2017 International Conference on Mathematics, Modelling and Simulation Technologies and Applications (MMSTA 2017)
ISBN: 978-1-60595-530-8

GARCH-LSSVM Coupled Predication Model and Its Application on Stock Index Forecasting

Xiao-xu HU[1,2], Meng-qi ZHANG[1,2] and Xin JIANG[1,3,*]

[1]Key Laboratory of Mathematics, Informatics and Behavioral Science, Ministry of Education, Beihang University, Beijing 100191, China;

[2]School of Mathematics and Systems Science, Beihang University, Beijing 100191, China;

[3]Beijing Advanced Innovation Center for Big Data and Brain Computing, Beijing 100191, China

*Corresponding author

Keywords: Generalized auto-regressive conditional heteroscedastic (GARCH), Least square support vector machine (LSSVM), Coupled prediction algorithm, Stock index forecasting, Technical indicators.

Abstract. It is usually a challenge for traditional time series prediction models to combine the linear and non-linear factors effectively, which might cause a problem that the trend forecast is not accurate. In this paper we propose a new prediction model based on GARCH and LSSVM. The GARCH model is used to deal with the heteroscedasticity of the residual series of the closing price of the stock index data. At the same time, a number of technical indicators are constructed to train the LSSVM model and the corresponding predicted values and residuals are obtained. Further, the predicted value and the residual obtained respectively from GARCH and LSSVM are used as the training set to modify the LSSVM model, with the predicted value of the logarithmic closing price of the stock index obtained. This coupled model not only contains the linear trend of historical information, but also integrates the non-linear features such as market volatility information which is closely related to the target data. Numerical experimental results show that the prediction accuracy of the model is 98.22% on the testing data set. This model performs also better than the GARCH 97.84% and the LSSVM 98.04% respectively in accuracy deviation.

Introduction

At present, how to build an efficient quantitative model to predict the price trend of securities has become a challenging and valuable work [1]. For the government macro management, accurately predicting the stock price index and stock price trend in the stock market can monitor and guide the smooth operation of the stock market at a certain level and reduce the market risk. For investors, building a quantitative model, has a good reference value for its effective investment strategy.CSI300 index is from Shanghai and Shenzhen two stock market 300 A shares as a sample compiled from the stock index and the stock market, as the representative of good, high liquidity, active trading ma-instream investment stocks, earnings can reflect the mainstream market investment. Therefore, the accurate prediction of CSI300 index has higher practical significance.

For the prediction of time series, scholars have proposed many prediction models, including linear models and nonlinear models. Generalized auto-regressive conditional heteroscedastic (GARCH) is a linear regression model specially designed for financial data. Ferenstein and Gasowski (2004) used AR-GARCH to model stock data [2]. Chen and Hu (2011) studied traffic flow forecasting by means of ARIMA-GARCH model [3]. But GARCH model is just a linear model that cannot capture nonlinear components.

Nonlinear model support vector machine (SVM) proposed by Vapnik have shown some advantages in time series prediction. Based on the principle of structural risk minimization, this method is supposed to construct the optimal hyperplane in the feature space, so that the learner can get the global optimal solution. However, the training of SVM needs to solve quadratic program problem, which slows down its calculation speed. In order to improve the speed of learning, Suykens[4]proposed least square SVM(LSSVM) approach, which used the least square method to transform the quadratic program problem into linear equations. In this paper, we construct a GARCH-LSSVM coupled optimization model and establish the verification performance index, which combines the regularization parameter and the kernel parameter in the model to achieve the prediction of the CSI300 index price.

Basic theory

GARCH Model

Most financial time series do not have the independent normal distribution characteristic, instead, they display peak thick tails, the negative bias and the characteristic clustering. In this case, the traditional econometric model is difficult to describe its fluctuation law, and the conditional heteroscedasticity model is needed. The ARCH model is considered as the most widely used conditional heteroscedasticity model that reflects the variance characteristics and is widely used in time series analysis of financial data. The mathematical idea of the ARCH model is as follows

$$Y_t = \beta X_t + \varepsilon_t \tag{1}$$

$$\sigma_t^2 = a_0 + a_1 \varepsilon_{t-1}^2 ... + a_q \varepsilon_{t-q}^2 + \eta_t, \quad t = 1,2,3... \tag{2}$$

where Y_t is the explained variable, X_t is the explanatory variable, ε_t is the residual sequence.

