Research on Stock Index Forecasting Based on Machine Learning

Yanyan Zhuo

Economics School, Tianjin University of Finance & Economics, Tianjin, China

Keywords: Stock index, Machine learning, Support vector machine

Abstract: Stock price index prediction is an important part of stock investment. Due to the highly nonlinear and highly noisy features of the volatility of stock market, it is extremely difficult to predict stock price trend. In this paper, we use machine learning method to give stock price index prediction model based on support vector machine. The whole prediction process of stock index forecasting based on machine learning includes the steps of data acquisition, data pre-process, eigenvector solution and normalization treatment. Among them, we should make use of the linear mapping and genetic algorithm to optimize the parameters to improve the Machine Learning method. Practice has proved that this method gives full play to the stability advantage of machine learning and improves the accuracy of prediction.

1. Introduction

The rapid development of Chinese society and capital market is both a challenge and an opportunity. At present, the domestic market still uses simple basic analysis or technical analysis as a means of investment analysis, under the condition of global capital market turbulence. Quantitative investment, which has been widely used in foreign capital market, is gradually being concerned by many investors. In fact, quantitative investment is a mathematical model of investment based on computer technology. In accordance with the mathematical model, the investment activities of the computer are independently carried out by the computer. In fact, the process of quantification is equivalent to converting the human experience logic into a model, which is expressed in a computer. The advantage is that people can't always be rational and the computer can. The mining of the financial temporal data in the capital market has been paid attention and developed in many aspects, such as investment decision, the pricing of financial products, the prediction of stock index futures and so on.

At present, analytical methods such as econometrics are still the mainstream methods in the application of financial data mining. Although these methods have their own advantages, but in the development of more and more restricted, especially for the nonlinear and non-stationary financial data, the traditional methods of solving these problems have become more and more weak in the solution of financial problems. Throughout the process of the development and progress of human history, learning ability has always been very important. By summarizing the lessons of the past, people can finally make reasonable predictions and inferences about the future. This often requires a summary of many known facts, which is a very heavy workload. But the emergence of computers makes learning easier. Using computer to simulate learning makes the problem simple. Machine learning is becoming an important part of the development of modern intelligent technology. In the case of known sample data, the computer can automatically simulate and estimate the potential relationship between its input and output, and then make a prediction of the output according to the new input. The existing learning methods are often based on the assumption that the number of samples in traditional statistics is infinitely large, but it is not usually the case, which not only restricts the development of traditional financial data mining, but also has a bad influence on machine learning.

In view of the above situation, this paper focuses on the application of machine learning in financial data mining to enhance its application in practice. Most of the researches on financial problems use traditional econometric methods, which all have their own theoretical assumptions and are not consistent with reality. Therefore, how to construct a machine learning algorithm suitable for financial data mining is of great theoretical significance for the research and analysis of financial data. Finally, a support vector machine model combining parameter optimization and feature extraction is constructed to demonstrate the actual financial problems. A support vector machine model is constructed to predict the fluctuation of stock market index based on the index of technical aspect.

2. Stock index and machine learning

2.1. Concept and characteristics of stock index

The stock price index is an index that reflects the general level and the change of all stock prices in the stock market. Usually referred to as the stock index. It is an indicator number compiled by the stock exchange or the golden drift service to help investors understand the stock market changes. The stock price is fluctuating frequently. Although it is easy for investors to understand the change of a stock, it is difficult to understand the changes in a variety of stocks. The financial structure makes use of its own business knowledge and the advantage of the market understanding to produce the stock price index. The publicly released index is used as the index of the price change in the market. Of course, investors can also define the change of the stock index market according to their investment experience. The stock price index is the relative number of stock price compiled and reflected by the general price level of the stock market and its trend of change. The average stock price or stock market value of the reporting period is compared with the selected base term average or stock market value and multiplies the ratio of the two to the index value of the base period, that is, the stock price index of the reporting period.

