Stock Return Prediction Based on CNN Model and Genetic Evolutionary Algorithm For A-share Medical Sector

Shirui Wang[1, †], Zixu Wu[2, †], Bo Yao[3, †] and Yuxin Yan[4, *, †]

[1]Department of statistics, Tianjin University of Commerce, Tianjin, China

[2]Zhuhai College of Science and Technology, Zhuhai, China

[3]School of automation, Chengdu University of Information Technology, Chengdu, China

[4]University of California, Santa Barbara, Goleta, The United States

*Corresponding author: yuxin_yan@ucsb.edu

[†]These authors contributed equally.

Abstract. Since the outbreak of the epidemic in December 2019, the real economy has been seriously affected, where the stock market has experienced a short-term crash. In this context, the financial market has become more turbulent and unpredictable. Previously, the traditional prediction method only captures the linear relationship between both the dependent and the independent variable. With the growing advancement of technology and science, the application of deep learning models to learn the nonlinear relationship between high latitude indicators and predictors has become the mainstream trend of stock prediction. Therefore, a convolutional neural network (CNN) model with fusion genetic algorithm is proposed in this paper to predict and analyze the stock return rate of China's A-share medical sector from January 1,2020 to February 2020. In order to facilitate the operation, this paper chooses the GPU with better performance to process the large-scale calculation in the neural network and builds the experimental environment under the Huawei cloud model arts notebook system. This paper selects the Sequential model belonging to the tensor flow library Keras package. Subsequently, this paper optimizes the parameters in the model by using genetic algorithm. According to the analysis, the mean square error of the optimal model is 0.0017, and the prediction results of the model in the test set are validated with well performances, which is of great significance to investors' decision-making. These results shed light on guiding further exploration of stock trend prediction.

Keywords: CNN model; stock return forecast; A-share medical sector; genetic algorithm; COVID-19.

1. Introduction

With the rapid growth in the national economy, the stock market and the national economic development more closely linked. The stock trend is not only related to the stability of the national financial market and the formulation of economic policies, but also has an important impact on the personal life of each investor. Contemporarily, due to the full-blown outbreak of the pandemic, the economy has suffered a severe downturn. On this basis, there has been a short-term crash in the stock market, resulting in a major reshuffle of stock classification indicator scores in 2020. During this period, owing to the large demand for medical devices and drugs and the government's support for vaccine research and development, the market has begun to re-examine the biopharmaceutical industry. Consequently, funds flow into the medical market and promote the development of related medical and technological companies. Many stocks in the medical sector continue to rise and stop, creating a momentum for the rapid development of the biopharmaceutical industry [1]. Therefore, the medical sector has also become a hot topic and focus of investors and researchers nowadays.

However, the stock market has been a very large and complex market, social, political, economic and other factors always affect the stock market [2]. Therefore, the prediction and rule analysis of stock trend has always been a difficult point in research. As an important factor to reflect the stock trend indicators, researches of stock return forecast have been the direction of scholars that continue to explore. The artificial intelligence has been developed rapidly since the 21st century, under this circumstance some scholars ingeniously start to combine artificial intelligence algorithms with financial market analysis to explore the trends and laws of financial data and establish financial trading strategies. By doing this, they are capable to obtain higher returns and avoid a certain degree of transaction risk to maximize returns [3]. As typical representatives of artificial intelligence algorithms, deep learning (DL) and reinforcement learning (RL) are mostly used in algorithmic trading, risk management, fraud detection, portfolio management and other fields. However, neural network models, time series models and decision tree models based on reinforcement learning perform well in stock forecasting, but few scholars have back-tested the accuracy of their model predictions [4].

Therefore, based on the stock data of China's A-share pharmaceutical industry from the initial outbreak of the epidemic (January 1st, 2020) to February 2022, the genetic evolution algorithm is applied in this paper to accomplish the research of stock return prediction. By combining the CNN model based on reinforcement learning, a stock return prediction method based on the prediction method using genetic evolution algorithm for back testing is proposed. It is expected that this back testing method can reasonably adjust the parameters of the CNN model to achieve the purpose of optimizing the model and improving the precision and accuracy of model prediction.

