Stock Market Trend Prediction Using Support Vector Machines and Variable Selection Methods

Hakob Grigoryan

Department of Economic Informatics and Cybernetics, Bucharest University of Economic Studies, Piata Romana 6, 010374, Bucharest, Romania

***Abstract*—In this paper, a prediction model integrating machine learning and statistical analysis tools is presented to predict the trend of stock market. The proposed approach consists of three stages, as follows. In the first stage, the system uses technical analysis to calculate useful indicators based on historical data. Then, two different variable selection methods are applied to select the most important variables that describe the given data set. Finally, support vector machines (SVM) has been used to construct the forecasting model. The hybridized approach was tested to solve the prediction task of directional changes in Dow Jones industrial average (DJIA) index. To evaluate the effectiveness of the use of variable selection techniques in construction of prediction models, this paper compares the performance of the proposed model with the standard SVM-based method. The study concludes that the use of a successful feature extraction technique can improve the forecasting accuracy of the prediction model.**

Keywords- machine learning; support vector machines; time series; technical analysis; data mining

I. INTRODUCTION

In recent years, rapidly developing markets and large availability of information technologies opened up new opportunities for investors and researchers interested in financial domain. Stock exchanges are regulated financial markets where securities are traded governed by the forces of demand and supply. In general, the behavior of stock prices is described as non-stationary and chaotic. Due to the different internal and external factors affecting financial markets, development of an efficient forecasting model is a complex task [1].

Various methods have been developed to forecast the behavior of stock market prices based on historical data. In the past, researchers have used traditional econometric tools, such as Autoregressive Integrated Moving Average (ARIMA) and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) to predict stock prices [2,3]. Stock prices prediction with conventional statistical methods has proven to be less effective as a result of non-linear characteristics of financial time series. On the contrary, machine learning-based methods, such as Artificial Neural Networks (ANNs) proved to be prone to inherent noise and more suitable for financial data prediction task [4,5].

Support Vector Machines (SVM), used mainly for solving classification and regression problems in time series domain, is a novel machine learning technique proposed by Vapnik [6]. It is based on the structural risk minimization principle and has a global optimum [7]. Compared with other supervised learning techniques, SVMs exhibit better generalization performance and tend to be resistant to over-fitting problem which makes them suitable for stock market prediction task [8,9].

In financial time series analysis with a large number of input variables, data pre-processing techniques are used to transform, prepare and reduce the number of features for the forecasting model. Among them, feature selection techniques are used to identify and select important variables from huge data sets that have the most influence on the target variable. Current studies in the field suggest that the use of different pre-processing techniques have huge impact on the accuracy of the prediction results [10-13].

This paper presents a stock market trend prediction model integrating support vector machines and variable selection techniques. In addition, this work analyzes the effectiveness of the use of technical indicators as input variables for the prediction model. The experiments are conducted based on real high-frequency data of the most widely quoted stock of the Dow Jones Industrial Average (DJIA) index..

The remainder of the paper is organized as follows. In Section 2, the theoretical background of support vector machines (SVM) is presented. Subsequently, Section 3 describes the variable selection techniques used in the studies. Then, a research methodology is proposed in the Section 4. Finally, experimentally established results are summarized and discussed in Section 5.

II. SUPPORT VECTOR MACHINES FOR REGRESSION

Let $S=\{(x_1, y_1),..., (x_n,y_n)\}$ be a set of training data, $x_i \epsilon R^d$, $y_i \epsilon R$, $i=1,...,n,$ where n is the number of training data points, d is the dimension of training dataset, x_i is the model's input vector and y_i is the desired target.

The kernel-based SVM discrimination hyperplane is defined by:

$$f(x)=w'\cdot\varphi(x)+b \tag{1}$$

where w is a weight vector and b is bias, and φ is the feature extractor .

In support vector regression, the main goal is to find a function $f(x)$ that has at most ε deviation from the actually obtained targets y_i for all the training data, and small w or most possible flat f.

