

# Stock Trend Analysis Based on Comprehensive Prediction Algorithm

Mingzhe Guo

School of Computer and Information Technology, Beijing Jiaotong University, Beijing, 100044, China

**Keywords:** Stock trend, Time series, Machine learning, Comprehensive forecast

**Abstract:** With the rapid development of the economy, the stock market has received widespread attention from investors. Combining the advantages of k-nearest neighbor algorithm, support vector machine algorithm and time-series algorithm, this paper proposes a comprehensive prediction algorithm and applies it to the CSI 300 index's ups and downs prediction. Firstly, a neighboring algorithm and a time series algorithm are combined to obtain a rising and falling variable of the future trend. The image is converted into a two-dimensional code number by image processing, and the nearest neighbor algorithm is used to analyze whether the curve of the time series prediction is up or down; then the variable is Input into the support vector machine algorithm to get the final synthesis algorithm; finally compare with a single time series algorithm to get the final accuracy. This method can make up for the shortcomings of the original algorithm and can predict the rising and falling trend of the stock market more accurately.

## 1. Introduction

Mastering the law of stock market changes and predicting its trend has always been a hot spot for investors and investment companies. The exploration of the method of predicting the future earnings of the stock market has never been interrupted. Since it involves the financial sector, it is only starting from the financial engineering and mathematical statistics to explore relevant methods. However, since the financial model is found to be unable to meet people's requirements for prediction accuracy to a certain extent, it is more based on the hypothesis for the financial model. Later, people proposed a machine algorithm model, which is more practical, wider in scope, and more accurate in model.

In the early 21st century, the financial market developed rapidly and more financial products were derived. Therefore, a single financial model algorithm and machine learning algorithm can not meet the market demand, people combine the financial model algorithm with the machine learning algorithm to achieve better results. For the quantitative analysis of the stock market in financial engineering, the advantage of the machine learning algorithm is that it can simulate the specific characteristics of the object to the greatest extent, and it also has a greater advantage in dealing with the amount of data and complexity. Another kind of prediction method is to combine various algorithms to compensate for the defects of the individual algorithms to some extent.

## 2. Time Series Model

This paper will use the advantages of k-nearest neighbor algorithm, support vector machine algorithm and time series algorithm to achieve the best stock index prediction effect.

As stock prices change over time, there is a very obvious chronological order. This series of observations arranged in time series is called a time series, so the time series can be considered to contain one or several statistical indicator features. A time series generally contains four kinds of information: long-term trend, cycle, seasonal transformation, and irregular transformation. Common time series models mainly include autoregressive model, moving average model, autoregressive moving average model, homogeneous non-stationary model and so on. The time series model is further divided into a traditional time series model and a modern time series model. The traditional time series model regards the time series as a complex of long-term trend, seasonal variation, cyclic

transformation, and irregular transformation. The modern time series model considers the time series as a random probability process, and the arima model algorithm is the representative of such models.

### 3. Comprehensive Prediction Algorithm

#### 3.1 Algorithm steps

The comprehensive algorithm prediction model is mainly divided into three parts. Firstly, the accuracy of support vector machine algorithm, time series algorithm and synthesis algorithm are calculated under the same data conditions, and then compared. The specific steps are:

- (1) First standardize the data;
- (2) After normalization, the accuracy of the support vector machine algorithm and the time series algorithm are calculated respectively;
- (3) According to the picture of the same pixel size, 20 consecutive 30-day Hushen 300 historical ups and downs are randomly taken as a sample set. Divide 20 graphs into two categories, namely up and down, and the number of samples in each category is 10;
- (4) According to the variable date selected by the support vector machine, draw a prediction curve of the time series for the next 30 days, and intercept the prediction curve according to the same pixel size;
- (5) Through the neighbor algorithm to compare, identify whether the time series of the next 30 days of the graph is up or down, and define the rise as 1, and the drop as -1;
- (6) The result obtained by the neighbor algorithm is input as one of the support vector machine variables, and finally the prediction accuracy of the integrated algorithm is calculated;
- (7) The accuracy of the support vector machine algorithm, time series algorithm and synthesis algorithm is compared, and the final conclusion is obtained.

#### 3.2 Algorithm principle

The integrated algorithm in this paper uses image processing technology as well as time series algorithm, nearest neighbor algorithm and support vector machine algorithm.

For image processing technology, the combination of image processing technology and machine learning algorithms can avoid the limitations of the naked eye and avoid duplication of work because it is not accurately judged by the naked eye at a certain stage of the stock market. The error rate brought. The color picture is binarized by the matlab software, so that the picture only displays black and white. Binary images are also called black and white images, because each pixel has only black and white, and there is no intermediate transition color, where 0 and 1 in the binary matrix of the output also represent the black and white value of the binary image, where black the pixel block represents 0 and the white pixel block represents 1.

