Research on the application of statistical learning method in financial data analysis

Mingchen Ye

Tangshan Normal University, Tangshan, China

Keywords: statistical learning, Support vector machine, Application to explore

Abstract: Adding statistical methods to financial data analysis, and using computer Internet big data processing technology, can quickly and effectively process financial information, for the development of the financial industry to bring technical revolution breakthrough, thus can effectively solve the data problems existing in the financial industry.

1. Introduction

There are a large amount of data in the financial markets, with the progress of the society, more and more data in a market economy, seem to be more complete, at the same time these data through the computer can quickly realize the data collection, and to study and explore the high-dimensional complex financial data, is the current financial data analysis to explore the main development trend.

2. The impact of time on financial data

2.1. Make the data have high noise

Due to the complex and unpredictable factors in the financial industry, the financial market is often affected and restricted by various aspects, such as the influence of national policies and the change of people's consumption concept, which may lead to certain noise of time on the financial data[1]. When time produces noise influence on financial data, it will cause great obstacles to the statistical financial data, which not only destroys the regularity of the market, but also makes many uncertain factors appear in the financial time series.

2.2. Make the data unsteady

Time will have a serious impact on financial data. At the same time, due to too many influencing factors, it is easy to lead to the possibility of (unstable) statistical characteristics of relevant data in the whole process. In the process of change without any starting point with time, it may even lead to extremely complex composition and relationship of relevant data in financial time series. Therefore, in the process of establishing data model, stationary data is needed as the main reference and research object.

2.3. Make the data have a certain periodicity

The development of anything has a certain periodicity law, so does the financial field. As a result, in the process of economic development, there may be the same periodicity as in the financial time field, so that there is a certain correlation. However, because the inherent regularity and periodicity are often unforeseeable, it provides many different methods and strategies for the periodicity research of financial serial data [2].

3. Statistical learning and supporting theory

3.1. Statistical learning theory

In the process of further development and application of statistical learning theory, its basic principle is based on the vector machine theory and improves its own processing performance by

  DOI: 10.25236/mmmce.2021.018

integrating the existing knowledge structure. In the process of forecasting and training the data of the preset model, there is a certain basis relationship. The relationship between the two is evaluated and analyzed, and the learning optimization is carried out. Then, the prediction effect of the learning method is counted when the parameters are adjusted and evaluated. In the concept of statistics, F(x,y) is used as the function representation, and N independent training population samples are distributed for the specific composition and possible unknown inside the function, and are represented by (x_1, y_1),(x_2, y_2),(x_3, y_3),...Extract it, and look for a possible relationship between the two F of x, w_0), while estimating the optimal result, the expected risk value R(w) is taken into account and the expected risk value is controlled to a minimum, and the formula is as follows:

$$R(w) = \int L(y, f(x, w))dF(x, y) \quad (1)$$

In the formula, as an independent set of sample prediction functions, and as a single sample function relationship as a prediction function, also known as the learning function, and the formula belongs to the generalized parameters, usually in the expression, can predict the loss of the learning prediction of function

$$Yf(x, w)wL(y, f(x, w))^{[3]} \quad (2)$$

In the approximate substitution of expected risk, the substitution method is used to predict the expected risk of the actual sample information of the statistical learning problem, and the principle of risk minimization is used to help the statistical learning, which will lead to the predicted sample size gradually approaching infinity, so that the statistical learning algorithm can also reach the optimal state. However, in fact, the independent sample size of many statistical learning problems is very small, and it is difficult to reach infinity. When confronted with key problems, there will be some drawbacks, making it difficult to make the best prediction of the location distribution samples outside the independent training samples.

3.2. Support Vector Machine Theory

3.2.1. Basic principle of support vector machine

In general, the basic principle of support vector machine is statistical learning model, and the construction of the model is concentrated in a specific space, so it can form enough interval linear separation. Different from perceptron, the existing interval feature function usually belongs to a part of the model, and is not the nonlinear classifier of support vector machine. As a result, the form of the maximum interval splitter of SVM is linear. After introducing different kernel functions, the nonlinear statistical problems of different degrees can be constructed.

3.2.2. Linearly separable SVM

The linear classifier of support vector machine is usually based on the most basic statistical learning model algorithm. When the support vector machine theory of the linear separable problem adopts two different samples to distribute, the distance between the classification intervals will result in the purpose of classifying samples with different distributions and finally achieving the highest accuracy [4]. Based on this, the distance of the classification interval needs to be kept to the maximum, so that the statistical learning support vector machine model algorithm has a better adaptation space in the number of extrapolated samples. In multiple sample Spaces with different dimensions, in order to obtain the optimal classification line, the solution of the optimal classification plane is deeply explored in the linear separable problem support vector machine model, and the formula w*x+b=0 is satisfied.

4. Forecast based on SVM stock index (Forecast stock index)

4.1. Selection of data and samples

In this study, the historical trading situation of the Shanghai Composite Index on January 2, 2020 and January 17, 2020 is selected as the training version.