The ARCH model usually analyzes the stochastic perturbation terms of the subject model in order to extract the information from the residuals adequately, so that the final model residual η_t become white noise sequence. If the square of the error is represented by the ARMA model, the ARCH model is transformed into the GARCH model (Bollerslev, 1986), and the GARCH(p,q) model is

$$Y_t = \beta X_t + \varepsilon_t \tag{3}$$

$$\sigma_t^2 = \omega_0 + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 . \tag{4}$$

Here $\sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 = 1$ is restriction.

LSSVM Model

LSSVM is an improved model of SVM, it changes the SVM inequality constraints into equality constraints. The error square and loss function are regarded as the empirical loss of the training set.so, the solution of the quadratic program is transformed into solving linear equations, improving calculation speed and convergence precision, reducing the difficulty of solving[5,6].

Given a training data set $\{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in R_n$ is the input of the system, $y_i \in R$ is the output of the system. The least squares support vector machine mathematical model is:

$$\min_{\omega,b,\xi} \; J(\omega, \xi) = \frac{1}{2}\omega^T \omega + \frac{C}{2}\sum_{i=1}^{n} \xi_i^2 \tag{5}$$

The constraints in the formula are

$$y_i = \omega^T \varphi(x_i) + b + \xi_i; i = 1,2,...,n \tag{6}$$

$\varphi(*)$ is a nonlinear mapping which maps x_i from the input space are mapped to high dimensional (even infinite dimensional) feature space, so as to realize the nonlinear regression in the input space into linear regression in high-dimensional feature space. By constructing the Lagrange function, the constraint optimization can be transformed into unconstrained optimization, and the optimization problem can be transformed into solving linear equations according to the KKT condition [7.8]:

$$\begin{pmatrix} 0 & I_v^T \\ I_v & H + c^{-1}I \end{pmatrix} \begin{pmatrix} b \\ a \end{pmatrix} = \begin{pmatrix} 0 \\ y \end{pmatrix} \tag{7}$$

where $y = [y_1, y_2, ..., y_n]^T$, $I_v = [1,1,...,1]^T$, $a = [a_1, a_2, ..., a_n]^T$, $H = \{H_{ij} \mid i, j = 1,2,..,n\}$, $H_{ij} = \varphi(x_i)^T \varphi(x_j) = K(x_i, x_j)$, $K(*)$ is a kernel function.

Coupling Model

This paper uses CSI300 index data from wind database (http://www.wind.com.cn) for empirical analysis, the time ranges from January 5, 2015 to June 30, 2017, with 607 samples. Each data sample includes five characteristics: the opening price, the highest price, the lowest price, volume, and amount. Select the above characteristics to construct seven indexes, which have strong correlations with closing price as the input vectors, the specific meaning is as follows. The output variable Y is the logarithmic closing price of the CSI300 index [9].

(1) VoluIndex: The ratio of today's trading volume to the average daily trading volume of M days

(2) CloseIndex: The ratio of today's trading close to the average daily trading close of M days

(3) TrAmIndex: The ratio of today's trading amount to the average daily trading amount of M days

(4) MeanIndex: The average closing price for the past M days

(5) RateChangeIndex: The ups and downs of past M days

(6) HighIndex: The highest closing price of past M days

(7) Lowdex:The lowest closing price of past M days

We apply window-type mobile standardization, which only uses the previous i-1 day mean and standard deviation to standardize the i-th day data:

$$x_i^{'} = \frac{x_i - \overline{x_i}}{\sqrt{Var(x_i)}}, i = 1,2,..., \tag{8}$$

where $\overline{x_i}$, $\sqrt{Var(x_i)}$ represents the mean and standard deviation of the sample starting point to the i-th day accordingly.

We first describe the CSI300 index trend in a given time period. It is shown that the trend and distribution of the index price have obvious clustering effect. The distribution of the histogram is characterized by peak and thick tail. Since there is a clear heteroskedasticity, we can establish a GARCH model for testing [10].