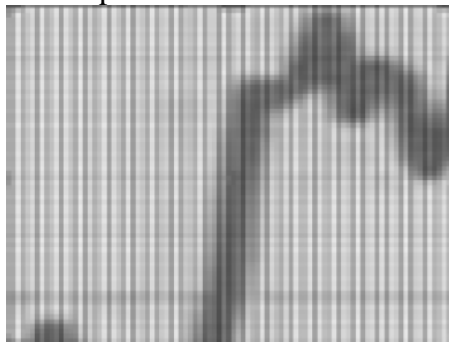


Figure 1. 300 ups and downs binarized image

According to the same pixel size, the picture of the ups and downs is taken, and the chart of the Shanghai and Shenzhen 300 is shown in Figure 1. In the figure, a clip taken image was randomly selected from the sample set, representing an upward trend. Then the black square corresponding to the figure is counted as 1, and the white square is counted as 0. In the algorithm, 20 consecutive

30-day Hushen 300 historical ups and downs are randomly selected as the sample set. The 20 graphs are divided into two categories, namely, ups and downs. The number of samples in each category is 10. Then, the 20 sample set images are converted into binary images, and then outputted by the program in the form of binary matrix data, and the labels of "up" and "fall" are respectively saved.

The time series algorithm predicts the stock market's rise and fall in the next 30 days, and then intercepts the same pixel's ups and downs, binarizes and converts it into a binary data matrix. Compared with other algorithms, time series algorithms have the ability to reflect the trend of future stock market fluctuations, and can maximize the stock market cyclical. Then, by intercepting the time series prediction graph, it is converted into binary data as a test set of the k-nearest neighbor algorithm.

The sample set is then trained by the k-nearest neighbor algorithm, and then the test set is input to determine whether the set of binary matrix data is up or down. The rising state is defined as 1, and the falling state is defined as -1, which is stored as one of the inputs of the support vector machine. The distance between the test set and the sample set array is calculated by the k-nearest neighbor algorithm to distinguish whether the state of the test set is up or down, and then the obtained experimental result is used as an input variable of the support vector machine. At the same time, the other variables of the support vector machine are used, that is, the daily closing price and daily trading volume of the first three days of the CSI 300 in the period from 2015 to 2016, and the index of the stock index on the fourth day is predicted, because under such conditions. The prediction results are more accurate and the experimental results are better. After the required variable data is obtained, it is normalized separately to make the experimental results more accurate.

Therefore, for the SVM algorithm, there are mainly three input parameters, the first parameter is the ups and downs predicted by the time series, the second is the daily closing price of 1, 2, and 3 days in advance, and the third is the advance 1, 2, 3 days of daily trading volume. The training set's single-day ups and downs are defined as 1 or -1. The test set is then calculated by an algorithm, and the final accuracy of the integrated algorithm is calculated from the results obtained by the test set.

## 4. Experimental Results

### 4.1 Time series simulation results

The time series algorithm must first smooth the prototype and de-cycle it before performing the operation. It is automatically modeled in SPSS Modeler to select the best first-order difference times and first-order seasonal difference times. This paper predicts the final stock market by entering the closing price and opening price of the CSI 300 in 2015. This paper does not consider the introduction of seasonal differences, and automatically obtains the ARIMA (1, 0, 2) model through software.

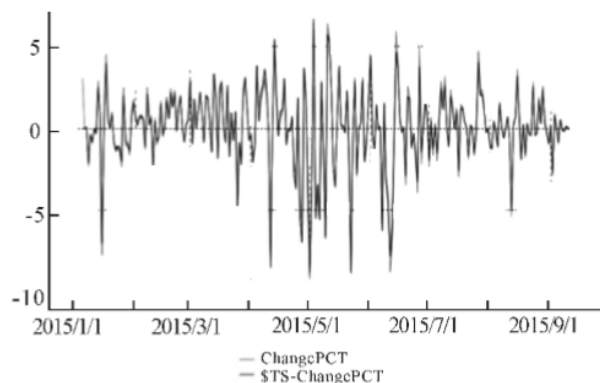


Figure 2. Time series training prediction results

Figure 2 is a graph showing the results of the training of the time series training on a daily basis. Since stock trading has weekends and holidays, the date is re-arranged for the daily cycle, so the weekend is skipped, and the final date is advanced compared to the monthly cycle. It can be seen from the figure that the training of the time series is better. But at the same time, the effect of the

prediction set is not very satisfactory. And the accuracy here refers to the probability of a rise and fall, and is not accurate to the ratio of the rise and fall. The possible reason is that the model is not seasonal, and the input of parameters may be too small to fully reflect the overall status of the stock market. For the time series algorithm, based on the historical data, there will be no certain accuracy for the future prediction range, and the randomness factor among them is also large. But the advantage of the time series algorithm is that it can roughly judge the overall cycle trend of the future stock market. Therefore, the time series algorithm needs to be combined with other algorithms to better improve the prediction accuracy.

