Transformer Actor-Critic with Regularization: Automated Stock Trading using Reinforcement Learning

Extended Abstract

Namyeong Lee
Hanyang University
Seoul, South Korea
leeny@hanyang.ac.kr

Jun Moon
Hanyang University
Seoul, South Korea
junmoon@hanyang.ac.kr

ABSTRACT
Recently, with the increasing interest in investments in financial stock markets, several methods have been proposed to automatically trade stocks and/or predict future stock prices using machine learning techniques, such as reinforcement learning (RL), LSTM, and transformers. Among them, RL has been applied to manage portfolio assets with a sequence of optimal actions. The most important factor in investing in stocks is the utilization of past stock price data. However, existing RL algorithms applied to stock markets do not consider past stock data when taking optimal actions, as RL is formulated based on the Markov decision process (MDP). To resolve this limitation, we propose Transformer Actor-Critic with Regularization (TACR) using decision transformer to train the model with the correlation of past MDP elements using an attention network. In addition, a critic network is added to improve the performance by updating the parameters based on the evaluation of an action. For an efficient learning method, we train our model using an offline RL algorithm through suboptimal trajectories. To prevent overestimating the value of actions and reduce learning time, we train TACR through a regularization technique with an added behavior cloning term. The experimental results using various stock market data show that TACR performs better than other state-of-the-art methods in terms of the Sharpe ratio and profit.

KEYWORDS
Reinforcement Learning; Sequence Modeling; Portfolio Allocation

We adopt the RL algorithm to automatically manage assets in various portfolio allocation problems by performing optimal actions on a daily basis [8, 15]. However, when an agent takes an action in the current state using RL in [8, 15], historical information cannot be considered due to the inherent Markov property of RL. Although the algorithms in [11, 14] are proposed as a combination of LSTM and transformers with RL, only state is considered among MDP elements, and LSTM has a vanishing gradient problem. Hence, there exists a limit when the model learns a long sequence. Moreover, the transformer is not used as a decision model to predict actions. To address this problem, the decision transformer [1] is a suitable model, which uses the GPT-2 [12]. By combining the GPT-2 with the RL, the GPT-2 is able to predict the current action based on historical MDP elements through the attention mechanism. Thus, when modeling based on a decision transformer in stock investment, it can be trained using historical data.

The decision transformer updates the transformer using the mean squared error (MSE), which is a loss function. Using only MSE, it is difficult to exceed the performance of suboptimal trajectories prepared with training data. Hence, in this paper, we propose TACR (Transformer Actor-Critic with Regularization), a new RL algorithm, which applies the critic to the decision transformer to improve the overall performance of RL using past MDP elements. Additionally, we train the model offline using pre-generated suboptimal trajectories for imitating good actions and reducing learning time. However, when only the critic is applied, the value function is inaccurate, leading to overestimation and reducing overall performance [7], since the actor does not interact with the environment. Hence, we apply the state-of-the-art regularization method [4] to prevent actions from being overvalued by the critic.

TACR guarantees the highest returns compared to other RL and transformer algorithms [2, 5, 6, 8, 13, 15, 16] using various datasets from Dow Jones, US 50, HighTech, NDX, MDAX, and CSI. Our code is available at: https://github.com/VarML/TACR

1 INTRODUCTION
Recently, various machine learning methods such as RL, LSTM, and transformers have been studied for investments in financial stock markets. The main purpose of such studies is to predict future stock prices and/or automatically trade stocks based on specified optimization criteria. In fact, since stock prices exhibit patterns, such algorithms show remarkable performance. Specifically, the RL methods [8, 15], the transformer, and LSTM [3, 10, 16] prove the necessity of considering past stock prices for stock market analysis.

2 TRANSFORMER ACTOR-CRITIC WITH REGULARIZATION
We construct the state space of MDP representing meaningful stock information such as opening/closing/high/low stock prices and technical indicators. The action space is the set of the allocation weight to invest in the $j$th stock for the current period $t$. The reward is scalar and is defined as the sum of the rate of daily returns for each stock $r_{t+1} = r(s_t, a_t) = \sum_{j=1}^{J} a_{jt} r_{jt}$.
where \( \pi \) is a critic network adding a regularization method to improve performance. Through \( L(1) \), the MDP elements are computed by the weights \( a \).

This term causes the transformer actor to follow the distribution of actions in the suboptimal trajectories included in the dataset. \( \lambda \) is a hyperparameter that simultaneously controls maximizing \( Q \) and minimizing the behavior cloning term.

3 EXPERIMENT

We use the same datasets as the state-of-the-art methods [8, 15, 16] for comparison of TACR. Furthermore, to demonstrate the general applicability of TACR, we provide the additional experiment results using NDX, MDAX, and CSI datasets which are the stock indices of the US, Germany, and China, respectively.

Transactions in all datasets are made on a daily basis. We adopt the most commonly used Portfolio value and Sharpe ratio [9] as metrics. We compare the performance of our model with other papers on portfolio allocation [8, 15, 16], and further include offline RL [6], off-policy RL [5], on-policy RL [13], and classic method (Equal weight strategy) as baselines. We construct baselines with other papers, all model-free RL algorithms to show the superiority of TACR, and EW to show whether RL is practical for stock trading application. As a result, Figure 2 shows that TACR has the highest performance compared with various baselines. Furthermore, when comparing the Sharpe ratio with our model and other algorithms, it is also high at least 13.1%, up to 177.7%, excluding the MDAX dataset. Furthermore, when increasing the sequence length \( u \), most datasets show good results. This means that the longer the past MDP elements are considered, the closer to optimal action is taken.

4 CONCLUSION

In stock trading, it is essential to analyze the market using historical data. We propose TACR, which includes the critic network to the decision transformer and the regularization method. In addition, we train our model efficiently using suboptimal trajectories offline. TACR enables the RL method to consider not only past stock data but also previous MDP elements when taking optimal actions, which is different from standard RL techniques formulated based on the MDP framework. Furthermore, TACR prevents the overestimation of the values of the actions. Compared with state-of-the-art methods, TACR shows good performance for various datasets.

Figure 1: Framework of TACR.

Figure 2: Comparison of the profit of TACR with other methods. TACR yields 13% (KDD 21), 1.1% (AAAI 20), 20.7% (NeurIPS Workshop 20), 54.8% (NDX), 274.8% (MDAX), and 54.5% (CSI) higher portfolio values compared with other algorithms.
ACKNOWLEDGMENTS
This work was supported in part by the Technology Innovation Program (20018112) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea), in part by the National Research Foundation of Korea (NRF) Grant funded by the Ministry of Science and ICT, South Korea (NRF-2021R1A2C2094350), and in part by Institute of Information & communications Technology Planning and Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2020-0-01373).

REFERENCES


