Support Vector Regression for Prediction of Stock Trend

Yaqing Xia [1], Yulong Liu [2], and Zhiqian Chen [3]
[1] School of International Trade and Economics, Central University of Finance and Economics, Beijing, China
[2] Department of Math, Ohio State University, Columbus, OH, USA
[3] Department of Software Engineering, Peking University, Beijing, China
yaqing.jane.xia@gmail.com, liu.2874@osu.edu, imczq@pku.edu.cn

Abstract— **Prediction of the trend of the stock market is very crucial. If someone has robust forecasting tools, then he/she will increase the return on investment and can get rich easily and quickly. Because there are a lot of factors that can influence the stock market, the stock forecasting problem has always been very complicated. Support Vector Regression is a tool from machine learning that can build a regression model on the historical time series data in the purpose of predicting the future trend of the stock price. In this paper, we present a theoretical and empirical framework to apply the Support Vector Regression (SVR) strategy to predict the stock market. Our results suggest that SVR is a powerful predictive tool for stock predictions in the financial market.**

Keywords-stock prediction; data mining; support vector regression; forecasting

I. INTRODUCTION

The financial market is a complex, evolutionary, and nonlinear dynamical system. The field of financial forecasting is characterized by data intensity, noise, non-stationary, unstructured nature, high degree of uncertainty, and hidden relationships [1]. Over the past decade, neural networks have been successfully used for modeling financial time series. A large number of successful applications have shown that ANN can be a very useful tool for time series modeling and forecasting. However, ANN has a difficult in explaining the prediction results due to the lack of explanatory power, and suffers from difficulties with generalization due to overfitting.

Based on the statistical learning theory, support vector machine (SVM) has become a hot topic of study due to its successful application in classification and regression task. It is a very specific type of learning algorithms characterized by the capacity control of the decision function, the use of the kernel functions and the sparsity of the solution. This paper aims to use support vector machine in regression task for predicting the financial time series.

The rest of this article is organized as follows. In Section 2, we will show the difficulty in winning money from the stock market by elementary analysis on stock market, in which the underlying properties in the stock market are discussed, such as stock randomness, zero expected return, and pink noise. In Section 3, we will give brief theoretical overviews of SVR. In Section 4, we focus on the application of SVR on the stocks market. Section 5 presents conclusion of our paper.

II. ELEMENTARY ANALYSIS ON STOCK MARKET

A. Random and Zero Expected Return

With the data of Standard & Poor's index (S&P 500) 2006-2011, which in a certain extent reflects the performance of the whole stock market in US, we plot the daily level change of the S&P 500 index over time in Figure 1. From Figure 1, we can hardly see any pattern of the trend of the stock market, which means that it is very hard to predict its trend.

Also we calculate the daily return rate by computing the derivative of the daily level (price) of the S&P 500 index. The return rate is the ratio of money gained or lost on an investment relative to the amount of money invested, which is a significant measure the investors most care about. Figure 2 (a) shows the fluctuation of the return rate of S&P 500 over the time from Jan 2006 to Jun 2011. It fluctuates around the return rate of zero and randomly goes up and down. The stock return histogram in Figure 2 (b) shows the distribution of stock return. The distribution is close to the normal distribution but has a high peak in the return rate of zero, which means that the expected return is 0, giving us a direct conclusion that it is very hard to win money in stock market if we can't hold some essential information to predict the change. Random investments just give you a zero even negative return. Also, stocks that have a larger variance have a better change of positive return, but it also means that risk goes with potential profit.

