Currency Investment Strategy Based on State-Action-Reward-State-Action

Penglan Liu*, Xuyu Hou

School of Electronic and Information Engineering, Liaoning Technical University, Huludao, 125105, Liaoning, China
**Corresponding author*

Abstract: *The investment products emerge in the market to earn the difference of profit in the space of currency appreciation. As the bull market peaked, bitcoin prices gradually fell. As a new thing, the lack of the substantial economic support, and the price fluctuations are inevitable. To avoid risk whenever possible, we propose an online reinforcement learning model-Sarsa (State-Action-Reward-State-Action). The model has the characteristics of online learning, sample small batch data from the past. At the same time, it also has the characteristics of intensive learning, which can try to explore in the past experiences, so as to learn to avoid risks and improve the ability of investment efficiency. we compare the our Sarsa model with the latest-based LSTM portfolio model. This experiment proves that the model has minimum regret value and small sensitivity to the proportion of transaction commission.*

Keywords: *Currency Investment; Portfolio Optimization; Reinforcement Learning; Transaction Commission*

1. Introduction

At present, the novel coronavirus is raging around the world, causing a huge blow to the development of the international financial markets. In 2021, the repeated outbreak has caused more serious situation after the shock of the international financial markets. As various countries issue money and print money, the purchasing power of the money fell, making the currency depreciate. Therefore, it also has an impact on gold prices and Bitcoin. The main purpose of market traders in financial investment is to gain benefits. Due to various factors, however, the level of its income and payment methods are also different.

There are two following main categories: the fixed income investment: it is the due income of some financial assets purchased by investors, with a fixed rate of return. This approach requires regular or due payments, and is fixed throughout the investment period, such as time deposits, bonds, etc. Such an investment is generally less risky.

Non-fixed income investment: it is the due income of certain financial assets purchased by investors, without fixed return rate. It will not necessarily be paid on schedule, such as stocks.

This kind of investment is generally risky, but it also has greater profit opportunities and higher returns. Therefore, in order to obtain higher profit opportunities and more returns, some market traders are willing to undertake certain investment risks for non-fixed income investment, such as stocks. However, stock prices will fluctuate with the exchange rate, the macro economy, the general environment, policies and the others, and there are certain risks. Therefore, in the process of buying and selling stocks, we need to make decisions to reduce the investment risks and maximize the total returns as much as possible.

When we use cash to purchase and sell gold and Bitcoin, the prices of gold and bitcoin will fluctuate with the exchange rate, the macro economy, the general environment, and the policy. Thus, there are some unknown risks in this approach. To give market traders a strategy to getting as many total returns as possible, we need to Based on the data, provided the best daily trade strategy by the statistical and descriptive analysis of gold and bitcoin price fluctuations in recent years to remove outliers.

2. Data process

2.1. Gold and Bitcoin price fluctuation background

Gold has long been an investment tool. High value, and is an independent resource, not limited to any country or trade market. The gold market operates similarly to the other investment markets and to the stock markets. The gold market trades at 100 ounces per volume, quoted at dollars. Gold prices have low volatility, people often reserve gold during inflation, and reserve gold can cause short supply, and prices rise. Gold is at the same market risk as all other currencies and commodities in the market. But the “scarcity” of gold determines that the general trend of past, present, and future gold prices will never change.

Bitcoin was presented on P2P foundation on November 1,2008, by Satoshi Nakamoto. Unlike most currencies, Bitcoin does not rely on specific monetary institutions, it according to a specific algorithm, through a large number of calculation, bitcoin economy using the entire P2P network node distributed database to confirm and record all transaction behavior, and use decentralized features and algorithm to ensure that not through a large number of manufacturing Bitcoin to artificially manipulate the currency value. The emergence of Bitcoin indicates the gradual collapse of the traditional financial system. Due to the attractive early performance of bitcoin, the number of new registered users and user activity have increased significantly in recent years, and the upward trend is obvious, indicating that more and more people are being exposed to bitcoin. The total number of Bitcoins is limited, and the total number will be permanently limited to 21 million, causing great price fluctuations, making bitcoin more suitable for speculation than anonymous trading. On November 10, 2021, Bitcoin approached $69,000 for the first time. In January 2022, bitcoin fell below $42,000.

