# Quantitative Investment Decision Model Based on PPO Algorithm

Xiaochen Xiao*

University of Electronic Science and Technology of China Zhongshan Institute Zhongshan, Guangdong, 528402 China

* Corresponding Author Email: 971851309@qq.com

**Abstract.** Bitcoin is sometimes called the new gold, replacing gold as a hedge against inflation, and research on the relationship between bitcoin and gold has important practical implications. This paper first calculates the correlation between bitcoin and gold. By introducing the calculation of dynamic penalty coefficient, the double stock portfolio investment problem is transformed into the single stock purchase investment problem, which greatly reduces the difficulty of feature engineering and model application. In terms of decision model, deep reinforcement learning (PPO algorithm) is used to make quantitative investment decisions, and the expected data in SLTM is taken as the input data of deep reinforcement learning, which is combined with deep reinforcement learning. Compared with machine learning quantitative investment decisions, after a period of learning, the accuracy rate and returns have been substantially improved.

**Keywords:** Deep reinforcement learning; Quantitative investment, time series analysis, PPO, LSTM.

## 1. Introduction

At present, the situation of low interest rate or even negative interest rate is the problem that the world's major developed economies must face, especially negative interest rate[1]. The total value of bonds with negative interest rates has exceeded $10tn.How to deal with the challenge of negative interest rates is indeed the main problem facing financial institutions[2].How to improve the asset management level of Our country to meet the diversified investment needs of customers is an urgent problem to be solved.Quantitative investment plays an increasingly important role in China's investment and financing system with its regularity[3], systematicness, accuracy, dispersity and timeliness.In this context, it is of great significance to study the quantitative investment scheme of bitcoin and gold.Compared with gold, bitcoin has a high yield, high volatility, free from regulation and taxation, and will have a huge development space in the financial field.Bitcoin is sometimes referred to as the new gold. It can replace gold as a hedge against inflation and become a new safe-haven asset. [4] It can complement gold as a hedge insurance, which is of great interest to investors of all kinds in the financial sector.Therefore, the research on the relationship between bitcoin and gold has important practical significance[5].

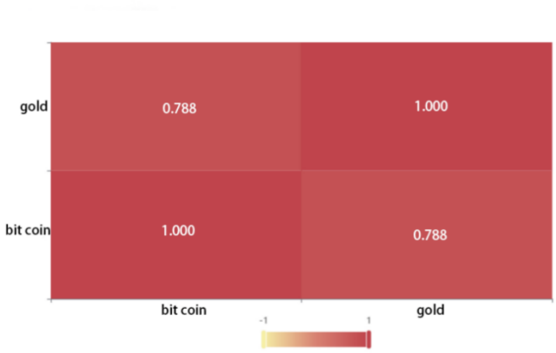
With the continuous development of artificial intelligence[6], the field of AI industry is gradually expanding.At the same time, due to the development of China's financial market and people's increasing demand for financial management, financial investment strategies and financial measurement models have gradually become important tools for people's financial management and investment, which makes the artificial intelligence industry may become the focus of asset investment.Asset appraisal is the inevitable result of market economic development to a certain stage[7]. Asset appraisal of listed companies is helpful to make accurate judgment on investment decisions.The analysis of asset evaluation in artificial intelligence industry is helpful to make up the deficiency of asset evaluation rules and regulations in China[8]. From a macro point of view, the AI industry related to the AI industry related to the AI portfolio analysis of AI concept stocks, can find the current AI industry deficiencies and improve, conducive to the development of the AI industry.Compared with gold, bitcoin has a high yield, high volatility, free from regulation and taxation, and will have a huge development space in the financial field.Bitcoin is sometimes referred to as the new gold. It can replace gold as a hedge against inflation and become a new hedge asset that can complement gold for hedging.It is of great significance to provide a solution for the quantitative trading of bitcoin and gold[8]

## 2. Data collection and analysis

Compared with gold, bitcoin has the characteristics of high yield, high volatility, free from regulation and tax-free collection, so it will have a huge space for development in the financial field.Bitcoin is sometimes referred to as the new gold, a hedge against inflation that can replace gold as a new haven asset that complements gold and is of great interest to investors of all stripes in the financial world.Therefore, the research on the relationship between bitcoin and gold has important practical significance[9].

