A Forecasting Model of Financial Assets' Price Based on Support Vector Regression

Junsheng Wang[1, 2, 3], Shaozhen Chen[1, 2, 3], Bo Wang[1, 2, 3], Ke Yang[1, 2, 3]

[1]State Grid Electronic Commerce Co., Ltd. Beijing 100053 China

[2]State Grid Xiong'An Fintech Corporation China

[3]Electric Finance and E-Commerce Lab. China

Keywords: gold price, support vector regression, BP neural network

Abstract: Gold is a very important financial asset. This paper briefly describes the price influencing factors in gold forecasting and the basic principles of support vector regression algorithm. The support vector regression algorithm and BP neural network are used to predict the gold price. Finally, we obtain the prediction of support vector regression algorithm. The effect is better than that of the BP neural network. The prediction error of the support vector regression algorithm in the sample is much smaller than the prediction error of the BP neural network algorithm in the sample, which provides guidance for the gold price forecast.

1. Introduction

In recent years, as the price of gold has continued to rise, it has spurred people's enthusiasm for investing in gold. Gold has become an investment and wealth management channel as important as stocks and government bonds. The gold market has the characteristics of high risk and high income, and one of its risks comes from whether it can make a correct analysis and judgment on the price trend of gold. Therefore, it is of great theoretical significance and application value to study how to effectively forecast gold prices. The research methods of gold price trend can be divided into two major schools: fundamental analysis and technical analysis. Fundamental analysis is the analysis of all factors that directly or indirectly affect the relationship between gold supply and demand [1, 2], such as international economic factors, international political environment, the US economy and the trend of the US dollar, international crude oil price trends, etc., thus yielding a probabilistic in conclusion. The technical analysis method is based on the three classic assumptions of the Dow Theory founded by Charles Dow, that is, the price runs in a trend mode; the price reflects everything; the history repeats itself. At present, the most commonly used technical analysis methods are k-line theory, morphological theory, wave theory, and Gann's law.

This paper uses the SVR algorithm for simulation experiments. Many scholars have done research on this. Nello Cristianini [10] and others introduced the support vector machine (SVM) for the first time. The support vector machine is a generation of learning system based on the latest development of statistical learning theory. SVM provides state-of-the-art performance in practical applications such as text categorization, handwritten character recognition, image classification, biological sequence analysis, etc. It has now become one of the standard tools for machine learning and data mining. Chen guo [8] et al. used genetic optimization algorithms to optimize parameters, thereby improving the accuracy of the SVR model. David Lindsay et al. [9] studied standard machine learning techniques (Neural Network, C4.5, K-Nearest Neighbor, Naive Bayes, SVM, and HMM) when testing on time series data sets in various problem domains. Output the validity of the probability prediction. The raw window data is converted to a pattern classification problem using the sliding window method, and the corresponding target prediction is set to some discrete future values in the time series sequence. Their results show that valid probability predictions can be generated on time series data.

In addition, many researchers have analyzed the influencing factors of gold prices [1-2]. The dollar index is often negatively correlated with the price of gold.

S. Zemke [3] analyzed that these two factors have obvious effects on the guidance and shock of the gold price trend. In response to this situation, China's coping strategy was put forward, which is to determine the reasonable scale of gold reserves and to capture benefits from exploration, mining, technology and structure. To a certain extent, these strategies will play a significant role in stabilizing China's economic development and improving international creditworthiness.

Vatsal H. Shah [4] and others analyzed relevant factors that may affect the price of gold, such as the US dollar index, crude oil price, and stock market. Using the organic combination of the neural network model and ARMA to achieve the prediction of the gold price, and finally compare and test the predicted results with the actual data. In addition, based on the established forecasting model and the conclusions obtained, some trading proposals for gold-related derivatives are given.