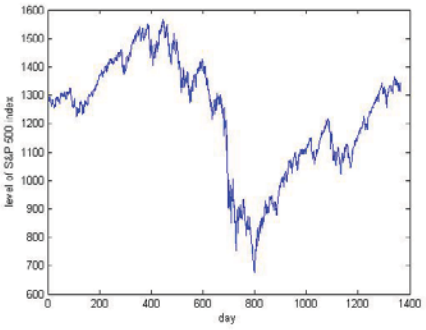
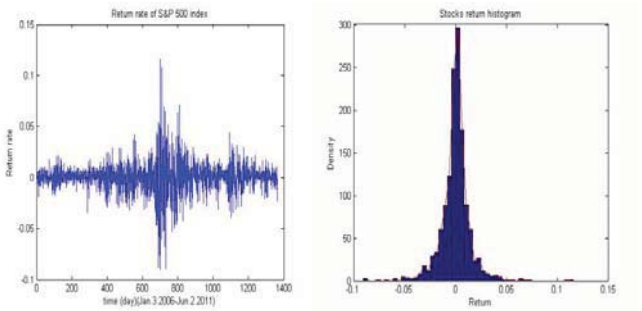


Figure 1. Time series data (S&P 500)

978-1-4799-0245-3/13/$31.00 ©2013 IEEE



(a) (b)

Figure 2 (a) Fluctuation of daily return rate and (b) stock return histogram

B. *Pink Noise*

Random noise having equal energy per octave, and so having more low-frequency components than white noise. It is defined as noise having a power spectrum that decreases like *1/f*. That is, the *f*-th Fourier coefficient of the noise has complex absolute value approximating $a \cdot \frac{1}{f} + b$, where a and *b* are constant. We can easily find pink noise pattern in the stock market by finding the best fit of the line:

$$-\alpha \cdot log(f) + c = log\ (power)$$

When this is a good fit, we can transform that equation into

$$power = \frac{\beta}{f^{\alpha}}, \qquad where\ c = \log(\beta).$$

That is

$$\text{power} \sim \frac{1}{\text{f}^{\alpha}}$$

When $\alpha = 1$, the stock is a pink noise. With the data of Standard & Poor's index (S&P 500), We used Least Squared Method to find a best fit on the power spectrum of the stock data. Figure 3 shows the fitting result and we get the following fitting function:

$$-0.9044 \cdot \log(\text{frequency}) + 5.6492 = \log\ (\text{power})$$

That means

$$\text{power} \sim \frac{1}{\text{f}^{0.9044}} \approx \frac{1}{\text{f}}$$

This examination tell us that the pink noise indeed exist in the stock market, which means the changed of the price in stock market is random like noise. Additionally, the property of pink noise (1) equal energy per octave and (2) more low-frequency components than white noise make the investors more difficult to get the dominant information from it.

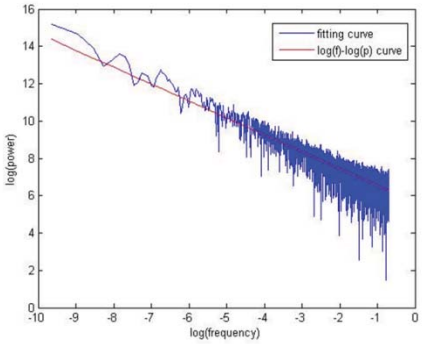


Figure 3. fitting result of Pink noise

III. SUPPORT VECTOR REGRESSION FOR PREDICTION OF STOCK TREND

SVM algorithm developed by Vapnik [3] is based on statistical learning theory [2,4,5]. It can be used for both classification and regression task. In the case of regression [6,7,8,9,10], the goal is to construct a hyperplane that lies "close" to as many of the data points as possible. Therefore, the objective is to choose a hyperplane with small norm while simultaneously minimizing the sum of the distances from the data points to the hyperplane. The main principle is the same as SVM classification, but we have a new function to be minimized. In the ε-insensitive support vector regression, our goal is to find a function *f(x)* that has an ε deviation from the actually obtained target y_i for all training data and at the same time is as flat as possible. Suppose *f(x)* takes the following form: $f(x) = wx + b,\ w \in X, b \in R$, then we have to solve the following problem:

$$\min \frac{1}{2} \parallel w \parallel^2$$

Subject to

$$y_i - wx_i - b \leq \varepsilon$$
$$b + wx_i - y_i \leq \varepsilon$$

In the case where the constraints are infeasible, we introduce slack variables ξ_i, ξ_i^*, this case is called soft margin formulation, and is described by the following problem.