Correlation analysis: According to the literature, for the correlation between gold and bitcoin, when one side is in the extreme market, the risk of contagion to the other side of the market is strengthened. In the downward market, gold is vulnerable to the influence of the Bitcoin market, while in the upward market, the bitcoin market is more susceptible to the influence of the gold market.That is, at a particular point in time, the correlation is commutative. [10]

According to the statistics of SPSS on the existing data, gold and bitcoin are highly positively correlated in the upward market. The thermal diagram and correlation analysis chart of bitcoin's correlation degree are shown in Figure 1.




correlation

| | | gold | bite |
|---|---|---|---|
| gold | Pearson correlation | 1 | .299** |
| | Sig.(Two-tail) | | .000 |
| | Number of cases | 477 | 477 |
| bite | Pearson correlation | .299** | 1 |
| | Sig.(Two-tail) | .000 | |
| | Number of cases | 477 | 477 |

**.At the 0.01 level (Two-tail), the correlation is significant.

**Figure 1.** Thermal map and correlation analysis chart of bitcoin correlation degree

There is a dynamic relationship between the two stocks.The problem of portfolio investment of these two stocks is transformed into a single stock purchase investment problem, which greatly reduces the difficulty of application of feature engineering and model.The two stocks have a certain dynamic correlation.The difficulty of feature engineering and model application is greatly reduced by transforming the portfolio investment problem of two stocks into the buying investment problem of single stock.The generalization ability of the model is improved from the point of view of data.According to the fluctuation of gold price, the corresponding events are found:

September 29, 2017 -- The 820-kilometer Lanzhou-Chongqing Railway (Lanzhou, Chongqing and Gansu) was put into operation.

In 2017, China's Sichuan province had two hydropower stations: Xiluodu Hydropower Station (Sichuan, with an installed capacity of 13,860 MW, the third in China) and Xiangjiaba Hydropower Station (Sichuan, with an installed capacity of 7.75 million kW, the sixth in the world).

A large number of mining machines have sprung up relying on hydropower stations.A month later, bitcoin saw a big rally (lower mining costs).In 2020, the annual outbreak of the novel coronavirus pneumonia caused many human economies to stall and triggered panic and volatility in global financial markets.Vaccination requires social distancing and changes people's lifestyles and habits.All kinds of contactless digital ways are accelerating and gaining popularity, such as online education and mobile payment.Digital currencies such as Bitcoin reached a new peak at the end of the year due to the interaction of the novel coronavirus pneumonia and the digital economy.

COVID-19 outbreak: More than 80 million confirmed cases worldwide December 27.Policy Implications:

On May 18, 2021, the Sanhui Issued the Notice on Preventing speculative Risks in Virtual currency Transactions.

On May 19, 2021, the price of bitcoin plunged 40% to a low of $30,000.

About penalty coefficient:

One is the punishment for the gold incident, and the other is the punishment for the international situation.

The penalty for a gold event is an event that moves gold.Positive event is set to 1, invalid event is set to 0.7 international penalty, positive event is set to 1, negative event is set to 0.5.

When the punishment for the gold event * the punishment for the international situation equals 1, it is determined to go up.At that point, if the gold price stabilizes, all the bitcoins will be bought.

When it is not equal to 1, it is identified as a declining market.At this point, we buy gold.

## 3. Model Overview

Artificial Neural Networks (ANNs) are also called Neural Networks (NNs) or connected models.It is an algorithmic mathematical model to imitate the behavior characteristics of animal neural network for distributed parallel information processing.This kind of network relies on the complexity of the system and processes information by adjusting the interconnections between a large number of internal nodes.Neural network structure is shown in Figure 2.The neural network structure is shown in Figure 3.

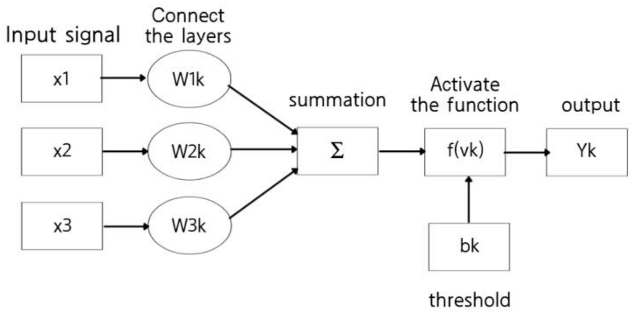



**Figure 2.** Neural network structure

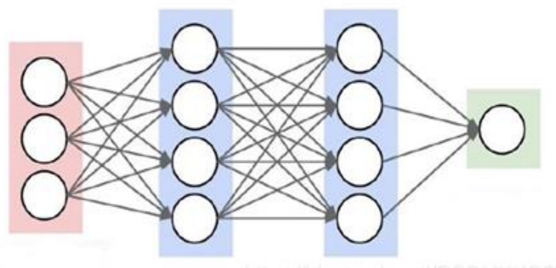


**Figure 3.** Neural network structure diagram

## 4. Deep reinforcement learning PPO algorithm

PPO algorithm thought:

