


PAPER • OPEN ACCESS

## A reinforcement learning model based on reward correction for quantitative stock selection

To cite this article: Haibo Chen *et al* 2020 *IOP Conf. Ser.: Mater. Sci. Eng.* **768** 072036

View the [article online](#) for updates and enhancements.



**EXTENDED ABSTRACT DEADLINE: DECEMBER 18, 2020**

**239th ECS Meeting**  
with the 18th International Meeting on Chemical Sensors (IMCS)

**May 30-June 3, 2021**

**SUBMIT NOW →**

# A reinforcement learning model based on reward correction for quantitative stock selection

Haibo Chen<sup>1</sup>, Chenyu Zhang<sup>2,\*</sup> and Yunke Li<sup>1</sup>

<sup>1</sup>School of Science, Zhejiang Sci Tech University, Hangzhou, China

<sup>2</sup>Essence Information Technology Co.Ltd, Building 4, No276, Huanghai Road, Tianjin, China

\*Corresponding author e-mail: itouchzcy@163.com

**Abstract.** A novel reinforcement learning model based on reward function correction is proposed to quantify stock selection. In this model, hidden markov model is used to predict the trend of the future state and the reward function of reinforcement learning is corrected according to the prediction, which makes it close to the long-term average gradually. Using the method, the prediction accuracy rate of HS300 in China A-market is 76.6%, Sharp ratio is 1.08%, and the average profit rate can reach 32.2%

## 1. Introduction

The application of reinforcement learning in quantitative finance has attracted many researchers' attention in recent years. Instead of predicting the price of future securities, it uses action value function to determine quantitative strategy[1]. [2] proposed a long short reinforcement method, which uses a continuous strategy to find stocks that exceed the market average.[3] proposed an RL framework in which Ensemble of Identical Independent Topology is taken as the kernel. [4] proposed a hybrid DL and RL model in which the DL senses the dynamic market condition, and the RL makes decisions in an unknown environment.[5] used PG learning in China's stock market and show that PG is more desirable in financial market than DDPG and PPO.

All the above methods based on reinforcement learning have a common key point is the iterative method of value function, whose rationality directly affects the actual effect of Markov decision-making. According to this characteristic, we propose a new iterative correction method of value function. Based on the hidden Markov chain, the EM algorithm is used to predict the state of the stock market, which reflects the trend of the stock value in a certain period. Furthermore, the prediction is used to correct the reward function of reinforcement learning, so that the reward function tends to be reasonable gradually.

The paper is organized as follows. Section 2 defines the process of hidden Markov state prediction. Section 3 introduces the value function correction method of reinforcement learning, and the backtest in China stock A market is carried and results are discussed in Section 4. Section 5 concludes with directions for future research.

## 2. State prediction of hidden Markov model

HMM model is a kind of time-series probability model. Discrete random variables are used to describe the process of generating an unobservable state sequence randomly from a hidden Markov chain[6], and then generating a random observation sequence from each state. The HMM model is defined as follows:



$$\lambda = (H, O, \pi, A, B) \quad (1)$$

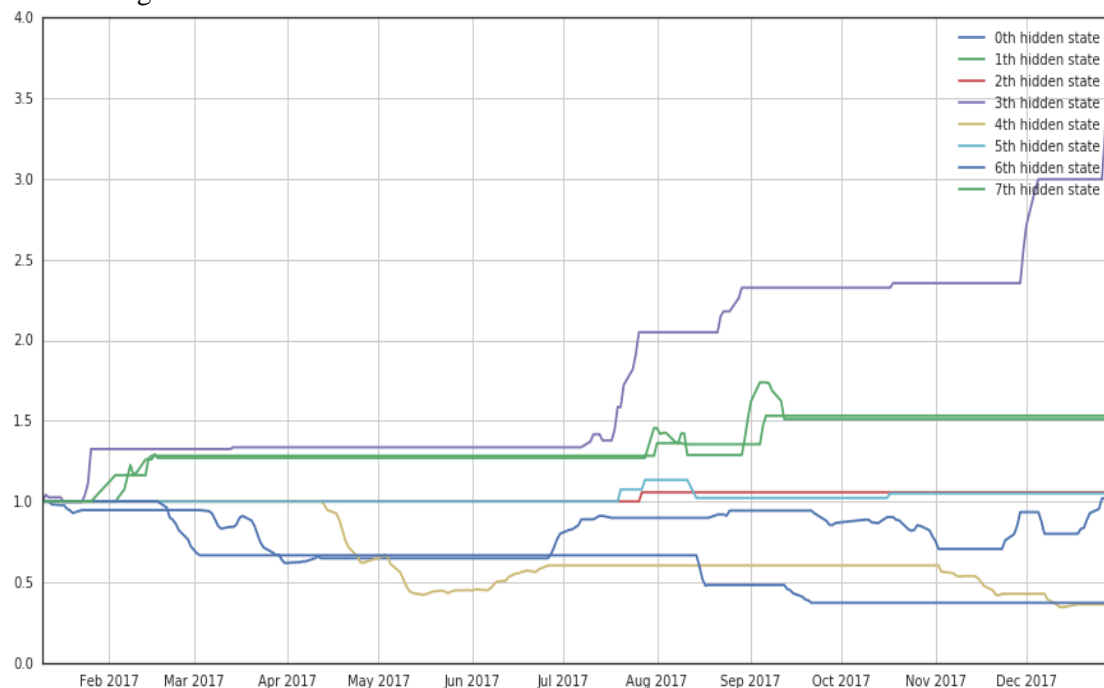
Where  $H$  is a set of hidden states, the random number between 8-12 is generally selected.  $O$  is the set of observation variables. In HMM model, it is associated with the hidden state and can be obtained by direct observation. Five observable states are selected: one-day logarithmic return difference, one-day high low logarithmic return difference,  $n$ -day logarithmic return difference, trading volume and total turnover.  $\pi$  is the initial state transition probability matrix,  $A$  is the hidden state transition probability matrix and  $B$  is the transition probability matrix of observation state.