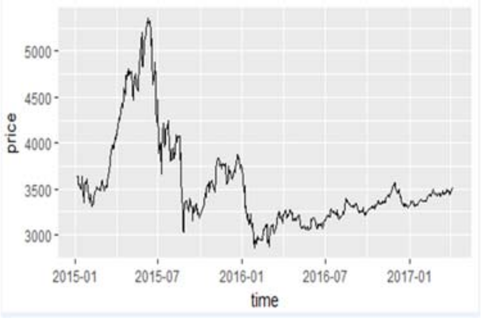

Figure 1. CSI300 index trend.

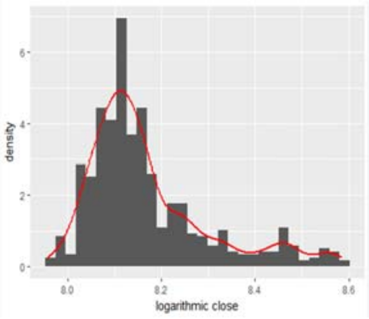

Figure 2. CSI300 index distribution.

Before the application of the model, we carry out the corresponding statistical analysis on the feasibility of establishing the GARCH model forecasting for the index closing price series [11,12].

1. Smoothness test. After the unit root ADF test with R language, it shows that the logarithmic closing price of the CSI300 index is at a significant level of 1% of the test P-value of 0.249, greater than 0.01, indicating a non-stationary sequence. The P-value of the logarithmic yield sequence of the CSI300 index is much less than 0.01 and the probability that the logarithmic yield sequence has a unit root is almost zero. There should be no obvious memory and the persistence of volatility for the smooth sequence. We set up the GARCH model for the logarithmic yield sequence of the CSI300 index.

2. GARCH effect test. The ARCH-LM test of the logarithmic yield series of the CSI300 index indicates that the coefficients of the residual autoregressive function are distinct from zero, and the sequence still has significant autocorrelation, i.e., ARCH effect. In summary, the logarithmic yield series of CSI300 has obvious fluctuation aggregation, stable sequence, significant ARCH effect, and can be modeled by GARCH.

3. GARCH model results output. We use the ugarchspec function of rugarch package in R language to establish the GARCH model and optimize the CSI300 index logarithmic yield sequence, the ugarchfit function to fit the GARCH model. The predicted results show that the P-value of the output model coefficients is significantly zero. The model sGARCH (1, 1) can be well fitted to the historical logarithmic yield series of the CSI300 index. The model is as follows:

$$y_t = 0.201 y_{t-1} - 0.992 y_{t-2} + \varepsilon_t - 0.206 \varepsilon_{t\text{-}1} + 0.991 \varepsilon_{t\text{-}2} \tag{9}$$

$$\sigma_t^2 = 0.062 \varepsilon_{t-1}^2 + 0.938 \sigma_{t-1}^2 \tag{10}$$

4. Model diagnosis. As can be seen from the histogram, the residual of the yield sequence satisfies the normal distribution. The Box-Ljung test shows that the autocorrelation values did not go beyond the significant (confidence) boundary in the lagged 1-20 order, and the P-value of the Box-Ljung test is 0.2619, indicating that the residuals are essentially white noise sequences. Therefore, the GARCH model is valid.

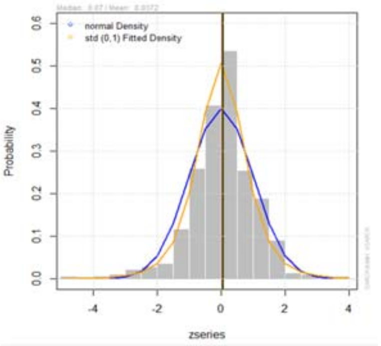

Figure 3. Residual normality test.

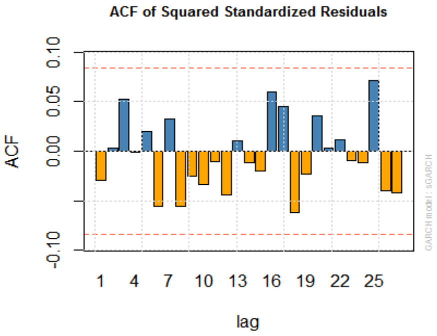

Figure 4. Residual autocorrelation test.