PPO algorithm is a new strategy gradient algorithm.The strategy gradient algorithm is very sensitive to the step size, but it is difficult to choose the appropriate step size.If there is too much difference between new and old strategies in the training process, it is not conducive to learning.PPO proposes a new objective function, which can be updated in small batches in multiple training steps, and solves the problem that the step size is difficult to determine in the strategy gradient algorithm.In fact, TRPO also aims to solve this problem, but the PPO algorithm is easier to solve than the TRPO algorithm.

Policy Gradient Review:

Review the strategy gradient algorithm again.Policy gradients do not use error backpropagation.It selects a behavior by observing information and transmits it directly back.Of course, it is surprising that there are no mistakes.Instead, rewards are used to directly increase and decrease the likelihood of choosing behavior, with good behavior increasing the probability of being selected next time, and bad behavior decreasing the probability of being selected next time.

Strategy τ's revenue forecast:

$$\nabla R = E_{\tau p_\theta}(\tau)\left[R(\tau)\nabla \log p_\theta(\tau)\right] \tag{1}$$

Importance sampling:

$$E_{x\sim p}[f(x)] = E_{x\sim q}\left[f(x)\frac{p(x)}{q(x)}\right] \tag{2}$$

Weight update of PPO algorithm:

$$R = \sum_{t=1}^{T} r_t \tag{3}$$

Path:

$$\tau = \{s_1, a_1, s_2, a_2, \cdots, s_T, a_T\} \tag{4}$$

Each S (state) corresponds to an A (action), and then to A reward (corresponding result)

The purpose of the whole function is that each step is accompanied by a reward, so we ask this question when the total reward is at its maximum

Now let's get this straightened out.How does the payoff for each step come about?Nature is what we do with each step.

The behavior of generating action is determined by theweight parameters in the whole neural network.So our approach is to optimize one of our network models so that it gives the right answer in every state.

Input data generation method:

$$p_\theta\left(a_t \mid s_t\right) \tag{5}$$

The operation record of the model is shown as follows:

$$\underbrace{p_\theta\left(\mathbf{s}_1, \mathbf{a}_1, \ldots, \mathbf{s}_T, \mathbf{a}_T\right)}_{p_\theta(\tau)} = p\left(\mathbf{s}_1\right)\prod_{t=1}^{T}\pi_\theta\left(\mathbf{a}_t \mid \mathbf{s}_t\right) p\left(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t\right) \tag{6}$$

The action record of the model is represented as follows: Objective analysis to be completed

First, we need to be clear about who we're targeting.Our approach is a neural network model.Therefore, we only need to specify the inputs and outputs.

$$\theta^\star = \arg\max_\theta E_{\tau\sim p_\theta(\tau)}\left[\sum_t r\left(\mathbf{s}_t, \mathbf{a}_t\right)\right] \tag{7}$$

The function is a gradient rise problem, that is, for the purpose of maximum reward. Why do I use expectation to express it? Because there is too much randomness in the process of this series, which is also a process of exploration.For example, as long as you reach the finish line, the reward is 100 points, so going around a corner straight is 100 points, if he a solution

If it's a detour, then his future solutions will be a detour.

$$\theta^{\star}=\arg\max_{\theta}\underbrace{E_{\tau\sim p_{\theta}(\tau)}\left[\sum_{t}r\left(\mathbf{s}_{t},\mathbf{a}_{t}\right)\right]}_{J(\theta)} \tag{8}$$

According to the law of large numbers, all solutions are exhaustive:

$$J(\theta)=E_{\tau\sim p_{\theta}(\tau)}\left[\sum_{t}r\left(\mathbf{s}_{t},\mathbf{a}_{t}\right)\right]\approx\frac{1}{N}\sum_{i}\sum_{t}r\left(\mathbf{s}_{i,t},\mathbf{a}_{i,t}\right) \tag{9}$$