2.2. Data analysis for the files

The data were analyzed and processed. Bitcoin is up and down sharply, choosing short-term trading as much as possible. Gold fluctuates less and is appropriate to use long-term trading. And bitcoin’s highest gain was much higher than gold prices.

Table 1: Data analysis of files

	DataSize	Range	Minimum	Maximum	AverageValue	Variance
GOLD	1265	941.45	1125.7	2067.15	1464.549	294.2918
BITCOIN	1826	62960.36	594.08	63554.44	12206.07	14043.89

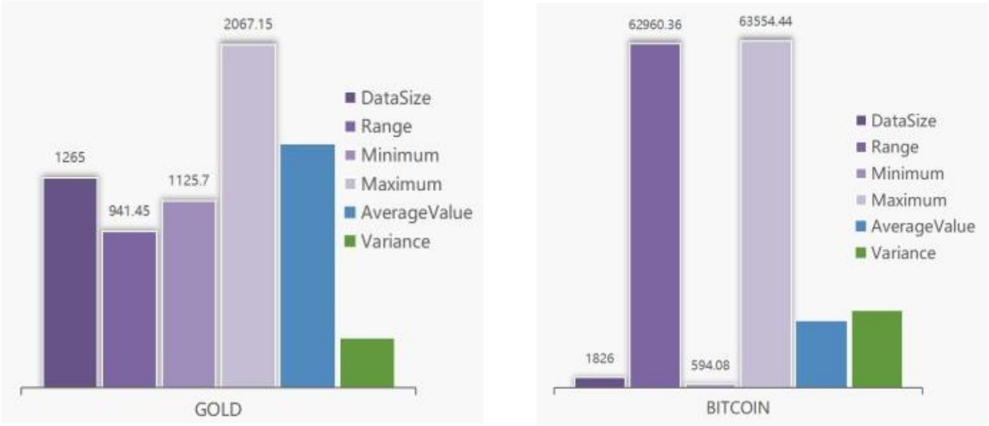


Figure 1: Gold and Bitcoin dataset analysis

3. Model build

Online learning and traditional statistical learning differ in the following ways:

(1) The sample is presented in an ordered manner, rather than in an unordered batch manner;

(2) We need to consider the worst case rather than the average situation, because we need to ensure that we are always in control of things during the learning process;

(3) The goals of learning are also different, online learning attempts to minimize regret values, while statistical learning needs to reduce empirical risk [1].

Through the analysis of our data and topics, our data and models should have the following characteristics:

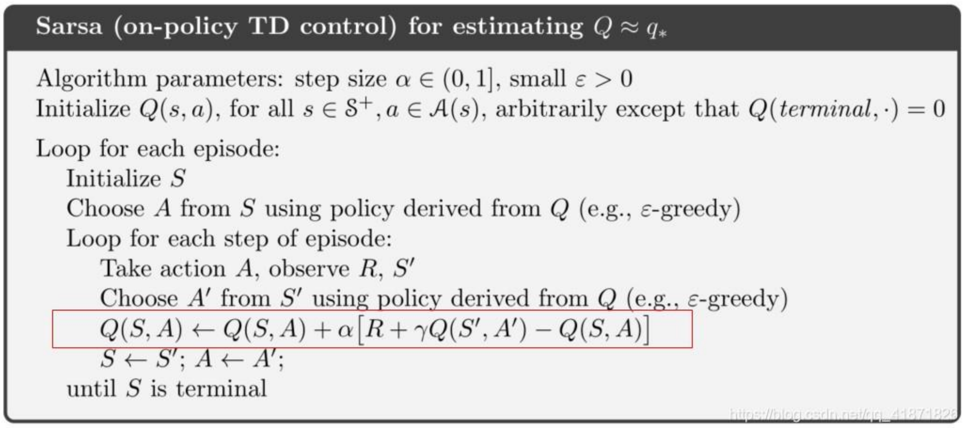
(1) Process time series data, in line with the characteristics of online learning1;