#### **4.2 Comprehensive algorithm simulation**

According to the principle of the integrated algorithm, 20 up and down charts are randomly selected and converted into binary matrix data, and divided into two categories, namely up and down. Then predict the image of the next month through the time series, the last month is the predicted image, and then push back one day, save the predicted one-month rise and fall chart through the spss software. A binary image is then obtained by image processing, and binary matrix data storing each image is output. Each set of data in the svm algorithm corresponds to an image that predicts a future rise and fall in a month.

To get the sample set and test set required by the knn algorithm, you can call the knn algorithm to determine whether the final class of each test set is up or down, and then save it as an input variable of the svm algorithm. Finally, the parameters required to be input by the SVM algorithm are the daily closing price and daily trading volume of 1, 2, and 3, combined with the time series algorithm and the knn algorithm to predict the trend of the next month, and the same as the previous svm algorithm. 60% of the data is used as the training set, and the remaining 40% is used as the test set. Only a set of variable parameters is added to it, and the overall framework has not changed. The final experimental results show that the accuracy of the single svm algorithm and the time series algorithm is not very high, generally only about 70%, and the accuracy of the integrated algorithm is close to 80%. Therefore, the synthesis algorithm works better than a single algorithm. Because the integrated algorithm combines the advantages of time series and svm algorithm, the svm algorithm can take into account the time characteristics of the time series, such as seasonality, and combines the advantages of the svm algorithm, that is, its own error rate is low, and the computational overhead Small, suitable for running large quantities of data, the results are more intuitive and specific.

#### **5. Conclusion**

Combining the advantages of various algorithms, this paper proposes a comprehensive algorithm combining machine learning and financial model algorithms, including k-nearest neighbor algorithm, support vector machine algorithm and time series algorithm. The k-nearest neighbor algorithm is mainly applied to classification. The similarity samples are classified into one class by the nearest neighbor algorithm. The svm support vector machine algorithm can effectively solve the local minimization problem that the neural network cannot avoid, and it is small sample size, nonlinear and high-dimensional. The number model, as well as the advantage of overcoming the dimension and over-fitting learning; the time series algorithm can well show the development trend and law of things in a certain period of time, so as to effectively predict the future changes. Applying it to the CSI 300 Index's ups and downs forecast has achieved good results. However, although the comprehensive algorithm has improved accuracy compared to a single algorithm, there is still much room for improvement in the future. It is necessary to explore and combine some more efficient algorithms to achieve higher accuracy.

#### **References**

[1] BURGERS B C.A tutorial on support vector machines for pattern recognition [J]. Data Mining and Knowledge Discovery, 1998, 2(2):121167.

- [2] Zhang Wenzhao, Dong Wei. Advanced tutorial on statistical analysis of spss [M]. Beijing: Higher Education Press, 2013:393395.
- [3] Lin Limin. Research on short-term stock price forecast based on support vector machine [D]. Hangzhou: Hangzhou University of Electronic Science and Technology, 2010:4550.
- [4] He Yongpei. Improvement of Arma Model Parameter Estimation Algorithm and Its Application in Stock Forecasting [J]. Journal of Chongqing Institute of Technology: Natural Science Edition, 2009(2):68.
- [5] Zhang Peng. Research on stock market forecasting and timing based on svr [J]. Journal of Chongqing University of Arts and Sciences, 2016(3):148155.
- [6] Sun Haibo, Wang Limin. The application of the trend factor bp model in stock market forecasting [J]. Statistics and Decision, 2015 (19):8789.
- [7] Li Yan. Stock price forecast based on arima model and neural network model [J]. Financial Business, 2014 (4): 7780.
- [8] Liu Haizhen, Bai Yanping. Analysis of time series model and neural network model in stock forecasting [J]. Management Science, 2011 (4): 2231.
- [9] Jin Debao. Research on stock market forecast based on support vector machine [D]. Hangzhou: Zhejiang University Press, 2010:210.
- [10] Wang Bo, Cheng Fuyun. Application of knn algorithm in stock forecasting [J]. Wuhan: Technology and Entrepreneurship Monthly, 2015(16):1015.
- [11] Zhang Chenxi, Zhang Yanping. Stock Prediction Based on Support Vector Machine [J]. Computer Technology and Development, 2006, 16(6):3435.
- [12] Pang ning tan, michael steinbach. Introduction to Data Mining [M]. Beijing: People's Posts and Telecommunications Press, 2011:157170.