$$\min \frac{1}{2} \parallel w \parallel^2 + C \sum_{i=1}^{l} (\xi_i + \xi_i^*)$$

Subject to

$$y_i - wx_i - b \leq \varepsilon + \xi_i,$$
$$b + wx_i - y_i \leq \varepsilon + \xi_i^*,$$
$$\xi_i, \xi_i^* \geq 0,$$
$$\text{C} > 0.$$

Where *C* determines the trade-off between the flatness of the *f*(*x*) and the amount up to which deviations larger than ε are tolerated. This is called ε-insensitive loss function $|\xi|_{\varepsilon}$ and is described by

$$|\xi|_{\varepsilon} = \begin{cases} 0 & if |\xi| \leq \varepsilon \\ |\xi| - \varepsilon & \text{if } |\xi| > \varepsilon \end{cases}$$

Compared to other neural network regressors, Support Vector Machine has three distinct characteristics when it is used to estimate the regression function. First of all, it estimates the regression using a set of linear functions that are defined in a high-dimensional feature space. Second, SVM carries out the regression estimation by risk minimization, where the risk is measured using Vapnik's-insensitive loss function. Third, SVM implements the SRM principle which minimizes the risk function consisting of the empirical error and a regularized term.

IV. Application of Support Vector Regression

Before performing the regression, we need to normalize the data value (see Figure 4). Our data set with 7 attributes is a big table of data like Table I. The regression analysis focus on how the Open price on the *(i+1)-th* day changes when the Open price, High price, Low price, Volume, and Adjusted close price on the *i-th* day vary. Our purpose is to find the following relationship

(Open(*i*), High(*i*), Low(*i*),Volume(*i*), Adj_Close(*i*)) => Open(*i+1*), by regression analysis.

TABLE I. PART OF THE DATA SET

Date	Open	High	Low	Close	Volume	Adj Close
2011-5-27	2789.02	2801.15	2788.29	2796.86	1641320000	2796.86
2011-5-26	2756.31	2787.33	2756.06	2782.92	1904400000	2782.92
2011-5-25	2739.99	2771.38	2739.85	2761.38	1894510000	2761.38
…	…	…	…	…	…	…
1990-2-6	425	425.2	422.2	424	119700000	424
1990-2-5	422.9	424.8	422.4	424.7	108350000	424.7
1990-2-8	428.8	430.2	427.1	427.3	140030000	427.3

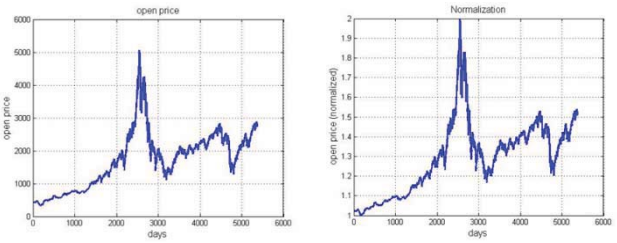


Figure 4. Original data V.S. Normalized data

In the process of applying SVR, in order to obtain a good effect of the regression result, two proper parameters for SVR need to be selected. They are punishment parameter *c* and kernel parameter *g*. The regression performances vary with different selection of these parameters. Here we use Grid search to find the best pair of parameters. A rough and fast Grid search performed at the first stage is for shrinking the searching region, and then a refined Grid search is given to that shrinked region for searching better solution. Figure 5 shows a contour map for SVR parameters selection, where x-axis represents the range of parameter *c*, y-axis represents the range of parameter *g*. and the number on the contour represents the cross validation mean squared error (CVmse) if selecting that pair of parameters located in that contour. Figure 6 is the 3D view of this CVmse distribution. Figure 7 and Figure 8 shows the CVmse distribution for the second stage of the parameter selection, from which we can see that the searching region is dwindled from [-8, 8] to [-4, 4]. And the final parameters selected for this case are c=0.25 and g=1.4142 with the CVmse 0.00086203.