There are many factors influencing the price of gold. First, there is a strong substitution relationship between gold and the dollar as reserve assets. Once a stronger dollar, as a hard currency status has been further consolidated, it shows that the U.S. economic situation is relatively good, so people will more readily assets investments like stocks and bonds market, this class has a higher yield, and less would consider assets used to store gold, so gold demand will decline, gold prices were falling. Conversely, when the dollar weakens, investors become concerned about the future of the U.S. economy, and more people tend to store gold. Secondly, at present, the price of most gold products in the world is priced in us dollars. Once the us dollar depreciates, due to the exchange rate fluctuations, many foreign investors will be more willing to buy gold as a reserve asset due to the advantages of the exchange rate. Crude oil as a globally recognized strategic material can be said to have penetrated into all aspects of the global political economy, and gold as a quality asset reserve, and crude oil is naturally closely related. On the one hand, crude oil, as a kind of important industrial raw material, with related industry covers all aspects of the political and economic life, so to speak, when oil prices rise, so the cost of the related industry will also rise, making the price of many products, the production cost of living increase will inevitably lead to inflation, at this point, the gold as a store of value is a hard currency will be more valued by people, makes the demand for gold increases, the rising value of their cause. On the other hand, the current global crude oil price is mainly priced in dollars, and the U.S. oil consumption ranks among the top in the world. Once the crude oil price fluctuates, it is bound to affect the trend of the dollar, and the rise and fall of the dollar will lead to the ups and downs of the gold price. Thus, the international crude oil price on the gold price is very significant. Normally, when the stock market falls, gold prices rise. This is because the stock market and the gold market are two very different styles of the field, the former aggressive, and the latter solid conservative. When the stock market is strong, investors are generally optimistic about the economic outlook, so they will focus their capital on the stock market to pursue higher returns. On the contrary, when the stock market is in a downturn and investors are generally concerned about the future economic situation, they will be more inclined to put money into the gold market to achieve the purpose of maintaining stability.

As a physical commodity, the impact of supply and demand changes on the gold price is more obvious. Therefore, to judge the medium-term trend of gold prices, we must start from the supply and demand of physical gold analysis. On the supply side, the output of mineral gold is not very sensitive to the change of external factors, and the supply elasticity is very small and relatively stable, while the official sales and recycled gold are closely related to the price of gold. Their increase and decrease dominate the supply of physical gold, which affects the price of gold. Therefore, to judge the change in the supply of physical gold, the amount of official gold sales and the amount of recycled gold is the main factor. In terms of demand, the demand of gold jewelry and manufacturing industry is mainly affected by the economic development of countries and regions. The main influencing factors of gold investment demand are the trend of some competitive investment returns, such as the us dollar exchange rate, standard & poor's index, federal funds rate, and oil price [3]. This article selects the official price of gold lag phase 1, the dollar index, the CRB commodity index, the CRB index, a metal net gold ETF holdings, brent crude oil spot prices during the month, silver spot prices during the month, yields on the ten-year United States Treasury note, the core CPI (year-on-year), U.S. non-manufacturing PMI, the M1, M2, the VIX, the us national financial conditions index, oil ratio of gold, gold and silver. As the dependent variable, support vector regression analysis was used.

2. Support vector regression algorithm

The basic idea of SVM is to map the training data from the input space to the high-dimensional feature space through the function f, and then construct a separate hyperplane with the largest margin in the feature space.

Give a training data set $x_i \in R^n, i = 1, \cdots, l, l$ corresponds to the size of the training data, $y_i = \pm 1$ is a classification label. SVM will find a hyperplane direction ω, and a compensation scalar b, making positive samples $f(x) = \omega * \Phi(x) + b \geq 0$ and for negative samples $f(x) = \omega * \Phi(x) + b \leq 0$. Therefore, although we can't find a linear function in the input space to determine the type of given data, we can easily find the best hyperplane that can clearly distinguish between the two types of data.

Consider a training data set $\{(x_1, y_1), (x_2, y_2), (x_3, y_3), \cdots, (x_l, y_l)\}$, and $x_i \in R^n$ annotated samples of the input space and have a corresponding target value $y_i \subset R$, where $i = 1, \cdots, l$, l is the size of the training data. The idea of a regression problem is to determine a function that can accurately approximate future values.

The general SVR estimation function takes this form

$$f(x) = (\omega * \Phi(x)) + b \tag{1}$$

Where $\omega \subset R^n$, $\mathrm{b} \subset \mathrm{R}$, Φ is a nonlinear transformation form in high dimensional space, Our goal is to find the values of ω and b so that x minimizes the risk of regression.

$$R_{reg}(f) = C \sum_{i=0}^{l} \Gamma(f(x_i) - y_i) + \frac{1}{2}\|\omega\|^2 \tag{2}$$

$\Gamma(\cdot)$ is loss function, C is a constant, Vector ω can be written as a series of data points

$$\omega = \sum_{\mathrm{i}=1}^{\mathrm{l}} (\alpha_\mathrm{i} - \alpha_\mathrm{i}^*)\, \Phi(\mathrm{x_i}) \tag{3}$$