2. Convolutional Neural Network Models

2.1 Introduction to Convolutional Neural Networks

The neural network is a learning algorithm composed of complex perceptron [5]. The general composition involves the input layer, the hidden layer and the output layer. Each layer contains a composition of several neurons, in other words, several values. The layers are connected by multiplying the weight matrix with the vector of the previous layer. Before reaching the final output layer, an activation function is added to perform nonlinear processing. In the actual operation, we obtain the gradient through the partial derivation and error sum of each parameter to be solved, and then iteratively update according to the negative gradient direction until a satisfactory result is obtained.

In this paper we will use convolutional neural networks to visualize the resulting individual stock data factors using convolutional neural networks to predict subsequent data trends. The convolutional neural network (CNN), which is widely used in deep learning, is proved to be the first multi-layer neural network learning algorithm. Its advantage is that it can capture the spatial characteristics and local correlation of information, and then extract the most representative global features for stock prediction [6]. In addition, Shen and Shafiq proposed three feature expansion methods, namely, maximum and minimum scaling, polarization and calculation of fluctuation percentage, in which the input data is the most commonly used technical indicator [7]. Spatial correlation is used in this approach in order to reduce the number of parameters and thus improve training performance.

Convolutional neural network in the original multi-layer neural network based on the addition of imitation of the human brain to the signal processing part of the classification, i.e.., feature learning. A partially connected convolution layer and a pooling layer are added in front of the original fully connected layer as the specific operation and add a level: input layer-convolution layer-pooling layer-output layer [8].

2.2 Convolutional layers

The convolution layer has parameters which are composed of a certain set of learnable filters. Each filter is relatively small in space, while the input data and its depth stay consistent. When determining the positional relationship between local features, features with local correlation are extracted. In the convolutional layer, the kernel extracts features when it moves at a certain interval on the input data. Moreover, since the convolution has the characteristics of weight sharing, it is able to reduce the number of parameters and the computational overhead and prevent overfitting that is caused by too many numbers of parameters. We use one-dimensional data and assume the input feature layer as a tensor, where the convolution operation is shown in follows [8]:

$$C_i = f(\omega \cdot x_{i:i+h-1}) \tag{1}$$

where C_i is the result of the convolution operation, f is a nonlinear function, and ω denotes the size of the convolution kernel, and $x_{i:i+h-1}$ denotes the sliding window consisting of the ith row to the i+h-1, and the number of words taken for the convolution calculation at each sliding window is the height h of the convolution kernel. Here, ω and $x_{i:i+h-1}$ must be the same dimension, and the size of the feature vector output by the convolutional layer is $(d-h+1) \times n \times 1$.

2.3 Pooling layer

The most representative feature samples are extracted from the convolution layer in the pooling layer. Average and maximum pooling are included in the sampling methods. Sampling is performed by extracting the average or maximum value of each interval. The pooling process is capable of removing unnecessary detail information and preserve features while compressing parameters. Therefore, it ensures that the important information only is successfully selected into the learning program. Through this process, the negative effects of noise data can be reduced, and better stableness can be obtained.

2.4 Fully connected layer

The fully connected layer makes connection between the neurons which are generated in the convolutional layer and the pooling layer with all the neurons in the previous layer. By making this connection, this layer can perform classification operations and prediction.

3. Genetic Evolutionary Algorithms

3.1 Introduction of Genetic Algorithm

Genetic algorithm is one kind of search heuristic algorithm used to solve optimization in the field of artificial intelligence. It is a type of the evolutionary algorithm. This solution is based on Charles Darwin's theory of natural evolution. The genetic algorithm embodies the process of natural selection, in which the most suitable individual is selected for reproduction to cultivate the next generation. After evaluating different individuals, different weights are used to enable individuals to mate with different probabilities to generate new offspring. Meanwhile, a constant mutation rate is used to ensure data diversity and prevent rapid local optimal solution.