The coefficients w and b can be determined by solving the following optimization problem:

$$minimize \quad \frac{1}{2}\|w\|^2 \tag{2}$$

$$subject\ to \begin{cases} y_i - w^T x_i - b \\ w^T x_i + b - y_i \end{cases}$$

To handle the feasibility issues of the optimization problem, soft margin slack variables ξ_i , ξ^*_i are introduced such that:

$$minimize \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{3}$$

$$subject\ to \begin{cases} y_i - w^T x_i - b \le \varepsilon + \xi_i \\ w^T x_i + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \\ C > 0 \end{cases}$$

where $(\xi_i+\xi^*_i)$ is the empirical risk, $½(||w||)^2$ is the structure risk, and C is called a regularization factor which determines the trade-off between empirical risk and structure risk.

The optimization problem proposed in (3) can be solved by introducing Lagrange multipliers a_i and a_i^* described in [7]. Consequently, the general form of the SVM-based regression function can be given by:

$$f(x) = \sum_{i=1}^{l}(a_i - a_i^*)K(x_i,x) + b \tag{4}$$

where $K(x_i,x_j)$ is the kernel function.

Any function that satisfies the Mercer's conditions can be used as a kernel function [14]. This research adopts the most widely used kernel function, a Gaussian radial basis function (RBF), described as:

$$K(x,x') = exp(-\|x - x'\|^2/2\sigma^2) \tag{5}$$

III. VARIABLE SELECTION

In high frequency financial data, the number of features is usually very large, some of these variables being meaningless when the prediction model is developed. Variable selection is a method to select significant features from the point of view of forecasting model. Selecting relevant input variables can simplify the prediction model, improve training time and forecasting accuracy. In the following, two variable selection techniques are presented.

A. Peeling Method

The Peeling method proposed by Võhandu and Krusberg [15] finds and selects the most influential variables that describe the main part of the given data set.

For a given correlation matrix C of size $n\times n$ and a certain threshold ε, the algorithm works as follows:

Step 1. For every column a measure of influence S_j is calculated: $S_j = \frac{\sum_{i=1}^{n} c^2_{ij}}{c_{jj}}$ where $c_{jj>}$ ε;

Step 2. Compute the maximal measure of S_j denoted by $S^{(k)}=maxS_j$ which identifies the most important variable, and the superscript k shows the iteration $(k=1,...,m \le n)$;

Step 3. Divide the correlation coefficients of the maximal variable c_j by the square root of the diagonal element c_{jj} and construct a new factor matrix B where the first vector is the transformed column vector $b_1 = c_j/\sqrt{c_{jj}}$;

Step 4. Find the residual matrix $C^{(1)} = C - b_1 b'_1$;

Step 5. Repeat the process r times, where $r\le n$.

According to the elimination order, the first most r important variables will be considered as input variables for the prediction model.

B. Cross-Correlation Based Method

Cross-correlation technique is a basic tool in the analysis of multivariate time series. The main idea is to estimate dependencies between time series by computing the cross-correlation function. The maximal value of the function closer to 1 shows the high dependency between analyzed variables.

The variable selection method using cross-correlation analysis is given as follows:

Let Tr be a given threshold value, XT the set of T input variables and Y the target. In other words, $XT(i)$ is the time series corresponding to the i^{th} variable and Y_t is the target variable's value at the moment of time t .

Step 1. For each i, $1\le i\le T$, compute the cross-correlation coefficient between $XT(i)$ and Y using:

$$r_{XT(i),Y} = \frac{\sum_{t=1}^{T}(XT_t(i) - \overline{XT(i)})(Y_t - \bar{Y})}{\sqrt{\sum_{t=1}^{T}(XT_t(i) - \overline{XT(i)})^2 \sum_{t=1}^{T}(Y_t - \bar{Y})^2}} \tag{6}$$

$$\overline{XT(i)} = \frac{1}{T}\sum_{t=1}^{T} XT_t(i)\,, \bar{Y} = \frac{1}{T}\sum_{t=1}^{T} Y_t$$

Step 2. Select the most correlated variables according to cross-correlation coefficients bigger than threshold value

$$r_{XT}(0)_Y > Tr.$$

The selected variables are considered the inputs of the prediction model.