We regard the stock state prediction problem as a learning problem of hidden Markov model, that is, how to solve the parameters of the model given a set of stock observation sequences, so that the probability of observation series is maximum under the model parameters. We select a group of observable parameter series, and use MLE to estimate the hidden Markov model  $\theta$  based on EM algorithm[7], so that the probability of the observation series appearing under the model is maximum. That is,  $\theta$  converges to the maximum value of  $L(\theta)$ ,  $L(\theta)$  is defined as:

$$L(\theta) = \log(P(O|\theta)) = \log\left(\sum_{i=1}^N P(O, I|\theta)\right) \quad (2)$$

Because the distribution of stock state data sets is often irregular, and there is no obvious center point of Gaussian distribution, it is assumed that the state space is a hidden Markov model composed of multiple Gaussian density functions.

We select the one-year data of China a stock market and set the number of hidden variables as 8. The result of hidden Markov model can be used for further reinforcement learning. However, the meaning of hidden state can not be fully defined in advance, and the accuracy of the state also needs to be further verified. For the meaning of hidden state, this paper uses two methods to infer. The first method is to generate a point chart of the closing price in which state 0 is always a trend of rapid decline and then rapid rise, state 1 shows a trend of bear market decline. In order to make the basic meaning of the hidden state more clear, we assume a simple strategy, which builds positions on the second day of each state, closes positions on the fifth day, and checks the return curve of each state, as shown in the figure 1.



**Figure 1** Simple strategic income curve based on state

Through the comparative analysis of a variety of stocks in different time periods, we can conclude that various states can be classified into several categories, including large-scale change state, the market first suppresses and then raises or first raises and then suppresses, in a short period of time (1

month), the market rises rapidly, slowly rises, violently shakes and rises, shakes and rises, rapidly falls, slowly falls, shakes and falls, repeatedly shakes.

After defining the meaning of the state, we need to further clarify the accuracy of the algorithm to predict the state. The definition of accuracy is as follows:

$$T(i) = \frac{\sum_{i=1}^N C(F_i(Close))}{\sum_{i=1}^N C(i)} \quad (3)$$

Where  $i$  is a certain state, the value of  $C(I)$  is 1 in  $i$ , otherwise is 0, and  $f(close)$  is the closing price change rate consistent with  $i$ . According to the calculation of China's A-share market in 2017, the accuracy rate can reach 76.6%.

### 3. Correction of value function in reinforcement learning

#### 3.1. Definition of state

In the learning algorithm, state includes market state and position state. The value of market state is based on Hidden Markov chain, so five observation variables are still selected. Position state includes position state and market value of each stock. The status is defined as:

$$S = \langle (W_1, \dots, W_5); (y_1, \dots, y_{50}); (M_1, \dots, M_{50}), F \rangle \quad (4)$$

Where  $W$  is the market observation variable,  $y$  is the position status of each stock, value 0 or 1,  $M$  is the position profit rate of each stock, and  $F$  is the available funds. The values of observation variables are normalized to -1+1 or 0-1, and 10 intervals are discretized by equal width discretization. The  $m$  value is also discretized into 9 intervals according to the same frequency.

This approach leads to a very large market state space ( $50^{120}$ ). Instead of using DQN[8][9] to solve the problem of state explosion, we design a method of dynamic calculation and modification of value function based on Hidden Markov model.

#### 3.2. Correction of value function

We set the action space to include only two actions: purchase and wait-and-see. The reason why we do not include sell action is that this paper focuses on stock selection in a certain time interval rather than actual trading.