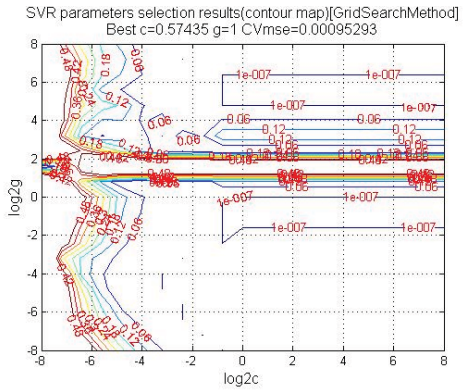


Figure 5. SVR parameters selection results (contour map)

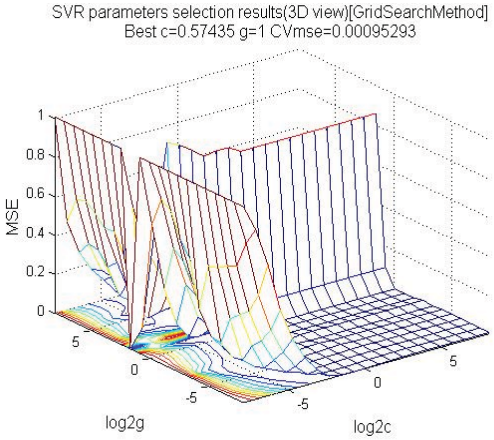


Figure 6. SVR parameters selection results (3D view)

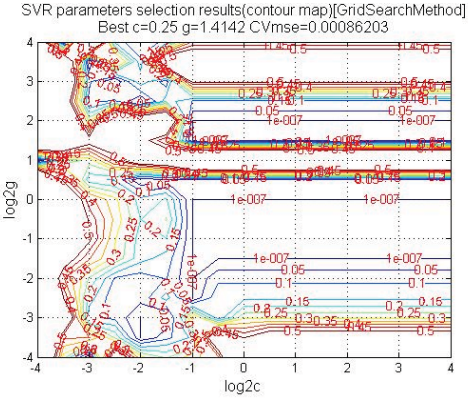


Figure 7. SVR parameters selection results (contour map)

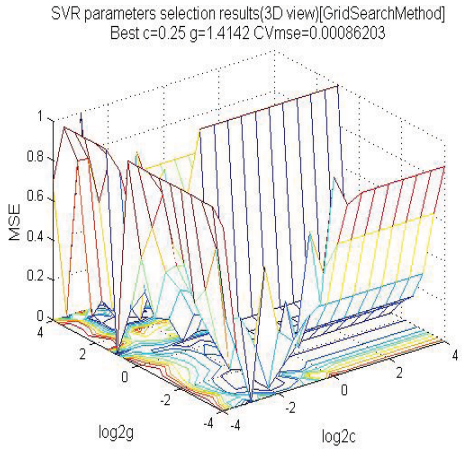


Figure 8. SVR parameters selection results (3D view)

Mean squared error (MSE) for regression is 2.53081e-005, and correlation R = 99.9373%. Figure 9 shows the effect of our regression by plotting the original data and regressive data. Figure 10 and Figure 11 show the absolute error and relative error of the regression separately.