3.2 Steps of genetic algorithm1

The steps of the algorithms are as follows, which is also illustrated in Fig. 1 [9]:

Coding the decision variables
Set the fitness function
Obtain initial population
Iterative computation
Get the ideal solution

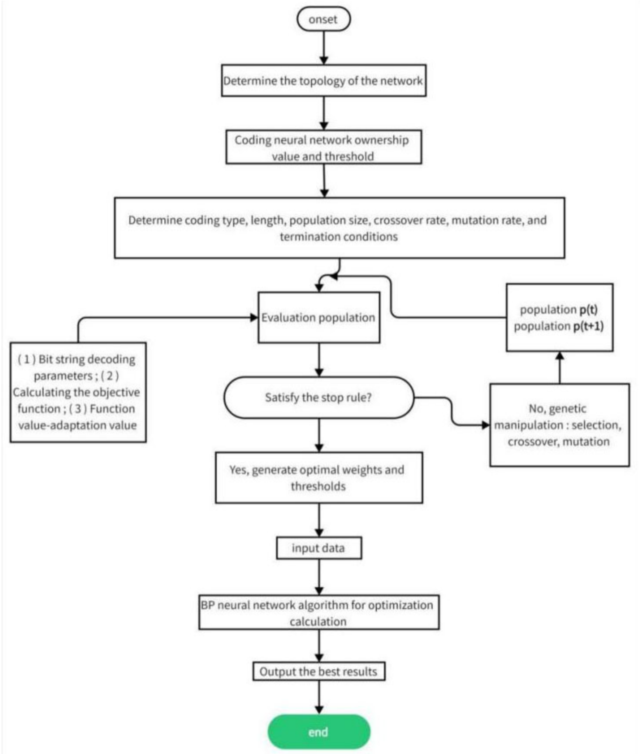

Fig. 1 Genetic algorithm operation process.

3.3 Application field of genetic algorithm

Because of its powerful searching ability and universality, genetic algorithm is widely used in many fields. The main application areas include: function optimization, scheduling problems, image processing, automatic control, machine learning, social and economic fields, artificial intelligence and scientific computing etc. The model accuracy can be improved by genetic algorithm through the process of optimizing parameters. Therefore, in this project, genetic algorithm will be applied to the field of function optimization. Genetic algorithm is applied to optimize the parameters of convolution layer and pooling layer (convolution kernel, filling, step size, pooling kernel, channel number) in convolutional neural network. In this way, the accuracy of the model is improved [10].

3.4 Characteristics of Genetic Algorithm

Genetic algorithm has the following characteristics [10]:

- Form a set of candidate solutions
- The fitness of candidate solution is calculated based on adaptability condition
- Retain several certain candidate solutions according to fitness while other candidate solutions are ignored
- Generates a new candidate solution for the reserved candidate solution operation

In addition to the common characteristics of the search algorithm, the genetic algorithm has the following four characteristics as well:

- It is initiated based on a set of problem solutions, rather than a single solution.
- It can deal with multiple individual parameters in one group at the same time.
- It does not involve the application of space searching or other auxiliary information, instead it merely uses fitness function value to evaluate the individual.
- It is capable of self-organizing, self-adaptive and self-learning.

4. Establishment of CNN Stock Return Prediction Model

Under the ModelArts notebook system, we set up the experimental environment and use the GPU to process the large-scale computation in the neural network. In terms of model, the Sequential model belonging to the Keras package of the tensor flow library is used in this paper. In terms of model, this paper selects a sequential model which belongs to tensor flow library Keras package. When using the Sequential model for data fitting, four hidden layers are established respectively and CNN and fully connected neural network are used to train eighty percent of the stock data, in other words, the training set data, and MSE is used as the loss function of this experiment. Then, Adam optimizer and dropout mechanism are used to prevent data overfitting and enhance the generalization ability of the model. The experimental flow chart is shown in Fig. 2.