πθ (τ) denotes the probability of the current sequence, and R (τ) denotes the reward:

$$J(\theta)=E_{\tau\sim\pi_{\theta}(\tau)}[\underbrace{r(\tau)}]=\int\pi_{\theta}(\tau)r(\tau)d\tau \tag{10}$$

Calculate slope:

$$\nabla_{\theta}J(\theta)=\int\nabla_{\theta}\pi_{\theta}(\tau)r(\tau)d\tau=\int\pi_{\theta}(\tau)\nabla_{\theta}\log\pi_{\theta}(\tau)r(\tau)d\tau=E_{\tau\sim\pi_{\theta}(\tau)}\left[\nabla_{\theta}\log\pi_{\theta}(\tau)r(\tau)\right] \tag{11}$$

And then expand according to the law of large numbers.The law of large numbers tells us that under certain conditions, all laws of large numbers are:

$$\frac{1}{n}\sum_{i=1}^{n}x_{i}\Longrightarrow E\left(\frac{1}{n}\sum_{i=1}^{n}x_{i}\right) \tag{12}$$

The convergence of random variables to a number means that the law of large numbers reflects the stability of the mean of random variables (the mean tends to be stable in the case of large samples). Generally speaking, this theorem means that the frequency of random events is close to its probability under the condition that the test remains unchanged.There is a certain inevitability in chance.

$$\nabla_{\theta}J(\theta)\approx\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\nabla_{\theta}\log\pi_{\theta}\left(\mathbf{a}_{i,t}\mid\mathbf{s}_{i,t}\right)\right)\left(\sum_{t=1}^{T}r\left(\mathbf{s}_{i,t},\mathbf{a}_{i,t}\right)\right) \tag{13}$$

Therefore, we get the expected value of probability in the selection, which forms a virtuous cycle and a network training process. The training process of PPO algorithm is shown in Figure 4.

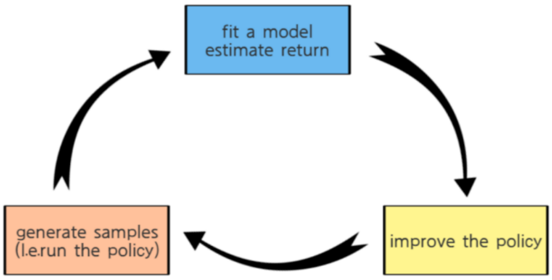



**Figure 4.** PPO algorithm training process

Instead, just keep training and plugging in the data.Then there's the problem that the total reward -- the total amount earned in a game -- looks like a weight term

$$R\left(\tau^{n}\right)-b\quad b=\frac{1}{N}\sum_{i=1}^{N}r(\tau) \tag{14}$$

We want to complete training with rewards and penalties, but some games may only have rewards, so we can do a de-mean operation on a total reward:

$$\nabla_{\theta}J(\theta)=E_{\tau\sim\pi_{\theta}(\tau)}\left[\nabla_{\theta}\log\pi_{\theta}(\tau)r(\tau)\right] \tag{15}$$

After removing the mean, it is symmetric with 0 as the origin and 0 as the center.One is positive, the other is negative.In this way, the effect of reward and punishment is clear.

OnPolicy and OffPolicy:

OnPolicy and Officy Policies OnPolicy is the training data obtained by the Agent's continuous interaction with the environment, while Officy is the result of the interaction with the environment

$$E_{x\sim P}[f(x)] = \int_x P(x) f(x) d_x \approx \frac{1}{N} \sum_{x_{i'}\sim P, i-1}^{N} f(x_i) \tag{16}$$

In the gradient strategy, we need to constantly generate sample numbers:

$$E_{x\sim P}[f(x)] = \int_x P(x) f(x) d_x = E_{x\sim P'}\left[\frac{P}{P'} f(x)\right] \approx \frac{1}{N} \sum_{x_i\sim P', i=1}^{N} \frac{P(x)}{P(x)'} f(x) \tag{17}$$

Sampling X continuously from the distribution of P, changing X to f of X, and finding the desired.F of X.F (X) p(X) DX=f (X) p(X))Q (X) DX, where p(X) can be used.