(2) We can only use the data before the trading day as the basis for the trading strategy of that day, which is in line with the characteristics of online learning2;

(3) Our trading cannot predict the future, cannot always try to get the highest profit, but minimize the trading regret value, in line with the characteristics of online learning3.

Secondly, the traders have only three trading actions per day: buy, sell, and take a position. Therefore, in line with the characteristics of online reinforcement learning models, we looked for the simplest online reinforcement learning algorithms Sarsa to build our models.

Sarsa is an algorithm that learns Markov's decision-making process strategies, often applied in the field of machine learning and reinforcement learning. It was introduced by Rummery and Niranjan in Modified Connectionist Q-Learning (MCQL) [2]. The pseudocode is as follows:



```
Sarsa (on-policy TD control) for estimating $Q \approx q_*$

Algorithm parameters: step size $\alpha \in (0,1]$, small $\varepsilon > 0$
Initialize $Q(s,a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(terminal, \cdot) = 0$

Loop for each episode:
   Initialize $S$
   Choose $A$ from $S$ using policy derived from $Q$ (e.g., $\varepsilon$-greedy)
   Loop for each step of episode:
      Take action $A$, observe $R$, $S'$
      Choose $A'$ from $S'$ using policy derived from $Q$ (e.g., $\varepsilon$-greedy)
      $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma Q(S',A') - Q(S,A)\big]$
      $S \leftarrow S'$; $A \leftarrow A'$;
   until $S$ is terminal
```

Figure 2: The pseudocode of Q-learning model

In the **Sarsa** algorithm, I need to initialize a table *Q*, the columns represent each state *S*, and the rows represent the action *A* performed by each state. Each action performed gives back a reward R for the current state. When we reach a new state, we attach a probabilityϵ to it. We select the maximum action of the last state in the table *Q* with the probabilityϵ. Randomly select an action as the current action with the probability $1-\epsilon$. The table *Q* is updated as shown in the red box. Among them, is the learning rate. γ is the extent to which the future would affect current decisions. A table value that represents a state execution action. $Q(S,A)$ indicates the value which the state *S* performs the action.

Next, we use the Sarsa to define the core variables in our model. They are as follows: We set each day to a state, assuming it is $S(i)$.

We try to consider as much as possible the situations we can encounter when trading, so we can choose from 7 types of each state:

A1: 50% of the current cash is used to buy gold and the other is used to buy Bitcoin.

A2: The total current value of cash and Bitcoin is used to buy gold.

A3: The total current cash and gold value is used to purchase Bitcoin.

A4: Sell all current gold.

A5: Sell all current Bitcoins.

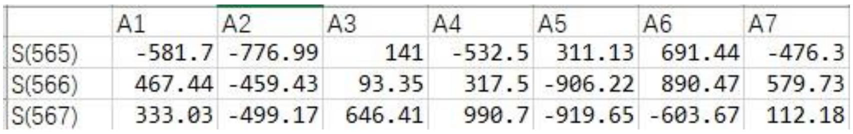
A6: Sell all current gold and Bitcoin.

A7: no buy and no sell.

On the one day, when the state is $s(i)$, the gold exchange rate is $g(i)$, the bitcoin exchange rate is $b(i)$, the number of gold held is $G(i)$, the number of bitcoin is $B(i)$, and the number of dollars is $D(i)$. On the second day, when the state is $s(i+1)$, the gold exchange rate is $g(i+1)$, the bitcoin exchange rate is $b(i+1)$, the amount of gold held is *G(i+1)*, the number of Bitcoin is *B(i+1)*, and the dollar number is *D (i+1)*. Next, we set a commission rate for gold of 1and a commission rate for bitcoin of 2. Meanwhile, we define the Reward for executing each Action.