IV. Research Methodology

In this paper, SVM technique is combined with variable selection methods to construct the stock market trend prediction model. The following steps were used to build the model: 1. Technical analysis is used as a processing step to determine the most important indicators based on the historical data. 2. Data normalization is done to normalize the original data into one scale. 3. Variable selection techniques were applied to choose the most influential variables for the prediction model. 4. Prediction model is constructed using support vector machines for regression. 5. Comparative analysis is conducted to examine the effectiveness of the proposed model.

The research data used in this paper consists of the daily observations of Dow Jones Industrial Average (DJIA) index operated by S&P Dow Jones Indices. The data set covers 400 trading days, from 6/9/2014 to 1/7/2016. The data collection includes daily traded volume, opening, closing, highest and lowest prices of the stock and 7 most widely used technical indicators obtained from the technical analysis of the stock. The target variable is represented by the closing price.

After data preprocessing, a variable selection is applied to select the significant variables to be fed to the SVM model. 11 variables are used in the process of variable selection implemented by cross-correlation based method and by Peeling method.

The cross-correlation coefficients between closing price and the input variables are presented in Table 1. We considered that the threshold value is set to 0.9. Then, the selected variables are: opening, highest and respectively lowest price, and moving average (MA).

TABLE I. LIST OF VARIABLES AND CORRELATION COEFFICIENTS

Name of attribute	Coefficients
Lowest Price	0.9879
Highest Price	0.9851
Opening Price	0.9652
Traded Volume	-0.1399
Moving Average (MA)	0.9105
Bollinger Bands (BB)	0.7916
MA Convergence/Divergence (MACD)	0.5360
Relative Strength Index (RSI)	0.4073
Momentum Oscillator (MTM)	0.3472
Williams' Percent Range (%R)	0.3273
Stochastic Oscillator (%K)	0.0245

The peeling algorithm is applied to the input data to select the variables which describe the main part of the variation of the given data set. By applying this method, two variables, namely technical indicators called Bollinger Bands and %K stochastic were selected.

Variables selected via peeling algorithm and cross-correlation-based variable selection algorithm are fed as an input to the SVM-based model for the training and prediction. The whole data set is divided into a training set (70% of the data) and a testing set (30% of the data).

The general forecasting model is constructed based on the following equation:

$$\hat{Y}_{(t+1)} = f(Y_t^{(d)}, X_t^{(d)}) \qquad (7)$$

where $\hat{Y}_{(t+1)}$ is the predicted value, d is the delay, and

$$Y_t^{(d)} = \{Y_t, Y_{t-1}, Y_{t-2}, \ldots, Y_{t-d+1}\}$$

$$X_t^{(d)} = \{X_t, X_{t-1}, X_{t-2}, \ldots, X_{t-d+1}\}$$

The evaluation metric used to measure the percentage of correct directional prediction is called Prediction Of Change In Direction (POCID) and is defined by:

$$POCID = 100\,\frac{\sum_{t=1}^{N} Trend_t}{N} \qquad (8)$$

where

$$Trend_t = \begin{cases} 1, & if\ (T_{(t+1)} \quad T_t)(P_{(t+1)} \quad P_t) > 0 \\ 0, & otherwise \end{cases}$$

and T and P are the vectors of target and predicted values, respectively.

V. Experimental Results and Conclusions

A series of experiments have been conducted to examine the effectiveness of the proposed integrated model, and compare the results of presented methodology against the model based only on SVM technique.

In Figure I, the forecasted values versus the real ones are displayed while the trend prediction accuracy is presented in Table II.