Bring (3) into (1), the equation can be written as

$$f(\mathrm{x}) = \sum_{\mathrm{i}=1}^{\mathrm{l}} (\alpha_\mathrm{i} - \alpha_\mathrm{i}^*) \left(\Phi(\mathrm{x_i}) \cdot \Phi(\mathrm{x})\right) + \mathrm{b}$$
$$= \sum_{\mathrm{i}=1}^{\mathrm{l}} (\alpha_\mathrm{i} - \alpha_\mathrm{i}^*) \left(\mathrm{k}(\mathrm{x_i}, \mathrm{x})\right) + \mathrm{b} \tag{4}$$

In (4), the dot product is written as $\mathrm{k}(x_i, x)$ labeled as a kernel function. The kernel function enables the implementation of a dot product in a high dimensional feature space using low dimensional spatial data input without knowing the transformationΦ. All kernel functions must satisfy Mercer's condition, which corresponds to the inner product of some feature spaces. RBF is usually used as the kernel of the regression

$$\mathrm{k}(\mathrm{x_i}, \mathrm{x}) = \exp\{-\gamma \lceil \mathrm{x} - \mathrm{x_i} \rceil^2\} \tag{5}$$

Some common kernels are shown in Table 1. In our research, we have experimented with these three kernels.

Table 1. Kernel experiment

Kernel	Function
Linear	$\mathrm{x} * \mathrm{y}$
Polynomial	$[(x * x_i) + 1]^d$
Radial Basis Function	$\exp\{-\gamma \lvert x - x_i \rvert^2\}$

The ε-insensitive loss function is most widely used for loss functions (5). The form of the function is:

$$\Gamma(f(x) - y) = \begin{cases} |f(x) - y| - \varepsilon, \ |f(x) - y| \geq 0 \\ 0, \ otherwise \end{cases} \tag{6}$$

By solving the quadratic optimization problem, the regression risk and the ε-insensitive loss function (6) in (2) can be minimized to

$$\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)\, k(x_i, x_j) - \sum_{i=1}^{l}\alpha_i^*(y_i - \varepsilon) - \alpha_i(y_i + \varepsilon)$$

Subject to

$$\textstyle\sum_{i=1}^{l}(\alpha_i^* - \alpha_i) = 0\ ,\ \alpha_i^*, \alpha_i \in [0, C] \tag{7}$$

Lagrange multiplier α_i and α_i^* represents the solution to the quadratic problem as the power to predict the target value. Only the non-zero value of the Lagrange multiplier in (7) can be used to predict the regression line and is called the support vector. For all points of ε, the Lagrange multiplier equals zero and does not contribute to the regression function.

As long as $|f(x) - y| \geq \varepsilon$ needs to be satisfied, the Lagrange multiplier may be a non-zero value and used for the support vector.

The constant C introduced in (2) determines the penalty for the estimation error. Larger C impose higher penalties on errors in order to train regressions to minimize errors and reduce generalization, while small C penalize fewer errors. This allows for minimizing the margin with errors and thus a higher generalization capability. If C reaches the maximum, the SVR will not allow any errors and lead to complex models, while C will become a lot of errors when the result becomes 0 and the model will be less complicated.

Now, we have used the Lagrange multiplier to solve the value ofw. For the variable b, it can be calculated by applying the Karush-Kuhn-Tucker (KKT) condition, in which case this means that the product of the Lagrange multiplier and the constraint must equal 0,

$$\begin{gathered} \alpha_i(\varepsilon + \zeta_i - y_i + (\omega, x_i) + b) = 0 \\ \alpha_i{}^*(\varepsilon + \zeta_i{}^* + y_i - (\omega, x_i) - b) = 0 \end{gathered} \tag{8}$$

And

$$\begin{gathered} (C - \alpha_i)\zeta_i = 0 \\ (C - \alpha_i{}^*)\zeta_i{}^* = 0 \end{gathered} \tag{9}$$

Where ζ_i and $\zeta_i{}^*$is slack variables, for $\alpha_i, \alpha_i = 0$, and $\zeta_i{}^* = 0$, $\alpha_i{}^\alpha \in (0, C)$,b can be compute as

b = $y_i - (\omega, x_i) - \varepsilon$ Where $\alpha_i \in$ (0, C)

b = $y_i - (\omega, x_i) + \varepsilon$ Where $\alpha_i^* \in$ (0, C)

In summary, we can use SVM and SVR without knowing the transformation. We need to try kernel functions; PenaltyC, which determines the penalty for estimating error; and the radiusε, it determines that the data within ε is ignored in the regression.