As a weight term:

Importance sampling:

$$D_{\mathrm{KL}}\left(\pi_{\theta'} \| \pi_\theta\right) = E_{\pi_{\theta'}}\left[\log \pi_\theta - \log \pi_{\theta'}\right] \tag{18}$$

The data are from p samples for t training θθ2e. In practice, they predict the results as much as possible through the network. We can directly take the previous parameter iteration of the training model:

$$\begin{gathered} \operatorname{clip}\left(\frac{p_\theta\left(a_t \mid s_t\right)}{p_{\theta^k}\left(a_t \mid s_t\right)}, 1-\varepsilon, 1+\varepsilon\right) \quad (Restrictions in PPO2 version) \\ \nabla_{\theta'} J\left(\theta'\right) = E_{\tau\sim\pi_\theta(\tau)}\left[\frac{\pi_{\theta'}(\tau)}{\pi_\theta(\tau)} \nabla_{\theta'} \log \pi_{\theta'}(\tau) r(\tau)\right] \end{gathered} \tag{19}$$

LSTM is used to predict the future trend of bitcoin and gold, and then these data are put into PPO algorithm as input data for enhanced learning.Lstm-pbo algorithm implementation flow chart is shown in Figure 5.

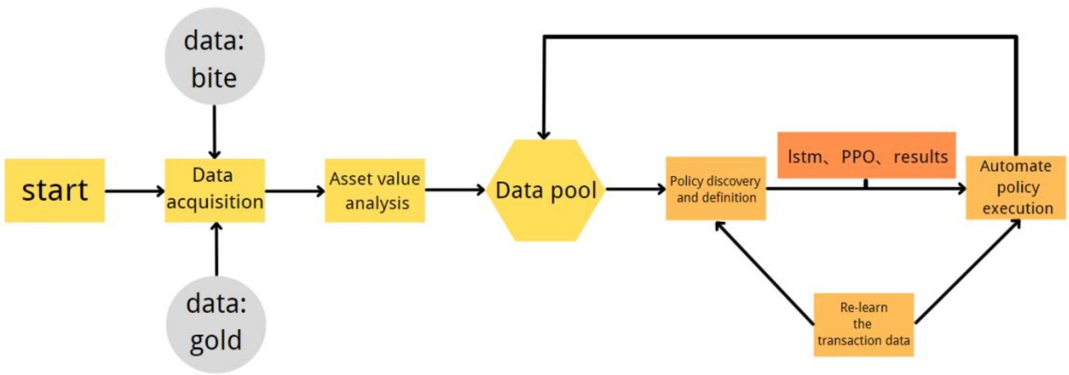



**Figure 5.** Flow chart of LSTM-PBO algorithm implementation

## 5. LSTM algorithm implementation

We put the data of Bitcoin and bitcoin time series into the LSTM algorithm, and its prediction accuracy reached 91% : The bitcoin prediction graph based on LSTM is shown in Figure 6. The gold prediction chart based on LSTM is shown in Figure 7.