TABLE II. PREDICTION ACCURACY

Model	Accuracy
SVM	82.5%
CC+SVM	84.1667%
Peeling+SVM	89.1667%

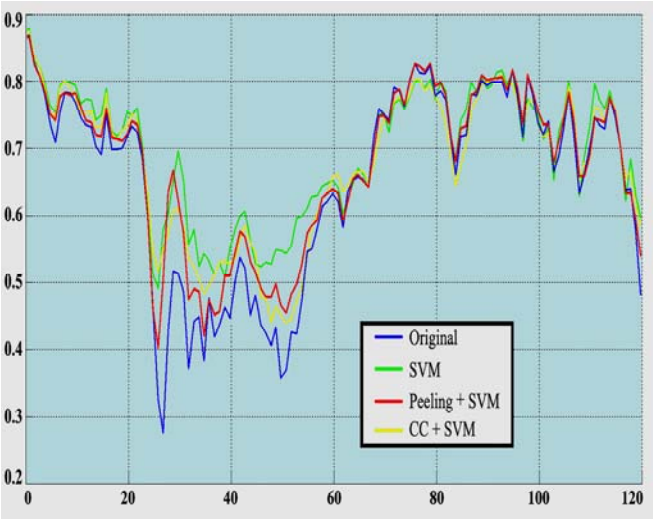

FIGURE I. PREDICTION VERSUS ACTUAL VALUES IN CASE OF DJIA STOCK

According to the experimental results, both hybridized models showed improvements in the prediction accuracy compared with the SVM model. As a result, the accuracy increased from 82.5% to 84.1667% and 89.1667% by using two different variable selection methods combined with SVM technique. The best trend prediction accuracy has been achieved when the integrated model combining Peeling algorithm and SVM technique has been applied.

ACKNOWLEDGMENT

The author would like to thank Mr. Leo Võhandu, the professor emeritus of Tallinn University of Technology and Mrs. Catalina Cocianu, the professor of Bucharest University of Economic Studies for their helpful suggestions and comments on using statistical and machine learning techniques.

REFERENCES

[1] Gujarati, D. N. (2003). Basic Econometrics. 4th. New York: McGraw-Hill.

[2] Cocianu, C. L., & Grigoryan, H. (2015). An Artificial Neural Network for Data Forecasting Purposes. *Informatica Economica*, *19*(2), 34.

[3] Chong, C. W., Ahmad, M. I., & Abdullah, M. Y. (1999). Performance of GARCH models in forecasting stock market volatility. Journal of Forecasting, 18(5), 333-343

[4] A.N. Refenes, A. Zapranis, G. Francis, Stock performance modeling using neural networks: a comparative study with regression models, Neural Networks (1994) 375–388.

[5] Adebiyi, A. A., Adewumi, A. O., & Ayo, C. K. (2014). Comparison of ARIMA and artificial neural networks models for stock price prediction. Journal of Applied Mathematics, 2014.

[6] Vapnik, V. (1998). Statistical Learning Theory. John Wiley&Sons. Inc., New York.

[7] Smola, Alex J., and Bernhard Schölkopf. "A tutorial on support vector regression." Statistics and computing 14.3 (2004): 199-222.

[8] Kim, K-J. (2003). Financial time series forecasting using support vector machines. Neurocomputing, 55, 307-319.

[9] Huang, Wei, Yoshiteru Nakamori, and Shou-Yang Wang. "Forecasting stock market movement direction with support vector machine." Computers & Operations Research 32.10 (2005): 2513-2522.

[10] Kazem, A., Sharifi, E., Hussain, F. K., Saberi, M., & Hussain, O. K. (2013). Support vector regression with chaos-based firefly algorithm for stock market price forecasting. Applied soft computing, 13(2), 947-958.

[11] Lee, M. C. (2009). Using support vector machine with a hybrid feature selection method to the stock trend prediction. Expert Systems with Applications, 36(8), 10896-10904.

[12] Tsai, C. F., & Hsiao, Y. C. (2010). Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. Decision Support Systems, 50(1), 258-269.

[13] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.

[14] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

[15] Võhandu L, & Krusberg H (1977). A Direct Factor Analysis Method. The Proceedings of TTU, 426, 11–21.