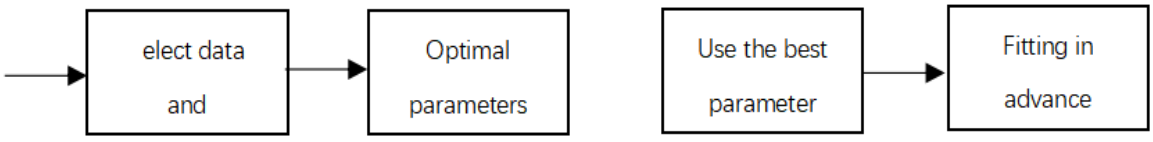
The specific sample data is shown in Figure 1.

time	Early morning	The highest	The minimum	The close	Rise and fall	% increase	Amplitude %	Total hands (ten thousand)
2020.1.2	3066.34	3098.1	3066.34	3085.2	35.08	1.15	1.04	2.92 E + 10
2020.1.3	3089.02	3093.82	3074.52	3083.79	1.41	0.05	0.626	2.61 E + 10
2020.1.6	3070.91	3107.2	3065.31	3083.41	0.38	0.01	1.36	3.13 E + 10
2020.1.7	3085.49	3105.45	3084.33	3104.8	21.39	0.69	0.685	2.77 E + 10
2020.1.8	3094.24	3094.24	3059.13	3066.89	37.91	1.22	1.13	2.98 E + 10
2020.1.9	3082.64	3097.33	3080.13	3094.88	27.99	0.91	0.561	2.43 E + 10
2020.1.10	3102.29	3105.22	3081.4	3092.29	2.59	0.08	0.77	2.1 E + 10
2020.1.13	3091.49	3115.57	3075.38	3115.57	23.28	0.75	1.3	2.11 E + 10
2020.1.14	3120.67	3127.17	3105.6	3106.82	8.75	0.28	0.692	2.3 E + 10
2020.1.15	3103.17	3107.94	3082.04	3090.04	16.78	0.54	0.834	2.02 E + 10
2020.1.16	3095.73	3096.37	3070.88	3074.08	15.96	0.52	0.825	2.03 E + 10
2020.1.17	3081.46	3067.25	3067.25	3075.5	1.42	0.05	0.803	1.9 E + 10

4.2. Model establishment

In the model building, there is a certain correlation between the index data of the previous day and the daily SSE stock index. Therefore, in the model building, the opening price of the day is taken as the dependent variable and can be included in the financial time series model. Usually, after the standardization of data units, the optimal algorithm should be used to select the SVM model as the optimal parameter to finally fit the prediction index of the model, and the corresponding model flow chart can be established, as shown in Figure 2[5].

Figure 2: Flowchart of opening trend forecast of Shanghai Composite Index



4.3. Data selection and processing

In setting up the model of the Shanghai composite index, the choice of different variables can be used as a model fitting data, the highest form of stock trading, the lowest price, the opening price, closing price, volume and total volume data as independent variables, and can according to the trend of time distribution, for the distribution of Shanghai stock index opened daily trend for data selection. In the process of data processing, in order to prevent data overflow, it is necessary to take the way of large value system leading small value to preprocess the data in order to reduce the difficulty of calculation. In the choice of parameters, the stand or fall of parameters selection of SVM in the problems in the financial time series forecasting, can consider the function of internal parameters optimization, the value of each parameter may permutation and combination, and using

cross validation method to evaluate its, can automatically adjust the optimal parameter combination, can achieve the optimal parameter values. The SVM model is trained with the best parameters for fitting prediction. The opening price of the Shanghai Stock Exchange Index and the predicted distribution trend of the opening price can be obtained through the network search method. The model fitting results are obtained through comparison, so the mean square error is generally MSE=2.125e-5 and the correlation coefficient R=99.951%.

4.4. Analysis of research results

Comparing the SSE index with the vector model and the ARMA model with the statistical method, the difference distribution among the three models can be predicted by applying the different model prediction methods to the opening price and forecast of the SSE index. Compared with the original data, the overall prediction of the ARMA model is in a state of failure and does not make any accurate prediction. In the comparison between the prediction trend of the SVM model and the original data, the prediction trend is basically the same as the rise and fall trend of the original data[6]. So can be seen by comparing the two kinds of model has relatively ideal prediction ability, for complex financial time series data, the original ARMA model is not able to determine the true development trend of forecast data, only the support vector machine (SVM) for nonlinear features strong skeleton complex data such as good adaptability, can achieve more accurate forecasts.

5. Conclusion

with the concept of large data into various industries, at the same time, the value of big data represent more and was deeply loved by all people, as a complicated financial market, there are a lot of data problems need to solve, so will the support vector machine (SVM) as the new statistical learning method is gradually applied to financial data research and forecast, can realize the analysis of the different data processing, it also makes the support vector machine (SVM) in the treatment of different financial problems, can form ideal study effect, its theoretical foundation and effective performance, better able to maintain its financial sector towards different directions.

References

[1] Xiao Ziyu. Application of Mathematical Algorithms in Big Data Analysis Technology. Digital World, vol.000, no.008, pp. 146, 2019.

[2] Liu Dongshuang. Exploration and Practice of CDIO Educational Concept in Teaching Financial Data Analysis Course. Forum on Informatization in Education, vol.003, no.004, pp.134-135, 2019.

[3] Liu Y N. Application research of statistical theory in big data analysis. National Business · Theoretical Research, vol.000, no.003, pp.136-137, 2019.

[4] Li C Y. The application of analytical big data in financial research. China Business and Commerce, vol.000, no.012, pp.184-185, 2020.

[5] Li Yanjie. Analysis of the application of big data in financial research. Fortune Today (China Intellectual Property), no.04, pp.46-47, 2020.

[6] Chen Linjiang. Analysis of Data Analysis Feedback Teaching Method -- Taking the Simulation Test Paper Data Analysis in College Entrance Examination Chemistry Review as an Example. Teaching Research, vol.013, no.020, pp.217-218, 2019.