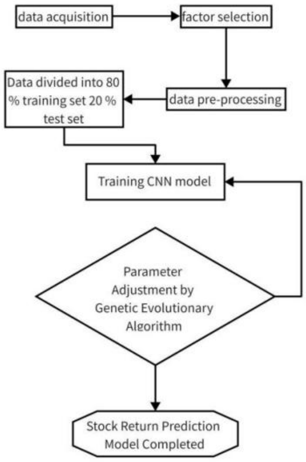

Fig. 2 Experimental flow chart.

4.1 Data acquisition and processing

In this paper, python is used to obtain 42 stock factor data of about 365 stocks in the A-share medical industry sector from January 1st, 2020 to Feburary 2022 from Akshare, Stockstats and Talib libraries. Python is used to process the missing values and infinite values of this data. Some of the data is shown in Figure 3.

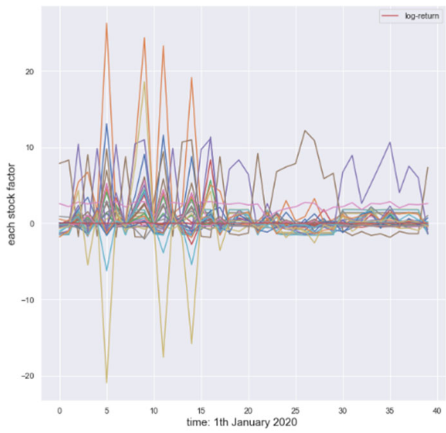

Fig. 3 Partial stock factor data chart

4.2 Selection of stock factors

In order to make the CNN model fitting achieve better results, this paper uses the AIC and BIC methods under multiple regression fitting to screen the existing stock factors, to obtain the stock factors with strong correlation with the dependent variable (stock return). The results are summarized in Table. 1. Seen from the results, the P values of the stock factors selected by AIC and BIC methods are less than 0.005. Therefore, the factors are significantly correlated with the dependent variable (stock return), i.e., these 22 stock factors can be used for subsequent model fitting.

Table 1. Results of AIC and BIC model factor selection

Stock Factor	Pr(>|t|)	Signif.codes	Stock Factor	Pr(>|t|)	Signif.codes
close	< 2e-16	***	low	0.000465	***

Volume	< 2e-16	***	rsi	0.0017	**
Close_2_s	< 2e-16	***	Rsi_6	3.22e-06	***
Stochris_6	1.29e-06	***	trix	6.17e-05	***
Middle_10_trix	1.19e-05	***	Middle_10_tema	0.00162	**
Vr_6	0.0003	***	wr	3.41e-12	***
Wr_6	8.27e-08	***	cci	9.02e-05	***
Cci_6	0.0463	*	pdi	9.01e-06	***
mdi	3.94e-14	***	dx	0.0046	**
Kdj_k	7.13e-16	***	Bias_6	2.48e-09	***
Bias_24	1.59e-10	***	Upper	6.68e-05	***

Note: * * *, * * , and * denotes for p<0.05, 0.01, 0.001, respectively

Since the autocorrelation between factors will cause over-fitting in the subsequent CNN model fitting to a certain extent, which will affect the prediction effect of the model. Therefore, this paper uses the Pearson correlation coefficient thermal matrix to screen the stock factors, remove the indicators with strong autocorrelation to guarantee the accuracy of the prediction effect of the subsequent model.