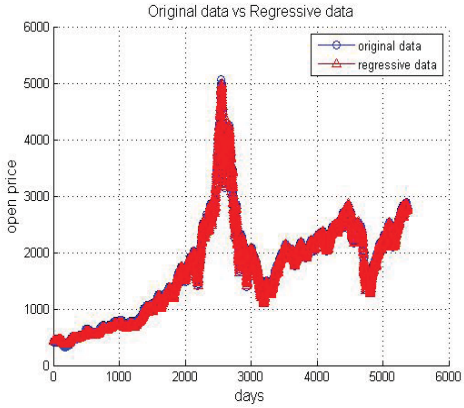


Figure 9. Regression result

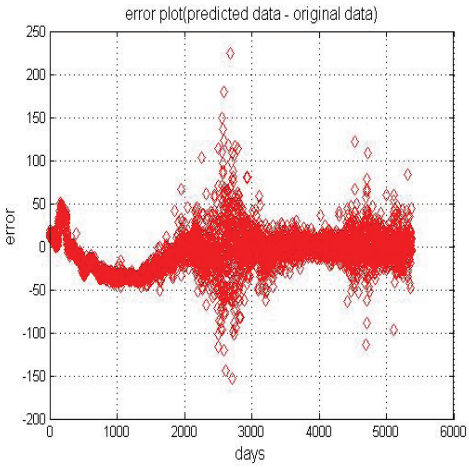


Figure 10. Absolute error

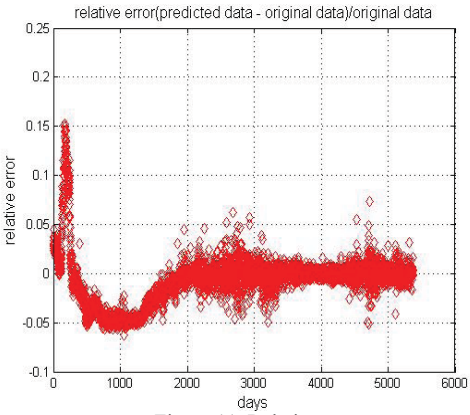


Figure 11. Relative error

V. CONCLUSIONS

In this paper, we study the use of support vector regression to predict stock movement direction. SVR is a promising type of tool for stock forecasting. This is a clear message for financial forecasters and traders, which can lead to a capital gain. However, each method has its own strengths and weaknesses. Although we can fit the historical data well and build the relationship between the attributes, the use of regression function for future stock prediction still remain to be careful, which just provides a useful tool for assisting us making decision on the stock market.

REFERENCES

[1] Wei Huang, Yoshiteru Nakamori, Shou-Yang Wang, "Forecasting stock market movement direction with support vector machine", Computers & Operations Research, Volume 32, Issue 10, October 2005, Pages 2513–2522.

[2] C.J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition", Data Mining and Knowledge Discovery, Volume 2, pp. 1-43, Kluwer Academic Publishers, Boston, 1998.

[3] C. Cortes and V. Vapnik, "Support Vector Networks", Machine Learning, 20, 273-297, 1995.

[4] M.Pontil and A. Verri, "Properties of Support Vector Machines", Technical Report, Massachusetts Institute of Technology, 1997.

[5] E.E. Osuna, R. Freund and F. Girosi, "Support Vector Machines: Training and Applications", Technical Report, Massachusetts Institute of Technology, Artificial Intelligence Laboratory, AI Memo No. 1602, , 1997.

[6] N. Ancona, Classification Properties of Support Vector Machines for Regression, Technical Report, RIIESI/CNR-Nr. 02/99.

[7] Jianwen Xie, Jianhua Wu, Qingquan Qian: Feature Selection Algorithm Based on Association Rules Mining Method. ACIS-ICIS 2009: 357-362.

[8] T. Joachims, Making Large-Scale SVM Learning Practical, Technical Report, LS-8-24, Computer Science Department, University of Dortmund, 1998.

[9] A.J. Smola and B. Scholkopf, A Tutorial on Support Vector Regression, NEUROCOLT2 Technical Report Series, NC2-TR-1998-030, 1998.

[10] S. Haykin, Neural Networks: A Comprehensive Foundation, Prentice Hall, Inc., New Jersey, 1999