidimensional matrix, which contains 500 data input (10× 5×10). The agent make action, the 10 weights of the stock, based on the observation and the neural network. Also, the reward systems are implemented based on the portfolio return and portfolio Sharpe ratio separately. The Sharpe ratio mathematical expression are shown in (4). For the simplicity of the model the Market Risk Free rate is the average risk-free rate in 15 years of time spam. The data from indicated that average RFR will be 1.17% [6]. The Standard Deviation is the represented in (3).

$$\text{Sharpe ratio} = \frac{\text{Annual Return} - \text{Market risk free rate}}{\text{Annual Standard Deviation of porfolio return}} \tag{4}$$

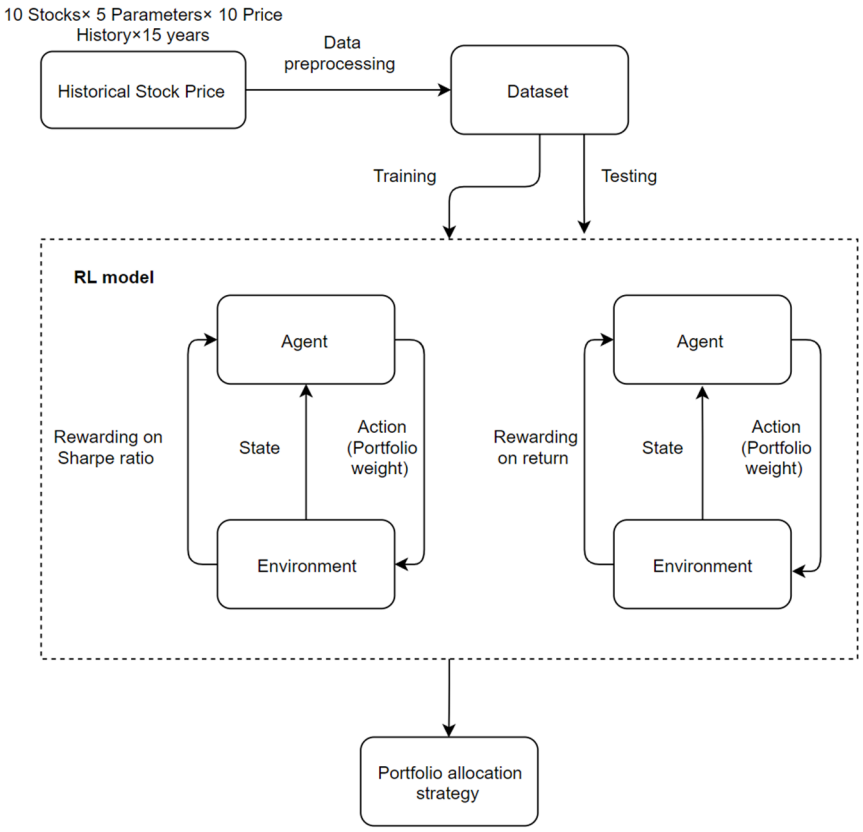

Figure.2 The flowchart of the Deep Q-learning model

4.1 MVO strategy

Mean variance optimization is widely implemented in the field of portfolio management. It is a strong strategy in controlling the risk while maintain a certain profit. In this paper, MVO and Naïve strategy are implemented as the benchmark. The Naïve strategy is simply investing equal portion of money into every stock to form a portfolio and hold the stock until the end of the period.

In this model, the standard deviation and average return of all the stocks were found using normalized data. Then portfolio Covariance Matrix was built by utilizing those foundation. Lastly, the weight can be determined by applying the formular (4) under the constrain of maximizing the Sharpe ratio. Rebalance policy was applied in the principle of making the weight constant, while any portion changes caused by the earning or losing. For the simplicity of the model, and due to the reason that tunning movement was negligible, transaction fee is not considered in this model. The result of the testing set is shown in Figure 3.

Figure. 3 Performance of MVO and Naïve model in testing set

TABLE 1. PERFORMANCE INDICATOR

	Annual return	**Sharpe ratio**	**MDD**
MVO	20.19%	0.700	0.403
Naïve	15.34%	0.650	0.365

5. DISCUSSION

Model-free Reinforcement Learning is promising in the portfolio management, since the algorism of machine learning can not only get all the data we can access, but also perform and form some function that can not be done or even understand by some non-complex mathematics. Moreover, Study by Mu-En Wu demonstrated that Sharpe Ratio based reward CNN and RNN can outperform return based model in 39% in return, their experiment was based on the general stocks in Taiwan [7].

Because the stocks are from same category (Blue chip company), so the feature difference is not significant, for example, the volatility of these stocks are all relatively low comparing to other tech stock or cryptocurrency and the correlations are quite big. Even though MVO was putting goal in maximizing the Sharpe ratio, the boost not significant. However, the annual return of MVO outperformed Naïve method for around 5%. In terms of MDD, MVO was having a higher risk in drawing back, this is probably due to the higher portion of capital invested in the more volatile stocks which draw down a lot during the Covid. One thing noticeable is that both training dataset and testing dataset composited a financial upend, 2008 financial crisis and 2019 Covid, so this model can be illustrative for the real-life situation.

6. CONCLUSION

It is important for financial institute to thrive in portfolio management. It is even more important for financial institute like pension fund to control risk and optimizing return, since the stable and robust of pension fund will simultaneously alleviate the government financial pressure and provide health society. Thus, pension funds are more willing to invest in blue chip companies to make profits within a certain risk range. In the paper, the DQN was employed based on Sharpe Ratio reward and return reward respectively. Tested on the 10 Blue chip stocks, which represent the stable and quality of stock, Sharpe Ratio based strategy can achieve slightly better portfolio value. If the final return of DQN Sharpe based (DQN-S) is outperforming the DQN Return based (DQN-R) and the benchmark of MVO, then it indicates that DQN-S is a suitable framework for the asset allocation. However, if the DQN-R is outperforming the DQN-S, it is an indication that the model-free framework is not suitable in the domain of less active market asset allocation, due to that the Sharpe ratio was well demonstrated to be a slightly stronger strategy.

ACKNOWLEDGMENT

Firstly, I would like to show my deepest gratitude to my teachers and professors in my university who have provided me with valuable guidance in every stage of the writing of this thesis. Further, I would like to thank all my friends and parents for their encouragement and support. Without all their enlightening instruction and impressive kindness, I could not have completed my thesis.

REFERENCE

[1] Jiang, Z. (2017). A Deep Reinforcement Learning Framework for the Financial Portfolio Management Problem. https://arxiv.org/pdf/1706.10059.pdf%EF%BC%89

[2] Sutton, R. S. (2018). Reinforcement Learning: An Introduction 2nd edition. MIT press, PP 26

[3] Mnih, V. Playing Atari with Deep Reinforcement Learning. https://arxiv.org/pdf/1312.5602.pdf?source=postpage

[4] Wigglesworth, R. (2018). “How a volatility virus infected Wall Street”. The Financial Times. https://www.ft.com/content/be68aac6-3d13-11e8-b9f9-de94fa33a81e

[5] Yahoo Finance (2021), Retrieved November 6th, https://finance.yahoo.com

[6] French Kenneth R. 2021, Fama/French 3 Factors, Retrieved November 6th, https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

[7] Wu, M. (2021). Portfolio management system in equity market neutral using reinforcement learning. https://link.springer.com/article/10.1007/s10489-021-02262-0