When the state shifts, the goods ($D(i)$, $G(i)$, $B(i)$) will change with the choice of action. We use the above formula to update the status and the reward, and then update the table Q in the **Sarsa** to achieve the purpose of building the model. The table is as follows.

Table 2: Computation table



	A1	A2	A3	A4	A5	A6	A7
S(565)	-581.7	-776.99	141	-532.5	311.13	691.44	-476.3
S(566)	467.44	-459.43	93.35	317.5	-906.22	890.47	579.73
S(567)	333.03	-499.17	646.41	990.7	-919.65	-603.67	112.18

In the data given in the question, the Sarsa is used to obtain the number of three currencies in the final state:($D(i)$, $G(i)$, $B(i)$) according to the formula:

$Dollar_{sum} = D(i) + G(i) * g(i) * (1 -\beta1) + B(i) * b(i) * (1 -\beta2)$ Finally, the initial \$1,000 grew to \$41,250.

4. Model evaluation

To validate the feasibility of the online reinforcement learning model based Sarsa model, we contrast the effects of the current predictive investment strategy-based LSTM model.

LSTM is a temporal neural network to solve the gradient disappearance problem which is often encountered in practical problems. In other words, it indicates the problem of the decreased perception of the later input point to the previous input point. It adds to the concept of "input gate", "output gate". The network structure is shown below.

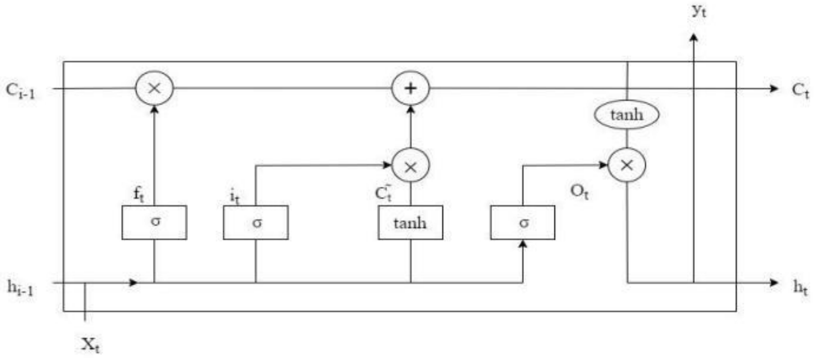


Figure 3: The network structure of LSTM

The LSTM model employs “forget gate” to determine which information needs to be discarded. “Forget gate” will read h_{t-1} and x_t, calculate by the $sigmo$d function, "forget gate" f_t:

$$f_t = \sigma(W_f * [h_{t-1}, x_t] + b_f)$$

"The Output gate” determines what value is updated,:"the output gate” i_t:

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i)$$

Finally, a new vector, $\tilde{c_t}$, was determined. It will be added to the state.

$$\tilde{c} = tanh(W_c * [h_{t-1}, x_t] + b_c)$$

Since then, we use the old state C_{t-1} to updateC$_t$, multiply the old state C_{t-1} with $f t$and add to $i_t * \tilde{c}_t$.. The last result is the new candidate value.

$$C_t = f_t + i_t * \tilde{c}_t$$

Then, we use another sigmod function to process the current state value.

$$o_t = (W_o * [h_{t-1}, x_t] + b_o)$$

W_i', W_f, W_c, W_o are the weight vectors. b_f, b_c, b_σ are the corresponding deviation vectors.

But the predictive model alone is not enough. LSTM can only predict the currency growth trend for some time to come, and still cannot give a specific trading strategy. So we adopt the classic $Dual - Thrust$ strategy.