The kernel function of the LSSVM model is a radial basis function:

$$K\left(x_i, x_j\right)=\exp\left[-\| x_i - x_j \|^2/\left(2\sigma^2\right)\right] \tag{11}$$

a and b in the linear equation (7) can be obtained by least squares, the linear regression function is:

$$f\left(x\right)=\sum_{i=1}^{n} a_i \mathrm{K}\left(x, x_i\right)+b \tag{12}$$

where n is the number of input samples, x is the input variable, a_i is not equal to zero. The input sample x_i is the support vector, and a is support vector coefficient [13].

10% cross validation method is selected when parameter optimization is done in LSSVM model. The data set is randomly divided into 10 disjoint sub-datasets, of which seven sub-datasets contains 61 observations, 3 sub-datasets contains 60 observations. In the subsequent 10 trials, each sub-datasets is taken as a test sample for each test and samples from the remaining nine sub-datasets are used as training samples. First, given the kernel parameters, we minimize the RMES to select the regularization parameter C in a certain range with a RBI method. According to the selected regularization parameters, using the RBI method to minimize the RMES to select the kernel parameter sigma, we select the last parameter, which is the optimal combination of parameters, to test the test sample [14, 15].

The GARCH-LSSVM coupling prediction algorithm is as follows.

Input: training sample set $\left(x_i, y_i\right)_{i=1}^{l}$, where $y_i \in R$ is the log closing price, $x_i=\left(x_{i,1}, x_{i,2}, ..., x_{i,N}\right)$ is the eigenvalue of the discrete feature set $A=\left(a_1, a_2, ..., a_N\right)$ associated with the target y_i and the input $x_{t+1}=\left(x_{t+1,1}, x_{t+1,2}, ..., x_{t+1,N}\right)$ $\left(N = 7\right)$.

Output: forecasting result y_{t+1}.

Step 1. The historical opening price, the highest price, the lowest price, the volume and the amount of the sample CSI300 index are combined to obtain seven indicators with high correlation with the index closing price, which is denoted as discrete numerical feature set A.

Step2. Normalize the feature set A by window movement, i.e., $x_i^{'}=\frac{x_i-\overline{x_i}}{\sqrt{Var\left(x_i\right)}}, i=1,2,...$ the CSI300's closing price sequence is logarized to get the sequence y_i.

Step3.The GARCH model is established for the logarithmic yield series of the sample CSI300 index, and then the GARCH model is used to obtain the fitted value of the logarithmic closing price series, noting as Y_1, according to formula $e^{'} = y_i - Y_1$, we get the residual sequence, noting as r_1.

Step4.Train the LSSVM on the training set, use the radial basis function, and select the optimal parameters of the LSSVM with the 10-fold cross validation method. Where the eigenvector Is 7 dimension input variables standardized. That is, the result of the standardization is highly correlated with the closing price.

Step5.In the trained LSSVM model to get the fitting results of training set, noting as Y_2.according to formulae, we get the residual sequence, noting as r_2.

Step6.the fitting numbers Y_1, Y_2 and the residual value sequences r_1, r_2 can be obtained from step3 and step5 as the input variables of the LSSVM model, logarithmic closing as an output variable, training model.

Step7.Add $x_{t+1}=\left(x_{t+1,1}, x_{t+1,2}, ..., x_{t+1,N}\right)$ into step3 and step5 respectively, we get the fitting values y_{t+1}^1, y_{t+1}^2,the residual values r_{t+1}^1, r_{t+1}^2.

Step8.The result of step 7 is taken into the model obtained by the Step 6 training to obtain y_{t+1}

Coupling model fitting results are shown in the figure.

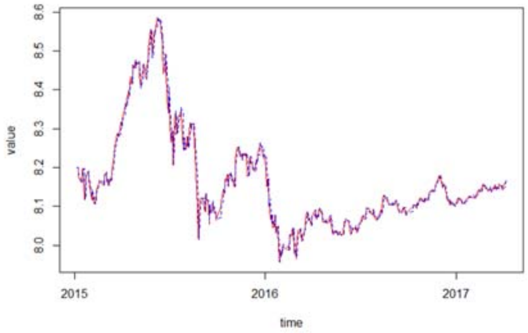

Figure 5. Coupling model fitting effect diagram.

The indicators used here to evaluate the performance of different predictive models are Root Mean Square Errors (RMSE),Relative Square Error(RSE) and Rsquare. Different model performance indicators as shown below, we can see that the performance of the coupling model is superior to a single GARCH model and LSSVM model.