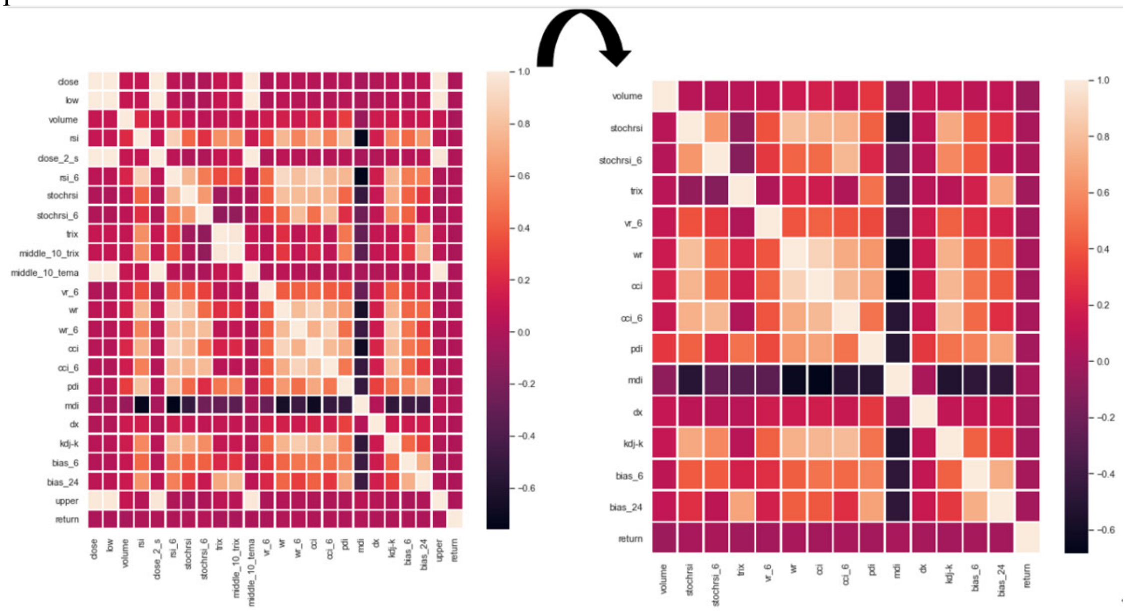

Fig. 4 Thermal matrix of Pearson correlation coefficient

The variance expansion factor (VIF) in this paper is used to test whether the selected factors have multicollinearity. The results are shown in Table 2. According to the results of Table. 2, the VIF values of the final selected stock factors are less than 10. Therefore, this paper believes that there is little multicollinearity among the 14 stock characteristic factors, and this data can be used for model fitting.

Table 2. Variance expansion factor VIF results

Indicators	Statistics VIF
volume	1.120
stochrsi	4.671
Stochrsi_6	3.635
trix	4.101
Vr_6	1.315
wr	7.184
cci	9.012
Cci_6	6.305
pdi	3.804
mdi	2.342
dx	1.182

Kdj_k	3.428
Bias_6	5.181
Bias_24	8.847

4.3 Definition of data factor index

As mentioned above, this paper uses the AIC and BIC methods under multiple regression fitting to, Pearson correlation heat matrix and variance expansion factor for initial screening of model factors and retains the following 14 stock factors as independent variables for this model training, and their factor indicators are defined as summarized in Table 3.

Table 3. Definition of stock factor

Volume	Volume	trix	Triple exponential smoothing moving average indicator
stochrsi_6	6 period stochastic relative strength indicator	Stochrsi	Stochastic Relative Strength Indicator
Pid	Uptrend direction line	Vr_6	Volume variability over 6 cycles
Cci	Homeopathic Indicators	wr	William Indicator
Cci_6	6 cycles of homeopathic indicators	Dx	Dynamic value
Kdj_k	K Series	mdi	Descent direction line
Bias_6	eccentricity	Bias_24	eccentricity

5. Analysis of Results

5.1 Genetic algorithm tuning parameters

The genetic evolutionary algorithm is constructed by GAFT framework and the quantity of convolution kernels, the quantity of neurons, the frozen weight and the learning rate in each layer of CNN model are adjusted iteratively. In terms of genetic algorithm parameter setting, this paper sets the crossover probability to 0.8 and the mutation probability to 0.1. On this basis, if the mutation does not occur in the next generation, the mutation rate will increase to 50 %.