If the position status is invalid (no position), the value function of buy action is determined by the following formula.

$$Q^{(t+1)} = Q^t * V_k + \log \sum_I p(I | \theta) * p(I | \theta) * (Q^t * T(i) + \max(q_k(s, a) * discount)) \quad (5)$$

The value function of observation action is calculated by:

$$Q^{(t+1)} = Q^{(t)} * V_k \log \sum_I p(O | I, \theta) * p(I | \theta) \quad (6)$$

The reward value calculated by Markov model is a priori value, and the maximum reward value of future actions is a posteriori value. Both formulas need to calculate the reward value calculated by hidden Markov state. The calculation method is to rank the number of times and occurrence ratio of each state in the past 10 days from high to low. The reward value is the sum of the set reward value and occurrence ratio of each state.

The biggest reward of future action involves the selection of the next state, which is quite different from playing games in dqn. The state is not completely determined by the action. In each iteration, the states within 30 days after the current state are selected are sorted from the largest to the smallest according to the number of occurrences. The states with the number of occurrences greater than 5 and the largest profit are selected. If not, the first state in the sorting is selected.

### 4. Experiment

We select three hundred stocks in China's stock market. The data from January 1, 2016 to Dec 31, 2016 are taken as training samples, and the data from January 1, 2017 to December 31, 2017 are taken as test samples. According to the rate of return within 30 days after buying the stock.

Based on the hidden Markov model (HMM), we constructs the financial stock selection model, and makes the stock selection strategy through the reinforcement learning. The strategy defines many observable indexes, then defines 8-12 hidden states, and calculates the probability distribution of hidden states by EM algorithm. Furthermore, the reward function of reinforcement learning is modified by hidden Markov prediction, and the execution of buying action is determined. the profitability and sharp ratio of the test are shown as follows.

**Table 1.** Observation indexes.

Name	Value	Name	Value	Name	Value
annual return	32.2%	Info ratio	0.15	Sharp ratio	1.08
volatility	26.5%	alpha	16.0%	Maximum Drawdown	6.5%

It can be seen from the figure below that the cumulative rate of return and the benchmark rate of return are generally on the rise. Among them, the blue line represents the cumulative rate of return. It can be seen that the cumulative rate of return fluctuates greatly, which is caused by the coexistence of market returns and risks. The black line represents the benchmark yield. It can be seen that the volatility of the benchmark yield is relatively gentle.



**Figure 2** Total return

## 5. Conclusion

The main contribution of this paper is to realize the optimization strategy of financial stock selection based on hybrid hidden Markov model and reinforcement learning, and to estimate the hidden variables through the establishment of the observation indicators and the corresponding state variables of the observation indicators. By hiding the factors, we can divide the stock into several States, correct the reward function to these States, and then get the maximum profit. The future research direction is to use CNN method to further optimize predictive ability.

## References

- [1] James Cumming, An investigation into the use of reinforcement learning techniques within the algorithmic trading domain. Master's thesis, Imperial College London, United Kingdoms, 2015
- [2] M.A.H. Dempster, V. Leemans. An automated FX trading system using adaptive reinforcement learning[J]. Expert Systems with Applications, 30(3):543-552.

- [3] Yue Deng, Feng Bao, Youyong Kong. Deep Direct Reinforcement Learning for Financial Signal Representation and Trading. *IEEE Transactions on Neural Networks & Learning Systems*, 2016, 28(3):1-12.
- [4] Yue Deng, Feng Bao, Youyong Kong, Zhiquan Ren, and Qionghai Dai. Deep direct reinforcement learning for financial signal representation and trading. *IEEE transactions on neural networks and learning systems*, 28(3):653–664, 2017
- [5] Zhipeng Liang, Hao Chen, Junhao Zhu. Adversarial Deep Reinforcement Learning in Portfolio Management[J]. *Papers*, 2018.
- [6] Hassan M R, Ramamohanarao K, Kamruzzaman J, et al. A HMM-based adaptive fuzzy inference system for stock market forecasting[J]. *Neurocomputing*, 2013, 104(Complete):10-25.
- [7] Benjamin Quost, Thierry Denœux. Clustering and classification of fuzzy data using the fuzzy EM algorithm[J]. *Fuzzy Sets & Systems*, 2015, 286(2):134-156.
- [8] Arulkumaran, Kai, Deisenroth, Marc Peter, Brundage, Miles. Deep Reinforcement Learning: A Brief Survey[J]. *IEEE Signal Processing Magazine*, 34(6):26-38.
- [9] Kiran Kalidindi, Howard Bowman. Using -greedy reinforcement learning methods to further understand ventromedial prefrontal patients’ deficits on the Iowa Gambling Task[J]. *Neural Networks*, 20(6):676-689.