3. Results analysis

This article USES data from April 2017 to December 2018 to make monthly forecasts. SVR and BP neural network are used for prediction, and the prediction accuracy indexes are defined as follows:

$$P_{ER} = \frac{1}{N}\sum_{k}^{N}\left(y(k) - y_p(k)\right)^2$$

Where $y_p(k)$the predicted value, is$y(k)$ is the true value, and N is the predicted sample number. See table 2

Table 2. Precision index

Error	SVR	BP neural network
Prediction error P_{ER}	6,586.788	1,183.275

It can be seen that the accuracy of the support vector machine fitting and prediction is significantly better than the other two methods. The fitting situation is shown in Figure 1.

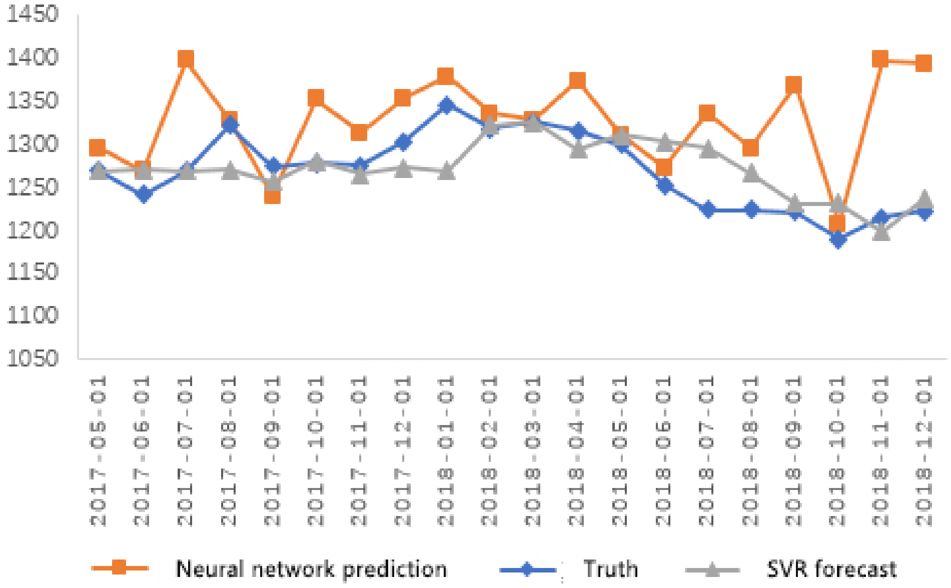

Figure 1. The fitting situation is shown in the figure

4. Conclusion

In the above results, we can see that the SVR method avoids many defects, and also obtains good results in prediction. The prediction effect is better than BP neural network in the sample. The prediction error of the support vector regression algorithm in the sample is much smaller than that of BP neural network. The prediction error of the network algorithm in the sample. Next we will consider how to improve the effectiveness of the SVR method.

References

[1] W. Huang et al., “Forecasting stock market movement direction with support vector machine,” Computers & Operations Research, 32, pp. 2513 – 2522005, 2005.

[2] J. Moody, et al., “Learning to trade via direct reinforcement,” IEEE Transactions on Neural Networks, vol. 12, no. 4, Jul. 2001.

[3] S. Zemke, “On developing a financial prediction system: Pitfall and possibilities,” Proceedings of DMLL-2002 Workshop, ICML, Sydney, Australia, 2002.

[4] Vatsal H. Shah, “Machine learning techniques for stock prediction,” www.vatsals.com. [5] Wu, Chun-Hsin & Ho, Jan-Ming & Lee, D. (2005). Travel-Time Prediction with Support Vector Regression. Intelligent Transportation Systems, IEEE Transactions on. 5. 276 - 281. 10.1109/TITS.2004.837813.

[5] Lin P, Su S, Lee T. Support vector regression performance analysis and systematic parameter selection [J]. IEEE, 2005, 2:877 - 882.

[6] Wang peng, Zhu xiaoyan. Model choice and application of SVM basing on RBF. Computer Engineering and Applications, 2003, (24).

[7] Chen guo. SVR time series model optimization basing genetic algorithm. Chinese Journal of Scientific Instrument, 2006, (9).

[8] David Lindsay, Sian Cox. Effective Probability Forecasting for Time Series Data Using Standard Machine Learning Techniques. ICAPR 2005, LNCS 3686, pp. 35–44, 2005. [7] Xuedong Wang, Haoran Zhang. Time Series Prediction Using LS-SVM with Particle Swarm Optimization. ISNN 2006, LNCS 3972, pp.747 – 752, 2006.

[9] Nello Cristianini, John Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Publishing House of Electronics Industry. 2000.