Table 1. Comparison of model performance.

Model	Insample			Outsample		
	RMSE	RSE	Rsquare	RMSE	RSE	Rsquare
GARCH	0.0191	0.0216	0.9784	0.0267	0.0938	0.9062
LSSVM	0.0182	0.0196	0.9804	0.0218	0.0803	0.9197
GARCH- LSSVM	0.0174	0.0178	0.9822	0.0209	0.0607	0.9393

Conclusion

Generally, GARCH model cannot effectively analyze the core nonlinear features of time series, in order to solve the problem, we study the nonlinear deviation of GARCH based on the LSSVM model with strong nonlinear predictive ability and generalization ability, and then train the model of the prediction of non-linear volatility pattern recognition. Meanwhile, the GARCH-LSSVM coupled forecasting model is proposed, and the over-fitting of non-linear terms by heteroscedasticity of the GARCH model is corrected. How to consider more core factors in the process of model training and further optimize the related parameters of LSSVM is the focus and direction of the next step.

Acknowledgments

This work is supported by NSFC (Grants No. 11201017, 11290141) and National Key Research and Development Program of China (Grant No. 2017YFB0701702).

References

[1] Peng T., Tang Z., A small scale forecasting algorithm for network traffic based on relevant local least squares support vector machine regression model, J. Appl. Math 2015, 9, 653–659.

[2] E. Ferenstein, M. Gasowski, Modelling stock returns with AR-GARCH processes, J. SORT-Statistics and Operations Research Transactions, 2004, 28(1): 55-68.

[3] Chen C., Hu J., Meng Q., et al. Short-time traffic flow prediction with ARIMA-GARCH model, C. Intelligent Vehicles Symposium (IV), IEEE, 2011: 607-612.

[4] J.A.K. Suykens, J. Vandewalle, Least squares support vector machine classifiers, J. Neural Processing Letters, 1999, 9(3): 293-300.

[5] Guo W., Zhao Z. A Novel Hybrid BND-FOA-LSSVM Model for Electricity Price Forecasting, J. Information, 2017, 8(4): 120.

[6] Chung S.S., Zhang S. Volatility estimation using support vector machine: Applications to major foreign exchange rates, J. Electronic Journal of Applied Statistical Analysis, 2017, 10(2): 499-511.

[7] Zhang Y., Dong Z., Liu A., et al. Magnetic resonance brain image classification via stationary wavelet transform and generalized eigenvalue proximal support vector machine, J. Journal of Medical Imaging and Health Informatics, 2015, 5(7): 1395-1403.

[8] Wu Q., Peng C. Wind power generation forecasting using least squares support vector machine combined with ensemble empirical mode decomposition, principal component analysis and a bat algorithm, J. Energies, 2016, 9(4): 261.

[9] Zhu B., Han D., Wang P., et al. Forecasting carbon price using empirical mode decomposition and evolutionary least squares support vector regression, J. Applied Energy, 2017, 191: 521-530.

[10] Lai L., Liu J. Support Vector Machine and Least Square Support Vector Machine Stock Forecasting Models, J. Computer Science and Information Technology, 2014, 2(1): 30-39.

[11] J. Contreras, Y.E. Rodríguez. GARCH-based put option valuation to maximize benefit of wind investors, J. Applied Energy, 2014, 136(C): 259-268.

[12] Badescu, R. J. Elliott, J.P. Ortega, Non-Gaussian GARCH option pricing models and their diffusion limits, J. European Journal of Operational Research, 2015, 247(3): 820-830.

[13] Ou P., Wang H. Financial Volatility Forecasting by Least Square Support Vector Machine Based on GARCH, EGARCH and GJR Models: Evidence from ASEAN Stock Markets, J. International Journal of Economics & Finance, 2010, 2(1): 337-367.

[14] Si X.S., Hu C.H., Yang J.B., et al. A New Prediction Model Based on Belief Rule Base for System's Behavior Prediction, J. IEEE Transactions on Fuzzy Systems, 2011, 19(4): 636-651.

[15] Zhang J.L., Zhang Y.J., Zhang L. A novel hybrid method for crude oil price forecasting, J. Energy Economics, 2015, 49: 649-659.