The core idea is that we define an interval. The upper and lower bounds of the interval are the support and resistance lines respectively. When the price exceeds the upper bound, if you hold short positions, first flat, and then open more; if there is no position, directly open more. When prices fall below the lower bound, if you hold multiple positions, then close your positions first, and then open your short positions; if there is no position, directly open short positions.

To verify the effectiveness of the **Sarsa** model, we trained the based **LSTM** and based $Sarsa$ models first on the dataset given in the question. Then, using the money market uncertainty and certain continuity, a random set of dataset is similar to the data format given in the question [4]. Observe the change in the daily total assets, as shown in the figure below.

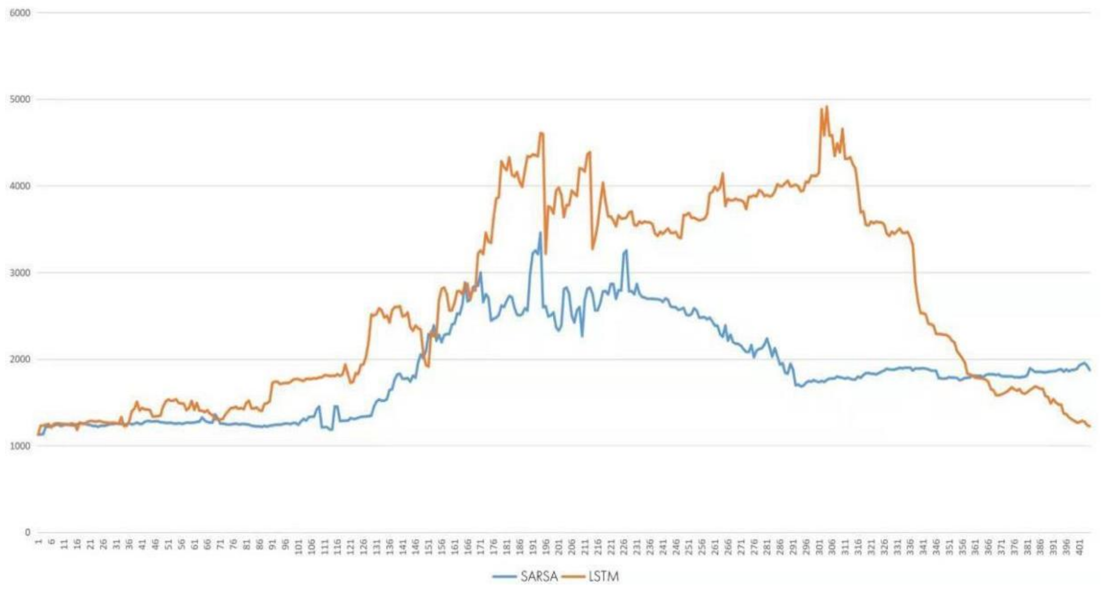


*Figure 4: **Sarsa** model and **LSTM** model data comparison*

*Table 3: **Sarsa** model and **LSTM** model comparative data analysis*

	Minimum	Maximum	Variance	average
LSTM	1126	4869.57	162..82	2416.3
SARSA	1126	3517.40	7119.25	1793.1

From the data and icon above, the variance of the LSTM model is larger and more unstable than the SARSA model. Despite the high revenue value, it is also at risk of bankruptcy at any time. Comparing their average assets, there is not much difference. Applied to the real currency trading market, the stable income increase brought by the SARSA model is the investment plan which is more favored by investors.

5. Model testing and optimization

In terms of decision planning, when establishing the optimization model, we look for appropriate decision variables, reasonably simplify the constraints, and non-convex functions have countless extreme points in the feasible domain. In order to jump out of the local optimal solution and find global optimal solutions, we study heuristic search algorithms.

Consider simulated annealing algorithm for planning model to speed up convergence, simulated annealing algorithm relatively, no constraints on decision variables, both continuous or discrete variables can solve, and jump out of the local optimal solution ability is very good, easy to find the global optimal solution, its disadvantage is only single-threaded work, cannot expand large search, when the decision variable dimension is high, the algorithm convergence speed is very slow.