5.2 Setting of experimental parameters

In this paper, the cubic fitting parameters in the genetic algorithm iteration are selected for display. The results are shown in Table 4. Experiment 3 is the best fitting parameter. At the same time, the remaining model parameters are set to the same, that is, using the 'tanh' activation function, Adam optimizer, in the genetic algorithm iteration CNN model training times set to 20 times.

Table 4. Model parameter settings

	Filter1	Filter2	Filter3	dense	dropout	Learning-rate
First experiment	585	612	1139	4077	0.2	0.0002
Second experiment	367	534	726	2582	0.01	0.0002
The third experiment (the best fitting)	342	631	874	2000	0.2	0.000004

5.3 Analysis of results

In this paper, the A-share medical sector data from January 1,2020 to February 3,2020 is considered as a training set for model training, and the stock data from February 4,2020 to February 11,2020 is used as a test set for model prediction. The model fitting results under these three sets of parameters are shown in Figure 7. compared with Experiment 1 and Experiment 2, the third experiment has the best fitting effect.

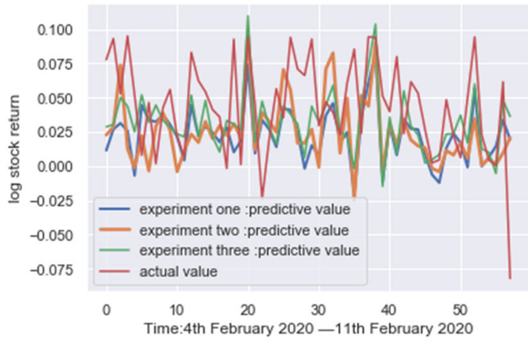

Fig. 7 Prediction result graph.

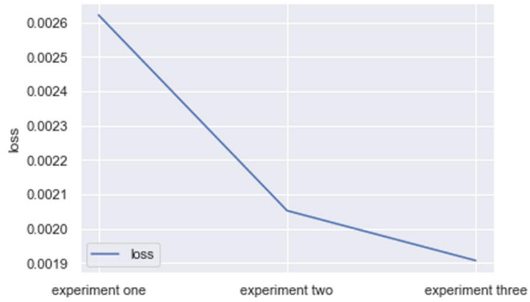

Fig. 8 MSE trend chart

5.4 Fitting the MSE of the model

In order to explore the difference of the model fitting effect in the three experiments, this paper selects MSE as the evaluation index of the model. The MSE value of the three parameters fitting is given in Fig. 8. According to the trend of MSE, with the cyclic adjustment of genetic algorithm, the MSE value of the parameters under optimal fitting is the lowest, that is, the fitting effect of CNN model is better under this parameter.

5.5 Summary

According to the above results, when the weight frozen by the model is 0.0017, the quantity of neurons is 4370, and the number of convolution kernels of the three-layer convolution layer is set to 68,185 and 530 respectively, the model achieves the optimal fitting effect, that is, the minimum MSE. The comparison between the true value and the predicted value is exhibited in Fig. 9. According to the results, the model has been effectively improved after genetic algorithm parameter optimization, which makes the model generalization ability improved and the fitting effect better, which has certain practical significance for investor decision-making.

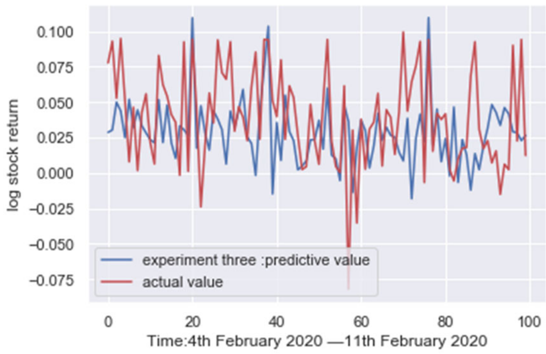

Fig. 9 Predicted value (red) and true value (blue)