The number of space features of a stationary sequence generally shows that the mean, variance, and so on does not change with the time. The randomness of the time series at each time point obeys a certain probability distribution and tends to fluctuate around the mean value in the graph. Non-stationary time series show different mean values in different time periods. The standard method for testing sequence stationarity is unit root test. The autocorrelation in the residual terms of the model is a notable feature of the nonlinearity of the financial time series. If the model's residual squared terms show obvious autocorrelation, the financial time series can only be described by nonlinear models. Under the nonlinear null hypothesis, all the residual squares terms of all linear models should be completely independent. Many scholars have used various mathematical methods to prove the nonlinear characteristics of stock index time series. The fluctuation frequency and amplitude of the time series of the stock market are constantly changing, even if there is no hype or significant news influence, it will also show a small amplitude of high frequency fluctuation because of the influence of random factors. This high frequency small amplitude fluctuation is regarded as noise. When predicting the time series, we should use mathematical methods to remove the noise as far as possible. The time series of stock index is a complex dynamic system with non-stationary, nonlinear, low signal to noise ratio and so on. It is necessary to select the appropriate model to analyse the characteristics of the sequence.

2.2. Machine learning and support vector machine

Learning is to use training data or experience to optimize the parameters of the model. Machine learning uses statistical theory when constructing mathematical models, because its core is reasoning from samples. In general, the input and output pairs reflect a function relation that maps the input to the output, but if there is noise interference, the intrinsic function often exists, which is called the objective function. The estimation of objective function by machine learning is also considered as a solution to learning problems. In classification problems, these functions are often referred to as decision functions. The object of the study is the system. When the input is given, the output can be obtained, and the learning machine in the graph can be regarded as an approximation of the inherent law of the system, and its output is the forecast output. In fact, sample data are used to simulate the relationship implied in the system, and the implicit relationship is used to predict the new sample data. Figure 1 shows the basic process of machine learning.

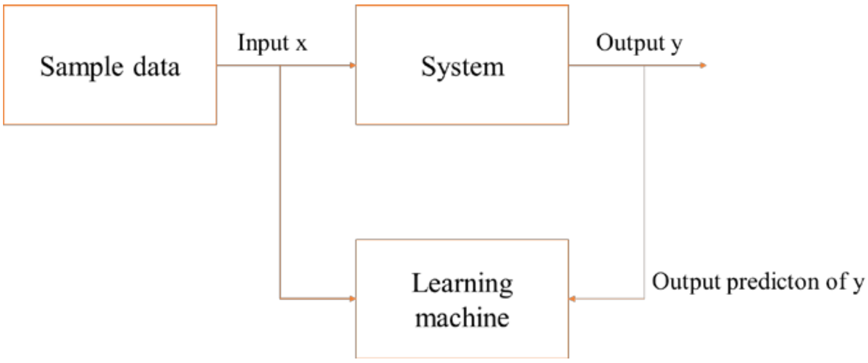

Figure 1. Basic process of machine learning

The support vector machine is based on the theory of statistical learning. The theory of statistical learning is the best theory recognized at present when it makes statistical and predictive learning of small sample data. This theory systematically studies the conditions of the principle of empirical risk minimization, the relationship between the empirical risk and the expected risk under the limited sample, and how to use these theories to find new learning principles and methods. Empirical risk minimization is defined as the average error rate defined on the training set, which is the expected risk of the whole sample set. But when it has enough samples, it has better generalization performance when the number of samples tends to infinity. This is the ability of learning machines to correctly predict future output. Obviously, in the case of small samples, the empirical risk minimization criterion does not have sufficient theoretical basis. The training error is not necessarily high, so there will be a learning phenomenon. The relationship between empirical risk and expected risk at least meet the following formula:

$$\mathrm{R}(\mathrm{a}) \le \mathrm{R}_{emp}(a) + \phi(\frac{h}{k}, \frac{\log(\eta)}{k})$$

The confidence interval is defined as

$$\phi\left(\frac{h}{k}, \frac{\log(\eta)}{k}\right) = \sqrt{\frac{h(\log\left(2 \times \frac{k}{h} + k\right) - \log(\frac{\eta}{4}))}{k}}$$

For the above training set, the confidence interval is related to the dimension of machine learning and the number of training samples.