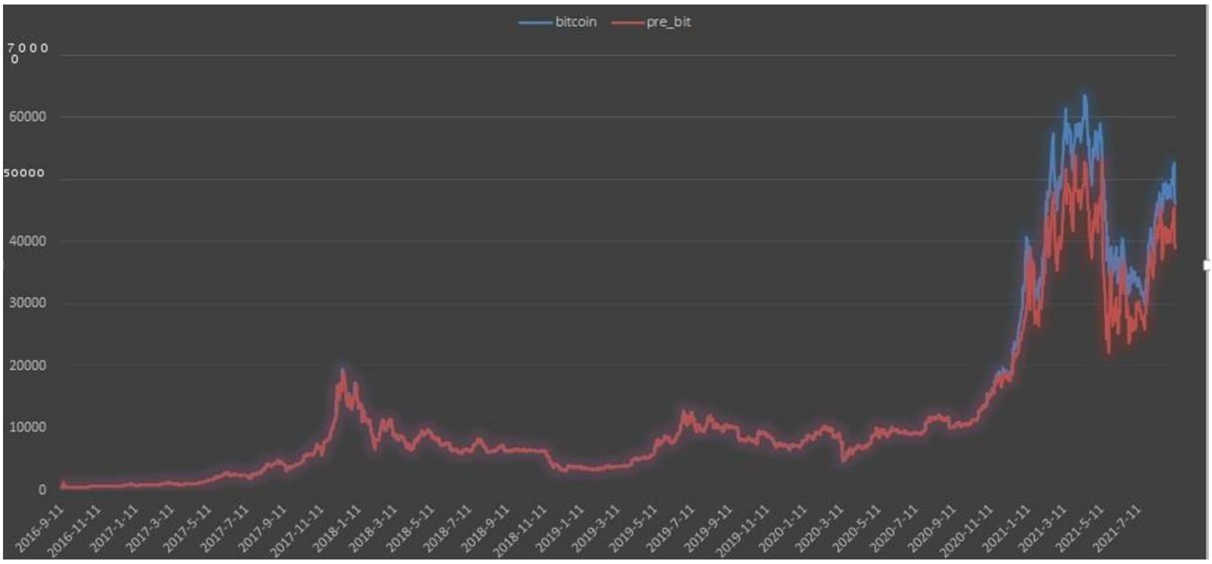



**Figure 6.** LSTM- Bitcoin forecast chart

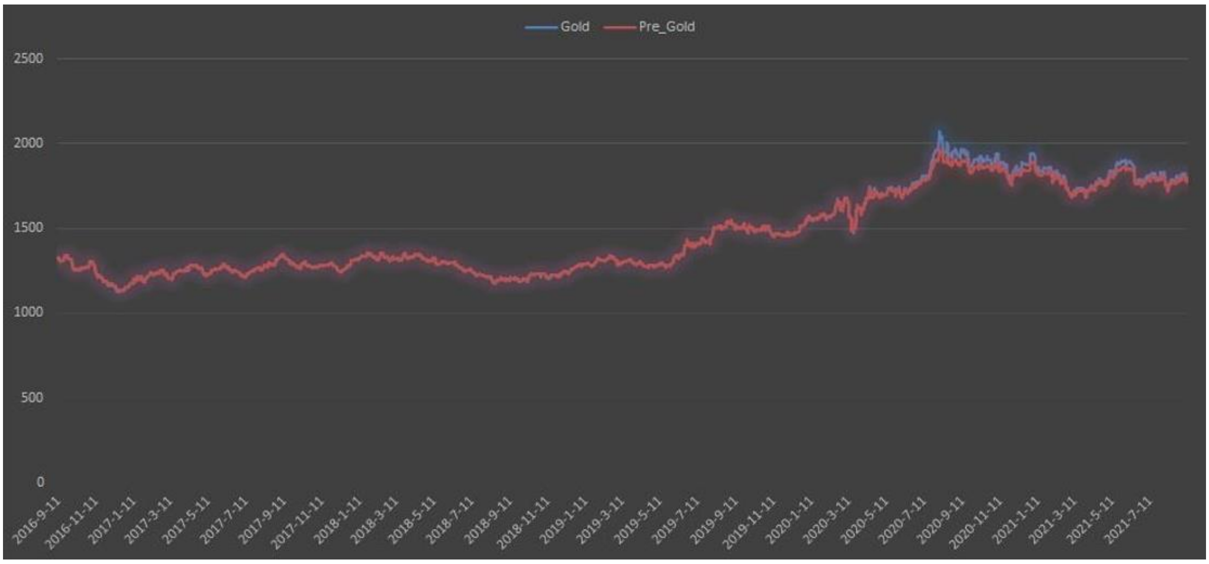



**Figure 7.** LSTM- Gold Forecast chart

The design of reward function is very important to the goal of reinforcement learning. In the context of stock trading, what we should care about most is current profits, so we use current profits as a reward function.

$$Reward = SELf.net\_WORTH - Initial\_account\_balance \; Reward = 1 \; If \; reward > 0 \; otherwise = -100$$

In order to make the network learn profit strategy more quickly, when the profit is negative, the network is given a large penalty (-100).At the same time, the introduction of penalty coefficient λ (when the punishment of the gold event * to the international situation of the punishment is equal to 1, determined as the upward market.If the gold price stabilizes, buy all bitcoin).In our deep reinforcement learning decision, the income of bitcoin and gold is as follows, and the result of our model is: total amount 157852.6585. Bitcoin trend is shown in Figure 8. Gold trend is shown in Figure 8.