6. Deficiency and outlook

On the other hand, there are several limitations in this study that need to be addressed. The data set selected in this paper is limited to a relatively short period of time, and some important features of longer periodicity may be ignored. The selection of parameter factors is mainly limited to technical factors, some macroeconomic indicators, enterprise indicators, social news and other influencing factors and included in the algorithm. As a result, in order to further improve the accuracy of the model, we hope to expand the sample size of the data set in subsequent studies. The convolutional neural network algorithm consists of a large amount of calculation and many parameters, which will greatly occupy computing resources. At the same time, the genetic algorithm without definite termination principle leads to long training time, which makes it difficult to process large amounts of data. To further improve the computational efficiency, a more optimized genetic algorithm should be studied.

There are many random factors affecting the stock market in real life. For example, before and after the announcement of economic indicators, stock movements mainly depend on the magnitude of investors' expectations and expected differences. These factors are difficult to be determined by artificial markers, so convolutional neural networks cannot identify or ignore certain features. Therefore, the neural network algorithm used in this paper cannot well understand the shortcomings of eigenvalues, thus affecting the model fitting degree. The data set image should be transformed to some extent to enhance the randomness of the training set.

7. Conclusion

In conclusion, this paper investigates Stock Return Prediction based on CNN Model and Genetic Evolutionary Algorithm. Specifically, the near optimal factors and parameters combination for CNN model are found in the context of genetic evolutionary algorithm, which was used in stock return prediction. According to the analysis, the difference between genetic evolutionary algorithm and global search algorithms in model training time and final model effect. It is believed that genetic evolutionary algorithm reduces the time to approach the optimal solution, while the result enhances the effect of the model classifier. Nevertheless, genetic evolutionary algorithm has some limitations. First, if the model has multiple near-optimal solutions and the parameter combinations vary greatly it is very likely to converge to a local optimum. Secondly, the potential capability of genetic evolutionary algorithm's parallelism mechanism is not fully exploited, and there is still a large optimization space in the convergence time. In the future, we can use genetic evolutionary algorithm for other artificial intelligence techniques methods for example case-based reasoning and decision trees. Overall, these results offer a guideline for solving the problem of model parameters adjustment.

References

[1] Gao Yixin,Li Khanru,Guo Yijun. Research on the impact of the new crown epidemic on stock investment. Journal of Hubei College of Economics (Humanities and Social Sciences Edition), 2021, 18(12): 48-51.

[2] Chiu D Y, Chen P J. Dynamically exploring the internal mechanism of stock market by fuzzy-based support vector machines with high dimension input space and genetic algorithm. Export Systems with Applications, 2009, 36(2): 1240-1248.

[3] Cen Yuefeng, Zhang Chenguang, Cen Gang, Zhao Cheng. Stock Price Forecasting Method Based on Proximal Reinforcement Learning. Control and decision, 2021, 36(04): 967-973.

[4] Zuo Chuan, Wang Yu, Li Zhen. Application of Deep Learning in Stock Investment. Journal of Zhengzhou University (Philosophy and Social Sciences), 2022, 55(02): 56-62.

[5] Haykin S. Neural Networks: A Comprehensive Foundation. Neural Networks A Comprehensive Foundation, 1994: 71-80.

[6] Hoseinzade E., Haratizadeh S. CNNpred: CNN-based stock market prediction using a diverse set of variables. Expert Systems with Applications, 2019, 129, 273-285.

[7] Shen J., Shafiq M. O. Short-term stock market price trend prediction using a comprehensive deep learning system. Journal of big Data, 2020, 7(1), 1-33.

[8] Kim Y. Convolutional neural networks for sentence classification. Empirical Methods in Natural Language Processing 2014. 2014:1746-1751.

[9] Jin Jingzhe. Principle and Application of Genetic Algorithms. Science and Wealth., 2019, (7): 249.

[10] Liu Haiyue. Analysis of genetic algorithm and neural network in stock prediction. 2011(05): 1-71.