3. Model of stock index forecasting based on machine learning

3.1. Data acquisition

To predict the stock index better, we use two data. The first group is the historical trading data of the Shanghai Stock Fifty index, which shows the change of the daily market index. The date is based on the day. The second group is the real time transaction data of each stock in Shanghai Stock Fifty Exchange. These data are changed in units of seconds. When we grab the data, we will merge the data. The processed data attribute includes four attributes: time, transaction price, turnover volume and turnover volume. In the experiment, we use stock index and stock historical data to predict stock index. The market index is used to get the label used in the experiment, so that we can reflect the last trend of stock index. When we extract eigenvectors, we get the historical transaction data of stock. Understand the real time trading data and find the change rule. The stock index can be better predicted. For data crawling, the main task is to capture historical transaction data of stocks. As the specific transaction data of each stock is a unit of second, the amount of data is large, and the time cost will be very huge if the traditional single thread is used to download. To save time and cost in the experiment and make the experiment widely applicable, we use batch processing to download. In the experiment, we need some batch files under the command line, which extends the file. To facilitate data processing in the future, we use traditional storage methods to place the captured files in local text files. This is not only convenient for viewing files, but also convenient for reading files. It has the advantages of convenient and fast processing. The drawback is that the storage requirements of computers are high when data are processed and calculated scientifically.

3.2. Data pre-process

To achieve the experimental results, we need to clean the collected data to merge and clean the data. To apply the data better, we need to make the data into a minute and then merge the data to the data cleaning. The purpose is to standardize the data, delete the information of the repeated information and the incomplete attributes, correct the errors in the data, and provide the data conformance data merging and cleaning. Hadoop is a distributed processing software framework, which can be divided into several groups according to the function of the distributed computing and distributed file system, so that the system has a very reliable architecture and a very good fault-tolerant system named Hadoop. When a node has a problem, its task will not fail, but it is transferred to other nodes based on the technology of rack perception to ensure the efficiency of the operation of Hadoop because it has a good platform extensibility. We can analyse it one by one and then summarize the results to get what we expect, although the internal operating system of Reduce is very complex, but our requirements command is very simple, and we can generate the data we want through a few lines of short code.

3.3. Eigenvector solution

Starting from the transaction distribution of each stock, the purpose is to understand the share of the stock on this day, which is the side to understand the favour of a stock. We start from the first fifteen minutes and thirty minutes per day according to the experience that most of the shareholders will choose the transaction within the two time, and the rest of the day. If people prefer a stock, the volume will be larger in the first fifteen minutes and thirty minutes. Conversely, the volume of the two time is taken as a feature vector, and the W is very significant compared to the characteristic values of the different stocks. We have learned about the share of stocks at fifteen minutes and thirty minutes per day at feature. If this is only the case, we can only understand whether the stock is favoured by most of the stock on this day and can't show whether the stock is concerned with the previous period. To solve this problem to solve this problem, we take the volume of the first fifteen minutes and thirty minutes in the first five days, ten days, twenty days and forty days respectively to reflect the stock's turnover in this day compared to the previous period. In the future, the stock is supposed to be good for the stock with higher value, and on the contrary, the stock is relatively poor. As a characteristic value, it can predict the trend of the stock index well.

3.4. Normalization treatment

Normalization is a dimensionless processing method, which makes the absolute value of physical system numerical value change into a relative value relation. It is an effective way to simplify the calculation and reduce the value of the quantity. For example, each frequency of the filter is normalized with cut-off frequency, and the frequency is the relative value of the cut-off frequency, and there is no dimension. After the impedance is normalized by the internal resistance of the power source, each impedance becomes a relative impedance value, and ohms has no such dimension. After all kinds of operations are over, anti-normalization is restored. We use the proposed algorithm to obtain the eigenvectors that reflect the change in the stock index. Because the eigenvectors are solved by different methods, the value produced is not on the same order of magnitude. When the model is trained, the data must have the same unit amount, otherwise a one-dimension feature may have a great influence on the model, thus affecting the accuracy of the model. We use the normalization method. The function of normalization is to transform data of different dimensions and different sizes into data with the same dimension and magnitude, which can make the data comparable between the dimensions. The usual normalization method is logarithmic function method.