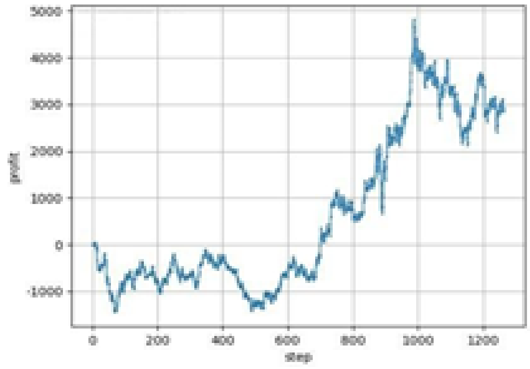



**Figure 8**. bitcoin trend

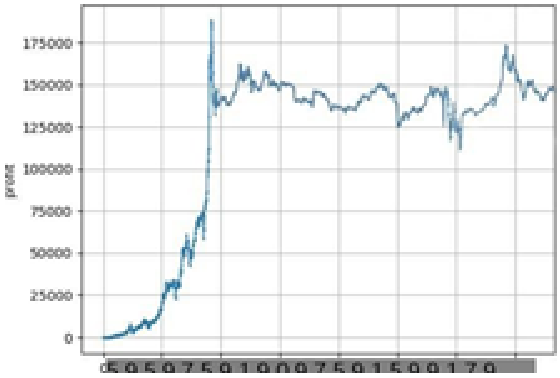



**Figure 9.** gold trend

## 6. Advantage analysis of the model

This paper evaluates the performance of the model by calculating the accuracy and loss rate of the data set of price changes of bitcoin and gold.

Compared with the traditional quantitative investment, the PPO+ LSTM model proposed in this paper is superior to the common neural network model.PPO+ LSTM had better predictive performance (Acc=0.95) and higher accuracy.

**Table 1**. Model advantage analysis table

| Model | Acc% | Loss% |
|---|---|---|
| PPO+lstm | 0.9562 | 0.2939 |
| Traditional neural network | 0.8524 | 0.6548 |

## 7. Conclusion

For the correlation between gold and bitcoin, there are: when one side is in the extreme market, the risk of contagion to the other side of the market is strengthened; when the market is on the downside, gold is vulnerable to the influence of the bitcoin market; when the market is on the upside, the bitcoin market is more vulnerable to the influence of the gold market.That is, at a particular point in time, the correlation is commutative (it may be positive or weak positive).Given bitcoin's high returns, we should buy bitcoin at the right time to get the most out of it.Then we need a penalty coefficient $\lambda$ to judge this time point. Bitcoin and gold are regarded as two stocks with a certain dynamic correlation.From the perspective of data, the generalization ability of the model is improved by transforming the portfolio investment problem of two stocks into the purchase investment problem of a single stock.Existing quantitative investment strategies do not require much transformation.About the penalty coefficient one is the punishment for the gold event, one is the punishment for the international situation.The penalty for a gold event is an event that moves gold.Positive event is set to 1, negative event is set to 0.7 international penalty, positive event is set to 1, negative event is set to 0.5 When the penalty for gold event is equal to 1, it is determined to go up.At that point, if the gold price stabilizes, all the bitcoins will be bought.Therefore, with sufficient data support, the effect of deep reinforcement learning is better than that of traditional methods.We propose to introduce deep reinforcement learning into quantitative investment decision making.

## References

[1] Zhou X. The application of Bren band trend breaking strategy in digital currency market [D]. Zhejiang industry and commerce university, 2021. DOI: 10.27462 /, dc nki. GHZHC. 2021.000281.

[2] Su Chunlin. (2019). Application research of multi-factor Model in digital currency Market (Master thesis, University of Electronic Science and Technology of China) [3] Huang Zepeng.(2017). Gold Price Prediction Modeling and Application based on Deep Learning (Master thesis, Shanghai Jiao Tong University) [4] Zhang Jing.(2013). Application of multiple models to quantitative investment. Analysis of stock market dynamics (32),87.

[3] Cao Yi.Design and Implementation of multi-factor quantitative Investment Management System [D].Hunan University, 2014

[4] Baek, C., & Elbeck, M. (2015). Bitcoins as an investment or speculative vehicle? a first look. Applied Economics Letters, 22(1-3), 30-34.

[5] Yu Liwei & NingLing. (2016). The stock market of quantitative investment strategies and empirically. Statistics and decision (6), 145-149. The doi: 10.13546 / j.carol carroll nki tjyjc. 2016.06.038.

[6] Guo, L. (2017). Research on the application of cross-sectional multi-factor quantitative model in csi 300 Index investment (Master thesis, Beijing Jiaotong University).

[7] Chen Lili. (2017). Quantitative Analysis of stock based on BP neural Network (Master's thesis, Zhejiang University of Technology)

[8] Meng Ye, Yu Zhongqing & Zhou Qiang, (2018). A safe quantitative stock selection strategy: Evidence from a-share market. Financial Theory and Practice (08),102-107.