Simulation annealing algorithm relatively speaking, no constraints on decision variable type, whether continuous or discrete variables can solve, and jump out of the local optimal solution ability is very good, easy to find the global optimal solution, its disadvantage is only single threaded work, cannot expand a large search, when the decision variable dimension is high, the algorithm convergence speed is very slow.

Particle Swarm Optimization particle swarm optimization algorithm (PSO) originated from the study of flock predation behavior. The basic core of the particle group optimization algorithm is to use the sharing of information by individuals in the population, thus allowing the movement of the whole population to produce a disordered to orderly evolution in the problem solution space to obtain the optimal solution of the problem. Particle group algorithm is mostly used for the optimization problem where decision variables are continuous variables, which has fast convergence, but its ability to jump out of the local optimal solution is relatively weak.

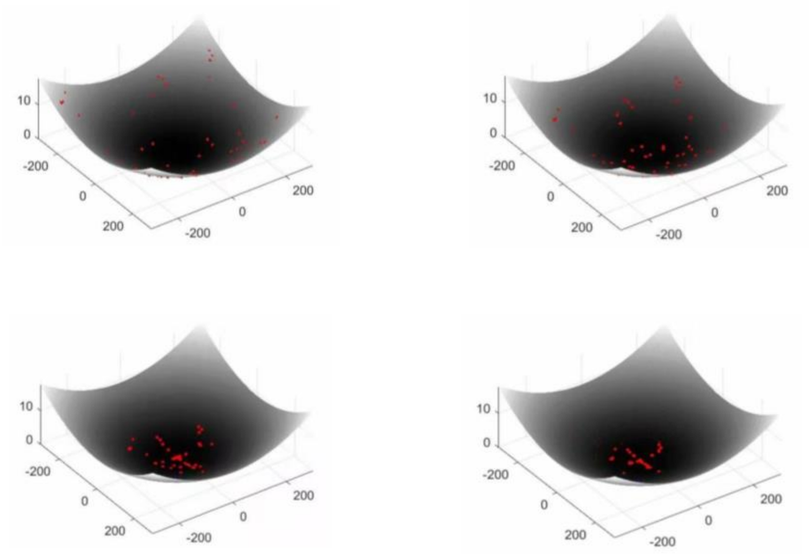


Figure 5: Optimize the model diagram

6. Conclusion

Our Sarsa-based online reinforcement learning model solves the most basic risk control in the monetary investment market, as well as the portfolio problems. However, due to time and data limitations. Because there are many factors affecting the currency value space in the trading market, there are also many types of currencies that can be bought. This leads to we cannot artificially customize trading behavior. If the stock market has more buyers than the seller, the stock market is called a bull market, the stock market sells more than the buyers, the stock market bearish is called a bear market. The bull markets and bear markets of the stock market are the traditional economic cycle indicators. As investors, only by identifying the bull market and the bear market of the stock market and mastering the long-term development direction of the market can we avoid the impact of short-term fluctuations in the market and be affected by the complex information of the market.

References

[1] W. Z. Song. Digital money portfolio strategy research-based on deep intensive learning methods. Nanjing University of Information Engineering, 2019.
[2] Q. Deng, S. Z. Chen, B. Hu, et. al. A study on Markov decision process in different wireless networks. Journal of Communications, 2020, 31(12): 25-36.
[3] S. J. Luo. Deep learning-based gold futures price prediction. Lanzhou University, 2020.
[4] H. Di, X. J. Zhao, Z. L. Zhang. Research on commodity futures investment strategy based on LSTM-Adaboost model. Southern Finance, 2021, (08): 62-76.
[5] J. C. Hu. Research on production schedul system based on learning effect. Shandong University, 2021.
[6] Y. Wang. Application of particle group algorithm in network structure damage recognition. Beijing University of Architecture, 2020.