For the sample data x(n), the data y(n) after the normalization treatment is:

$$\mathrm{y(k)} = \log\big(\mathrm{x(k)}\big)\ \ k = 1,2,3 \dots n$$

4. Parameter optimization of machine learning algorithm

4.1. Optimization of parameter σ

There is a great advantage in the application of support vector machines: only the original sample data can be used to complete nonlinear mapping, and the precision is high. If the parameter is changed, the nonlinear mapping structure will change correspondingly, so the parameter selection problem has always been one of the research directions. In this paper, the main parameters to search are two parameters that have important influence on support vector machines. The function parameters, such as parameters and error penalty factors, are used when kernel functions are used. If these two parameters are not selected properly, the over learning phenomenon will be caused. The advantage of support vector machine, as a new machine learning algorithm, is that it does not need to calculate the specific form of nonlinear mapping by introducing the kernel function. It only needs to optimize the kernel function of the mapping by using the point product conversion. Based on arbitrary distribution samples, kernel function is proved to be the best application in the actual situation, because its corresponding characteristic space is infinite dimension, so it can be linearly separable after mapping.

$$\mathrm{k}(x_i, \mathrm{x}) = \exp\left\{-\frac{\|x - x_i\|^2}{\sigma^2}\right\}$$

In the formula, we can see that the unknown parameter is σ. The parameters mainly affect the complexity of the sample data in the high dimensional feature space after mapping. In support vector machine algorithm, there is another parameter, that is, error penalty factor. Its role is to control deviations, in fact, that is to adjust the proportion of empirical risk and confidence range. The greater the value of the value is, the greater the loss of the target function, and the greater the loss of the target function. This time it is not willing to eliminate the deviation point, which makes the loosening of her variable as small as possible, which will cause more emphasis on the complexity of the model of the training data and may lead to overlearning; but the value of the too small, emphasizes the model complex. The minimization of the complexity will make the model too simple and lead to the phenomenon of less learning.

4.2. Optimization of parameter c

Genetic algorithm is a computational model that simulates natural evolution process and searches for optimal solution. This unique algorithm needs to be completed in the coding space, so it is necessary to convert the original problem into the coding space to replace the parameter space of the problem, and then to implement the selection process. The selection process is often realized through the evaluation of the design fitness function, and the iterative process is finally established through the genetic mechanism. Through continuous crossover and mutation, the individuals of the group are evolved. Finally, the optimal solution is found to satisfy the convergence condition in the last generation population. In the process of simulation of evolution in nature, genetic algorithms are random genetic operations, but it does not mean that the genetic algorithm is a completely random process. In fact, the goal of genetic algorithm is to get individuals who can survive in the existing environment. The goal is based on the historical information that can be used to find the next generation that has a stronger survivability. The historical information of the previous generation will become the basis for the generation of the next generation and output the optimal individual.

After obtaining the corresponding points of coding space, fitness function constitutes the living environment of individuals to measure individual adaptability. The size of fitness is the judgement of its viability in this environment. So, it is very important to design the fitness function reasonably. We should follow the principles of single value, continuity, non-negativity, rationality, consistency, small amount of calculation and universality. Selection, crossover and mutation are the three steps of genetic manipulation. These three steps are the core of the whole genetic algorithm's ability to search. Each operator has different functions. The use of the selection operator is to imitate the survival of the fittest mechanism in nature. The use of the crossover operator is to simulate the reproduction and hybridization mechanism in the heredity. The function of the mutation mechanism is to simulate the mutation phenomenon in the heredity. These three operators can be used as tools for controlling the genetic process. Mutation is the simulation of chromosomal gene mutations in nature, resulting in changes in structure and physical shape. In operation, we often replace the values of chromosomes in individuals and form a new individual. The advantage of this method is that the local search ability of genetic algorithm can be improved based on unchanged population diversity.

5. Conclusions

The prediction of stock price index is the top priority of the researches of stock investors. Machine learning can effectively predict stock price index to help financial investors avoid risks and maximize investment profits. The main conclusions of this paper are as follows:

(1) The fluctuation of stock price index is nonstationary, nonlinear and has low signal noise ratio, which is suitable to be predicted by machine learning.

(2) The whole prediction process of stock index forecasting based on machine learning includes the steps of data acquisition, data pre-process, eigenvector solution and normalization treatment.

(3) The linear mapping method can optimize the parameter σ and the genetic algorithm can optimize the parameter c, which makes the prediction more accurate.

References

[1] Xiong T, Bao Y, Hu Z. Multiple-output support vector regression with a firefly algorithm for interval-valued stock price index forecasting[J]. Knowledge-Based Systems, 2014, 55: 87-100.

[2] Singh P, Borah B. Forecasting stock index price based on M-factors fuzzy time series and particle swarm optimization[J]. International Journal of Approximate Reasoning, 2014, 55(3): 812-833.

[3] Nayak S C, Misra B B, Behera H S. Impact of data normalization on stock index forecasting[J]. Int. J. Comp. Inf. Syst. Ind. Manag. Appl, 2014, 6: 357-369.

[4] Chen Y S, Cheng C H, Tsai W L. Modeling fitting-function-based fuzzy time series patterns for evolving stock index forecasting[J]. Applied intelligence, 2014, 41(2): 327-347.

[5] Wei L Y. A hybrid ANFIS model based on empirical mode decomposition for stock time series forecasting[J]. Applied Soft Computing, 2016, 42: 368-376.

[6] Cervelló-Royo R, Guijarro F, Michniuk K. Stock market trading rule based on pattern recognition and technical analysis: Forecasting the DJIA index with intraday data[J]. Expert systems with Applications, 2015, 42(14): 5963-5975.

[7] Laboissiere L A, Fernandes R A S, Lage G G. Maximum and minimum stock price forecasting of Brazilian power distribution companies based on artificial neural networks[J]. Applied Soft Computing, 2015, 35: 66-74.

[8] Das S P, Padhy S. A novel hybrid model using teaching–learning-based optimization and a support vector machine for commodity futures index forecasting[J]. International Journal of Machine Learning and Cybernetics, 2018, 9(1): 97-111.

[9] Sun B Q, Guo H, Karimi H R, et al. Prediction of stock index futures prices based on fuzzy sets and multivariate fuzzy time series[J]. Neurocomputing, 2015, 151: 1528-1536.

[10] Rubio A, Bermúdez J D, Vercher E. Improving stock index forecasts by using a new weighted fuzzy-trend time series method[J]. Expert Systems with Applications, 2017, 76: 12-20.

[11] Zhang B, Wei Y, Yu J, et al. Forecasting VaR and ES of stock index portfolio: A Vine copula method[J]. Physica A: Statistical Mechanics and its Applications, 2014, 416: 112-124.

[12] Chiang W C, Enke D, Wu T, et al. An adaptive stock index trading decision support system[J]. Expert Systems with Applications, 2016, 59: 195-207.

[13] Barak S, Modarres M. Developing an approach to evaluate stocks by forecasting effective features with data mining methods[J]. Expert Systems with Applications, 2015, 42(3): 1325-1339.

[14] Jothimani D, Shankar R, Yadav S S. Discrete wavelet transform-based prediction of stock index: a study on National Stock Exchange Fifty index[J]. arXiv preprint arXiv:1605.07278, 2016.

[15] Junior P R, Salomon F L R, de Oliveira Pamplona E. Arima: An applied time series forecasting model for the bovespa stock index[J]. Applied Mathematics, 2014, 5(21): 3383.

[16] Rather A M, Agarwal A, Sastry V N. Recurrent neural network and a hybrid model for prediction of stock returns[J]. Expert Systems with Applications, 2015, 42(6): 3234-3241.